Tools for stochastic weather series generation in R environment

Emanuele Cordano[1,2], *Emanuele Eccel*[1*]

Abstract: *The "R" packages RMAWGEN and RGENERATEPREC aim to generate daily maximum and minimum temperature and precipitation series preserving the meteorological coherence with observations, required for many agro-ecological applications. The implemented methods are designed to work with existing tools implemented in the R environment, such as vector autoregressive models (VAR) and other extensions of generalized linear models with logit regression, for generation of daily precipitation series; one aim of the algorithm is the conservation of the temporal and spatial correlations among variables. The internal parameters of the weather generator are calibrated from observed time series. The article describes the main features of the presented packages and an application to a dataset of daily weather time series recorded at 28 sites in Trentino (Italy) and its neighbourhood.*

Keywords: *Multisite weather generators, vector auto-regression, logit regression, temperature, precipitation, climate.*

Riassunto: *I pacchetti "R" RMAWGEN e RGENERATEPREC generano serie giornaliere di temperature massime e minime e precipitazioni preservando la coerenza meteorologica con le osservazioni, richiesta per molte applicazioni agro-ecologiche. I metodi si avvalgono dell'uso di librerie già implementate in ambiente R, in particolare modelli vettoriali auto regressivi (VAR), per mantenere le correlazioni temporali e spaziali tra le variabili, ed altre estensioni di modelli generalizzati con regressione logistica, per la generazione delle serie di precipitazioni. Uno degli scopi dell'algoritmo è la conservazione della correlazione spaziale tra le variabili. I parametri interni dei* weather generator *sono calibrati dalle serie osservate. L'articolo descrive le principali caratteristiche dei pacchetti presentati e le applicazioni ad un archivio di serie meteorologiche giornaliere registrate in 28 siti in Trentino e regioni limitrofe.*

Parole chiave: *Generatori meteorologici stocastici multisito, auto-regressione vettoriale, regressione logit, temperatura, precipitazione, clima.*

1. INTRODUCTION

Stochastic generators of weather variables, called "Weather Generators" (WGs), have been widely developed in the recent decades for hydrological and agro-ecological applications (Richardson, 1981, Racsko *et al.*, 1991, Semenov and Barrow, 1997, Parlange and Katz, 2000, Liu *et al.*, 2009, Chen *et al.*, 2012, Chen and Brissette, 2014). Applications in agricultural meteorology require the contemporary generation of series of more than one quantity, at least temperature and precipitation, and possibly more (Rocca et al., 2012), e.g. for some pest modelling, or for water balance modelling. A typical application of WGs is the reproduction of daily weather time series from downscaled monthly climate predictions (Mearns *et al.*, 2001, Wilks and Wilby, 1999, Qian *et al.*, 2002, Semenov and Stratonovitch, 2010). If high-resolution application models are to be used with downscaled series, a meteorological consistence of generated series is required, suggesting the use of a multi-site weather generator. Algorithms to represent historical spatial dependences of weather variables have been developed e.g. by Wilks, 1998, Khalili *et al.*, 2009, Serinaldi, 2009, Bàrdossy and Pegram, 2009, Kleiber *et al.*, 2013. Wilks (1998) simulated rainfall occurrences through a generation of combinations of Gaussian random variables and established a relationship for each pair of rain gauges between Gaussian variables correlation and binary (precipitation occurrence) values. In this way, weather generators can reproduce at least partially spatial correlations; this approach is widely cited in literature (Mehrotra *et al.*, 2006, Brissette *et al.*, 2007, Serinaldi, 2009, Thompson *et al.,* 2007, Mhanna and Bauwens, 2011, Kleiber *et al.*, 2012).

Recently, statistical methods useful for weather generation, originally developed in environmetrics and econometrics, were made available in the R platform (R Core Team, 2014). In this context, a suite of two weather generator tools was developed within the R environment through the creation of two packages: RMAWGEN and RGENERATEPREC. In particular, RMAWGEN (R Multisite Auto-regressive Weather Generator - Cordano and Eccel, 2011) was developed to cope with the demand for high-resolution climatic

[*] Corresponding author's e-mail: emanuele.eccel@fmach.it

[1] Sustainable Agro-ecosystems and Bioresources Department, IASMA Research and Innovation Centre, Fondazione Edmund Mach - San Michele all'Adige (I).

[2] Rendeana100, Tione di Trento.

Received 6 February 2016, accepted 10 June 2016

DOI:10.19199/2016.3.2038-5625.031

scenarios but also to create a flexible tool for engineers and researchers in agro-environmental modelling. For this reason, RMAWGEN uses existing R tools for vector auto-regressive models (Pfaff, 2008b), employed for generation of weather variables (Adenomon and Oyejola, 2013; Luguterah *et al.*, 2013; Shahin *et al.*, 2014), and to let the users work with other R spatio-temporal tools for data analysis and visualization (Bivand et al., 2008; Loecher and Berlin School of Economics and Law, 2012; Kahle and Wickham, 2013).

The paper is organized in the following sections: a description of the mathematical methods applied in the weather generators; the usage of the weather generator with an example on a multi-site dataset, including model validation through statistical tests; and, finally, some concluding remarks.

2. METHODS

The following methods were utilized for generation of daily temperature series:

- Vector Auto-Regressive models (VAR);
- Gaussianization of continuous variables;

while precipitation generation is based on:

- Single-Site Logistic Regression for Precipitation
- Occurrence based on Generalized Linear Models (GLMs);
- Multi-site generation of random values based on Wilks' correlation matrix;
- Random Generation of Precipitation Amount for each site using simultaneous precipitation occurrences as predictors.

2.1 Vector Auto-Regressive models (VAR)

The basic idea of RMAWGEN consists on the generation of daily precipitation and temperature by using Vector Auto-Regressive Models (VARs). A set of K random variables can be described by a Vector Auto-Regressive Model VAR(K,p) as follows (Lütkepohl, 2007; Hamilton, 1994; Pfaff, 2008a):

$$x_t = A_1 x_{t-1} + \dots + A_p x_{t-p} + u_t \qquad (1)$$

where x_t is a K-dimensional vector representing the set of weather variables generated at day t by the model, called "endogenous" variables, A_i is a coefficient matrix K x K for i = 1…p and p is the auto-regression order. In absence of exogenous variables, u_t is a standard white noise.

The methods to estimate the parameters of a VAR model, as defined by equation (1), are properly illustrated in Lütkepohl (2007). The auto-regression order must be found before VAR parameter estimation; some information criteria were proposed to assess the best value of p (Akaike and May, 1981; Hannan and Quinn, 1979; Schwarz, 1978; Lütkepohl, 2007). Pfaff (2008b) suggests four indices: AIC(p), HQ(p), SC(p), FPE(p), where p is the lag order. This information criteria to select the optimal value of p are implemented in R package "vars", and then utilized by package "RMAWGEN" and in the examples shown in this article. In RMAWGEN, the implemented methods for VAR parameters refer to the function VAR of "vars" package (Pfaff, 2008b). Further theoretical details about VAR models can be found in Lütkepohl (2007), Hamilton (1994), and Pfaff (2008b).

2.2 VAR diagnostics

Once built the VAR(p) model and estimated its coefficients, the model residuals can be processed through diagnostic tests, which can be summarized in the following, borrowed from econometrics (Pfaff, 2008a):

- Multivariate Portmanteau and Breusch-Godfrey (Lagrange Multiplier tests), which verify the absence of time-autocorrelation of the VAR(p) residuals;
- Jarque-Bera and multivariate skewness and kurtosis tests, which validate the multivariate Gaussian probability distribution of the VAR(p) residuals;
- ARCH-LM tests, which verify the absence of heteroskedasticity of the VAR(p) residuals.

The former verifies that the residuals, calculated by instrumental time series, are not auto-correlated. The second verifies that they are multi-normally distributed, i.e., its skewness and kurtosis are null. In case of stochastic generation, this condition is generally satisfied because the random part of the model is a white noise. The third test, the ARCH-LM (Auto-Regressive Conditional Heteroskedasticity - Lagrange Multiplier) test, analyzes the covariance matrix or the variance (in a univariate case) of residuals in order to verify if the variance of residuals is constant or if it varies with time, i.e. its homoskedasticity or heteroskedasticity, respectively. Lütkepohl (2007) proposed also an ARCH-Portmanteau test for residual heteroskedasticity. Further details about the tests can be found in the already quoted works and in Peña and Rodriguez (2002), and Mahdi and Ian McLeod (2012).

2.3 Gaussianization of continuous variables

VAR models work correctly for normally distributed variables. This requires a normalization procedure for daily meteorological values of both temperature and precipitation, which can be formally expressed as:

$$x_t = G_m(z_t) \tag{2}$$

where z_t is the meteorological time series and G_m is a univariate or multivariate function (so that x_t is multi-normally distributed and can vary according to time, month, and season: subscript m is an indicator of the month. G_m contains all the information on the variable transformation and can be inverted:

$$z_t = G_m^{-1}(x_t) \tag{3}$$

In this case, the function G_m transforms a meteorological variable, whose probability distribution is different for each month, into a random variable with equal probability distribution for all months of the year. The function G_m resumes the "climatic" properties of the m-th month and is operationally determined and parameterized according to the meteorological variable, by sampling monthly the historical time series. This function invokes a coordinate transformation which is called "Gaussianization" (Erdogmus *et al.*, 2006; Laparra *et al.*, 2009).

The one-dimensional Gaussianization is a transformation from a random variable z, whose cumulate probability distribution is $F_z(z)$, into a Gaussian-distributed one x. It can be analytically expressed as follows:

$$x = F_G^{-1}(F_z(z)) \tag{4}$$

where F_G is the Gaussian cumulative function with zero mean and standard deviation equal to 1. If $F_z(z)$ is an invertible function, the transformation between x and y is also invertible:

$$z = F_z^{-1}(F_G(x)) \tag{5}$$

Finally, the transformation between x and z is done by applying the inverse cumulated probability function of x and z, respectively, and vice versa. In case z is a seasonally-changing weather variable, the Gaussianization process can take into account its dependency on month or season: in (2), G_m is defined as depending on month (Fig. 1).

Similarly to the one-dimensional case, it is intuitive that an invertible function exists between any multi-dimensional random variable and a Gaussian variable with the same dimensions (Chen and Gopinath, 2000). Nevertheless, an explicit form of such a function is not as easy to find as in the one-dimensional case and might require high computational effort. Laparra *et al.* (2009) proposed an iterative method based on Principal Component Analysis (PCA). This method is organized in a loop where each iteration contains the following steps:

1. One-dimensional Gaussianization of each component of x, i.e., Marginal Gaussianization (in this step, only the marginal probability distributions are taken into account);
2. Orthonormal transformation of the coordinates based on the eigenvector matrix of the covariance matrix according to Principal Component Analysis (PCA).

The variables become marginally Gaussian and then, by rotation, the marginal Gaussianization is calculated in the other directions of the variable coordinate space. Later, Laparra *et al.* (2011) generalized this transformation and called it "Rotation-based Iterative Gaussianization (RBIG) transform", where the iterative process can be coupled with other kinds of rotation like Independent Component Analysis (ICA) or Random Rotation (RND). The convergence of these methods is explained in Laparra *et al.* (2011). However, as suggested by Laparra *et al.* (2009), in package RMAWGEN, the Gaussianization is performed through PCA rotation (GPCA), because of the lower computational burden and the acceptable quality of results.

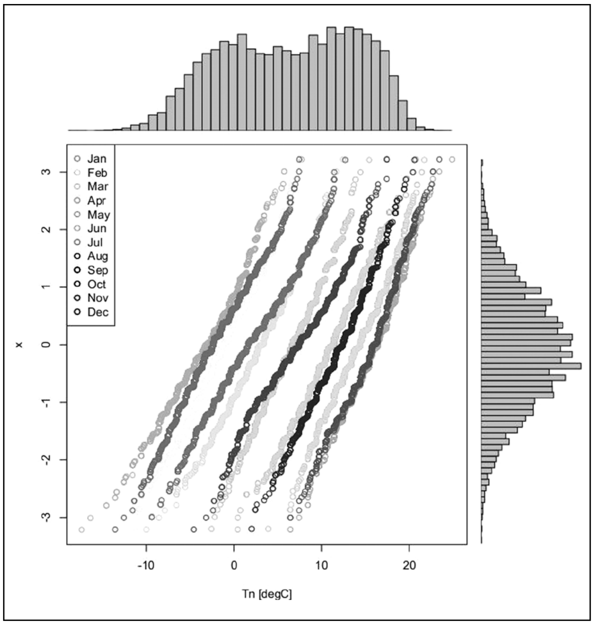

Fig. 1 - Monthly relationship between minimum daily temperature (bimodal) and the transformed Gaussian series x with the respective frequency histogram.

Fig. 1 - Relazione mensile tra la temperature minima giornaliera (bimodale) e la serie gaussiana trasformata x con i rispettivi istogrammi di frequenza.

2.4 Multi-dimensional Generation of Precipitation

The marginal gaussianization of intermittent weather values, e.g. in presence of zeros, as previously described, does not take into account the inter-site correlation: the zeros (no precipitation days) are randomly Gaussianized and the Gaussianized value cannot be affected *a priori* by the precipitation occurrence at another site. If this aspect is not dealt with, precipitation generated with simple Gaussinized VAR models will lack information and the inter-site correlation will be underestimated. To fix this shortcoming, Wilks (1998) introduced an algorithm to transform correlation of a couple of binary values into correlation of the corresponding Gaussian variables. This approach is widely applied to the random generation of daily precipitation occurrence. The dependence within a rainfall gauge network is well known, especially as concerns precipitation occurrence, whereas the correlation in a VAR is strictly connected to the continuous values, like precipitation depth. As can be observed, precipitation occurrence is better spatially correlated than the value of daily precipitation depth.

In order to describe the correlation among precipitation occurrences among several sites, RGENERATEPREC was developed on Wilks' approach for the estimation of the cross-correlations among raingauges and its results compared with that of RMAWGEN. RGENERATEPREC generates precipitation occurrence and precipitation depth separately. According to Wilks' approach, daily precipitation occurrence can be generated by generating normally (Gaussian) distributed random numbers as follows:

$$[X_i]_t = \begin{cases} 0 & P_G([x_i]_t) \leq [p0_i]_t \\ 1 & P_G([x_i]_t) > [p0_i]_t \end{cases} \tag{6}$$

where $[X_i]_t$ is the binary state of precipitation occurrence in the ith site and on the tth day: 0 (dry day/no precipitation occurrence) or 1 (wet day/ precipitation occurrence); $[p0_i]_t$ is the probability of no precipitation occurrence for the ith site and the tth day; $[x_i]_t$ is a normally distributed random variable and $P_G(x)$ is the cumulate probability function of the normalized (Gaussian) distribution. In this work, the probability value $[p0_i]_t$ is conditioned to the state of the previous day(s) and other factors, namely it is calculated by a logistic auto-regression with the following predictors: (I) precipitation occurrence in the previous day(s), and two exogenous variables: (II) daily maximum temperature anomalies (see next section "Case study"), and (III) the day of the year. The logit regression is implemented through a generalized linear model (McCullagh and Nelder, 1989, Chambers and Hastie, 1992), which generates time series of daily precipitation occurrence probability similarly to VAR implementation in package "vars". Finally, the probability value $[p0i]_t$ is calculated singularly for each site and is not conditioned to the precipitation occurrence at the neighbouring sites.

The computation reported in RGENERATEPREC makes use of an R implementation of the normal copula (Yan, 2007, Kojadinovic and Yan, 2010) and the marginal probability values $p0_k$ for each station on a monthly base. Finally, precipitation amount in the wet days is generated by inverting a parametric probability distribution (Breinl *et al.*, 2013, Furrer and Katz, 2008, Li *et al.*, 2012) or by inverting the non-parametric frequency distribution obtained by observations (Cordano, 2015a, 2015b).

3. CASE STUDY

3.1 Geographical description and meteorological series

A dataset containing daily time series is included in RMAWGEN dataset. The dataset, called "*trentino*", contains daily minimum and maximum temperatures and precipitation form 1950 to 2007. The weather data had been recorded and homogenized at 59 sites in Trentino region (North-Eastern Italian Alps) and its neighbourhood (Eccel *et al.*, 2012).

The geography of the area is characterized by a valley system, ranging between 70 m a.s.l. (Lake Garda) to 3769 m (Mount Cevedale). The area covered by the weather station network can be ascribed to a Köppen class ranging from "Cfb" ("temperate, middle latitudes climate, with no dry season") to "Dfc" classification ("microthermal climate, humid all year round") in the more elevated, mountain areas. Precipitation amounts are mostly distributed over two maxima, in the autumn (main) and in the spring (secondary), although in some mountain areas rainfall peaks in summer (Di Piazza and Eccel, 2012).

The region is illustrated in Fig. 2, obtained by package RgoogleMaps (Loecher and Berlin School of Economics and Law, 2012) and easily reproduced by the example script *trentino_map.R* contained in RMAWGEN directory. In this application, RMAWGEN model is calibrated for a 30-year long reference period from 1961 to 1990 and then utilized for a random generation of daily

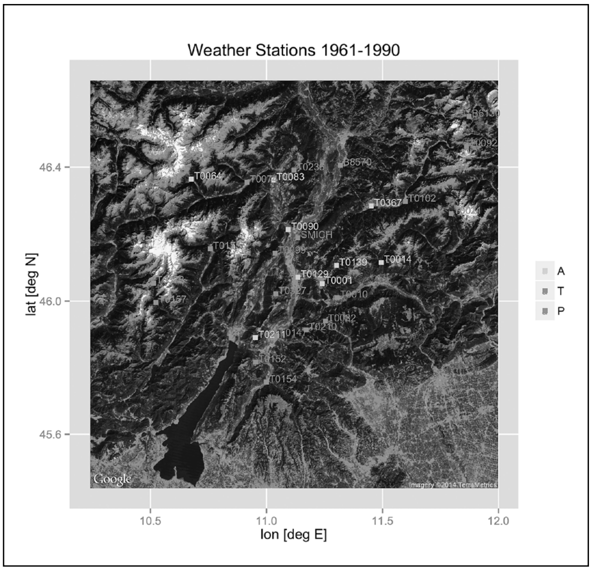

Fig. 2 - The weather stations of *trentino* dataset operating in the period 1961-1990 with time series of daily precipitation and temperature: (T) only temperature time series are complete with no gaps; (P) only precipitation time series are complete with no gaps; (A) both temperature and precipitation time series are complete with no gaps.

Fig. 2 - Le stazioni meteorogiche del data set trentino *funzionanti nel periodo 1961-1990 con le serie temporali di precipitazione giornaliera e temperatura: (T) solo le serie di temperature sono complete; (P) solo le serie di precipitazione sono complete senza lacune; (A) entrambe le serie sono complete senza lacune.*

precipitation and temperature for a 30-year long period with the same climatic properties.

3.2. Temperature generation

3.2.1. Model setting

The daily minimum and maximum temperature are generated for the reference period 1961-1990. The new variable z_t is pre-processed as a vector of anomalies containing the observed mean daily temperature anomaly and the observed daily temperature range, and it is:

$$z_t = \left(\frac{Tx_t + Tn_t}{2} - \frac{Tx,s_t + Tn,s_t}{2} \right) \cup \left(Tx_t - Tn_t \right) \quad (7)$$

where Tx_t and Tn_t are the observed daily maximum and minimum temperature, Tx,s_t and Tn,s_t are the mean daily climatic values of maximum and minimum temperature at each site; they are the result of a daily spline interpolation from the monthly values; and U is the vector-append function. With this deseasonalization, z_t is far from a Gaussian distribution, but the constraint $Tx_t > Tn_t$ is always respected. Consequently, a Gaussian variable is found through Gaussianization, as expressed in (4), which samples the variable z_t monthly at least at the first iteration (eq. 2).

Once a time series vector is created for x_t, the parameters of the GPCA-VAR or VAR model are calibrated. Then, a normality test on VAR residuals is required and residuals are subsequently Gaussianized and normalized through GPCA if they are not Gaussian.

To test the effectiveness of the PCA Gaussianization and the optimal auto-regression order, four different models for random generation were used:

- [P01] auto-regression order p equal to 1, no PCA Gaussianization;
- [P06] auto-regression order p equal to 6, no PCA Gaussianization;
- [P01GPCA] auto-regression order p equal to 1, PCA Gaussianization;
- [P06GPCA] auto-regression order p equal to 6, PCA Gaussianization.

In all the four listed models the observed time series were previously deseasonalized ad marginally gaussianized. With function "VARselect" in package "vars", it is possible to test auto-regression order from 1 to 20 days or more to find the optimal values according to the information criteria; the value of auto-regression order equal to 6 was selected, according to the results of Akaike's AIC criterion.

3.2.2. Results: model validation

As proposed by Pfaff (2008), Luetkepohl (2007), Hamilton (1994) and other authors, the residuals of four VAR models were tested for normality and seriality (autocorrelation). The test results (summarized in Tab. 1, where significance p-values of 0.05 or higher give the same qualitative results) indicate that both normality and serial tests are successful in case of GPCA with autoregression order p equal to 6 (P06GPCA), which, therefore, can be considered the best choice, among those tested. Other cases yielded very low p-values, resulting in unsuccessful verifications. The tests generally highlight the importance of gaussianization pre-processing of temperature.

	Normality test	Seriality test
P01	unsuccessful	unsuccessful
P06	unsuccessful	unsuccessful
P01GPCA	successful	unsuccessful
P06GPCA	successful	successful

Tab. 1 - Results of normality and seriality tests for temperature.

Tab. 1 - Risultati dei test di normalità e serialità per la temperatura.

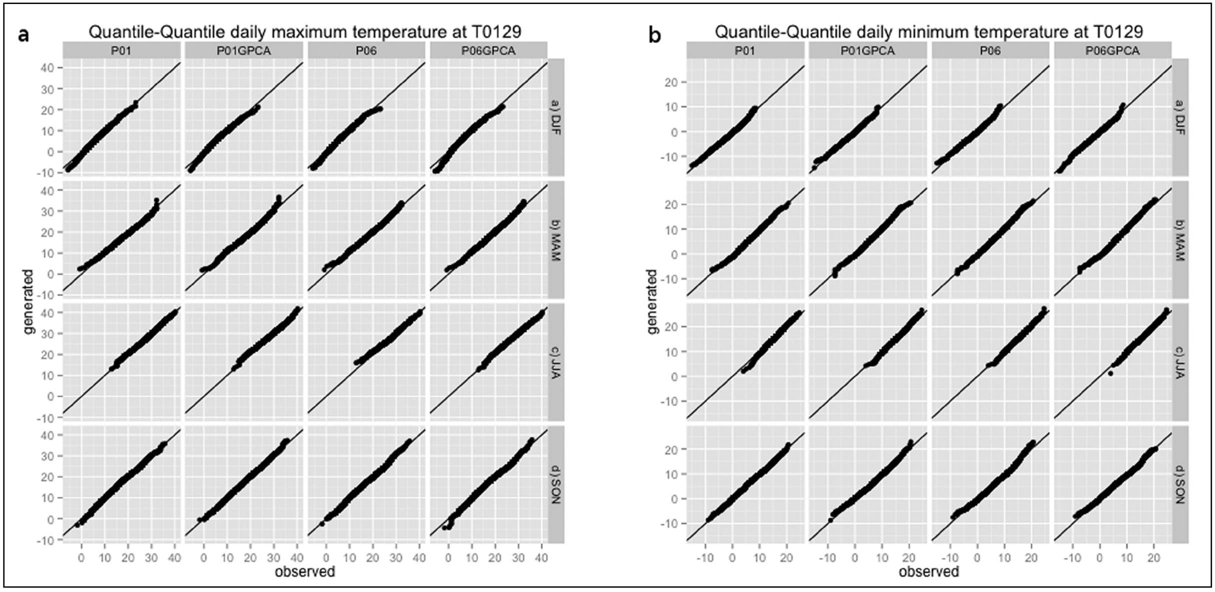

Fig. 3 - Q-Q plots for daily maximum (a) and minimum (b) temperature (generated vs observed) at station T0129 in the four seasons.
Fig. 3 - Diagrammi Q-Q per le temperature massime e minime giornaliere (generate contro osservate) per la stazione T0129 nelle quattro stagioni.

The PCA-Gaussianized VAR model aims to preserve the probability distribution and spatio-temporal correlation of the corresponding observed data. Any comparison of daily values being meaningless, due to the random generation, the validation of a weather generator can only consist on comparisons between statistics of observed and modelled series. Fig. 3 shows the Quantile-Quantile plot ("Q-Q plots") between the generated and observed daily maximum and minimum temperature at station T0129 (see Fig. 3). "Q-Q plots" are particularly useful, showing distribution fittings between

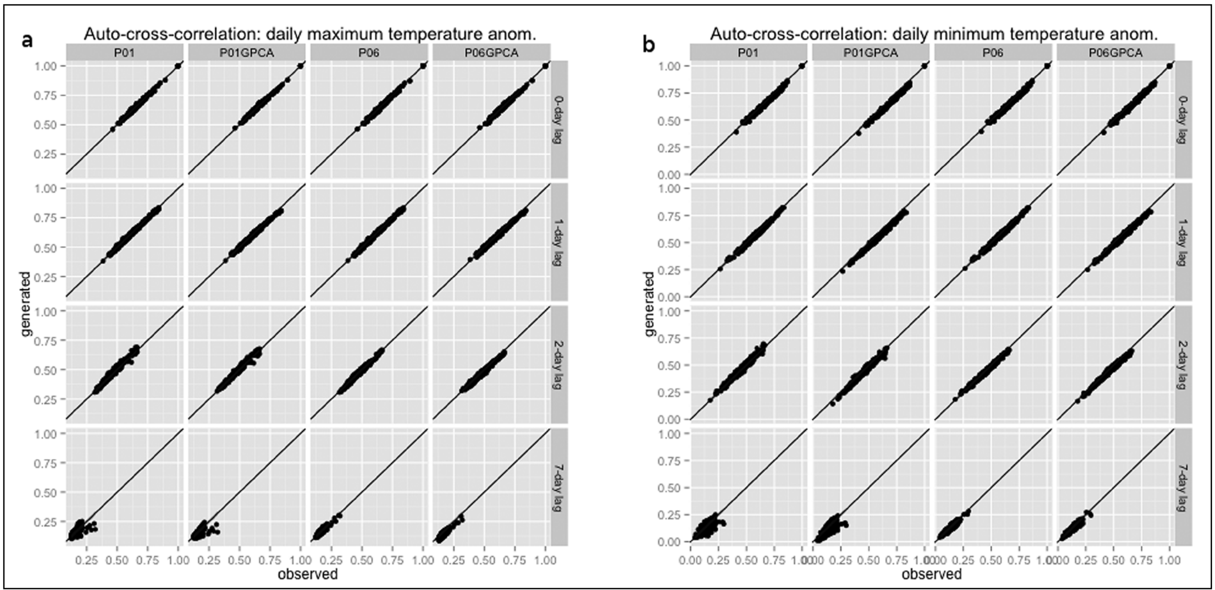

Fig. 4 - Auto- and Cross-correlation for daily maximum (a) and minimum (b) temperature anomalies (generated vs observed). Columns correspond to the four VAR models used in RMAWGEN generations. Rows correspond to the different lags used for the computation of crosscorrelation: no lag (0 days), 1 day, 2 and days 7 days.
Fig. 4 - Auto- e cross-correlazione per le anomalie di temperatura massima e minima giornaliera (generate contro osservate). Le colonne corrispondono ai quattro modelli VAR usati nelle generazioni di RMAWGEN. Le file corrispondono ai differenti lag *per il calcolo della correlazione: nessun lag (0 giorni), 1 giorno, 2 giorni e 7 giorni.*

observed and generated series. The quantiles of observed and simulated data are plotted for each VAR model type and each season (Fig. 3). The goodness-of-fit of observed versus generated daily temperature is quite satisfactory: most of the points appear along the quadrant bisector for all the four considered VAR models.

Fig. 4 shows the autocorrelation and the cross-correlation of temperature anomaly (observed or simulated daily value rescaled to the typical climate value, spline-interpolated by monthly climatology) referred to each couple of stations and for four different lags expressed in days: 0 (cross-correlation), 1, 2, and 7 days. The correlations are plotted with respect to both maximum and minimum daily temperature and are calculated for each of the implemented VAR models. The coefficients obtained from both observed and generated time series are very similar and stay on the quadrant bisect, especially in case of simultaneous correlation (lag 0) and one-day autocorrelations, and range between 0.6 and 1.0. As expected, a less good agreement is found after a 7-day lag, in the cases with p=1 (P01 and P01GPCA); nevertheless, correlations are lower (less than 0.25), and become less significant.

3.3 Precipitation generation

3.3.1 Model setting

Daily precipitation is separately generated for the reference period 1961-1990. through a random generation with an auto-regression based on generalized linear models (Chambers and Hastie, 1992) implemented in the RGENERATEPREC

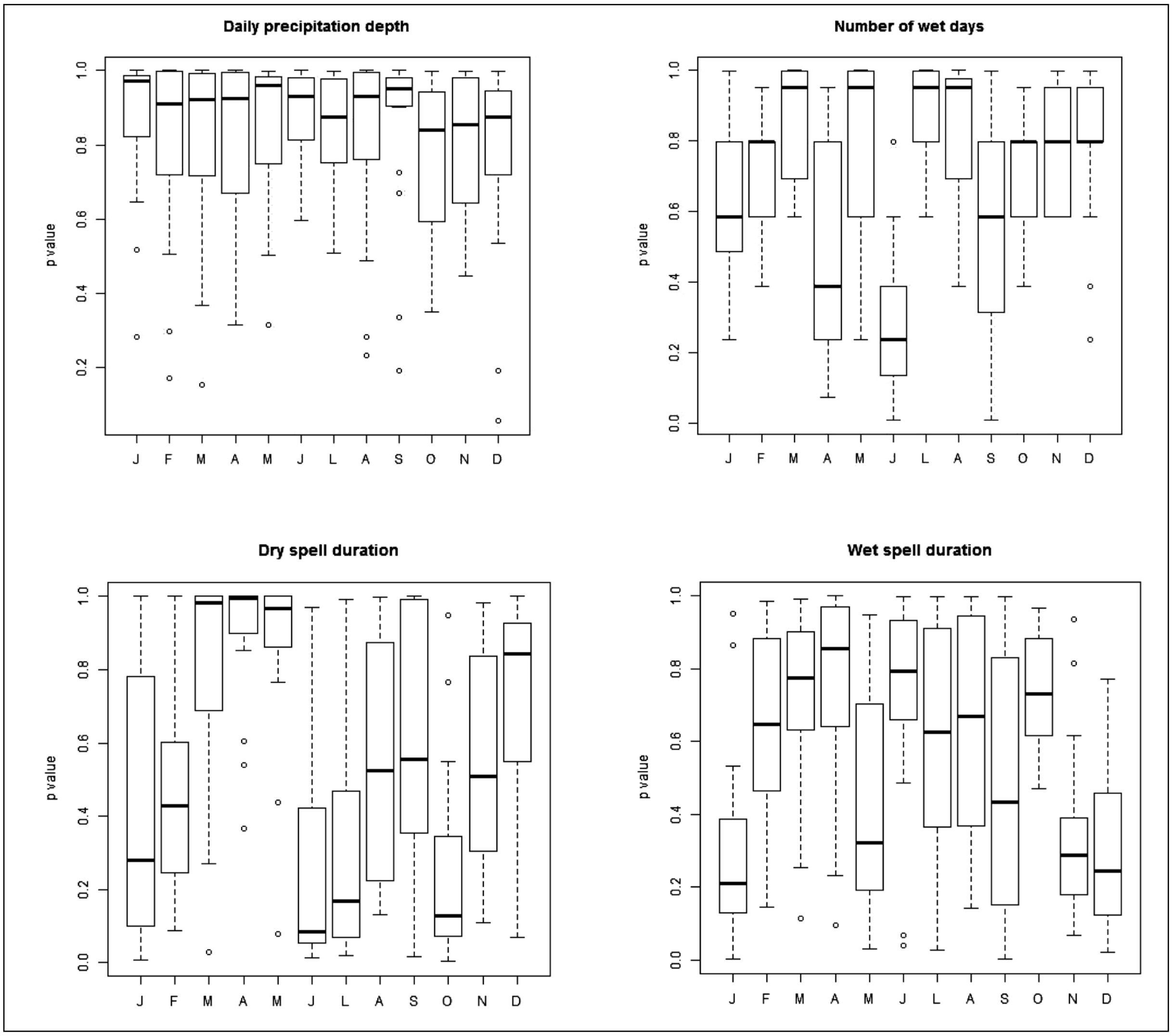

Fig. 5 - Monthly boxplots of fitting Kolmogorov – Smirnov's tests (generated vs. measured series).
Fig. 5 - Boxplots mensili dei test di adattamento di Kolmogorov – Smirnov (serie generate contro misurate).

package for a single site and following Wilks' approach for spatial (inter-station) correlations. After generation of precipitation occurrences, Gaussianized precipitation is calculated through a simple linear regression of occurrences making use of the observed frequency distribution with the addition of a random white noise. Finally, precipitation depth is obtained through an inverse Gaussianization with the use of the monthly non-parametric distribution from the observed samples.

3.3.2 Results: model validation

The following quantities are considered at each station in the different months and seasons of the year:

- precipitation depth (Fig. 6);
- dry spells (Fig. 7);
- monthly number of days with precipitation for each station (Fig. 8);
- probability that no daily precipitation occurs at each station or at each pair of stations (Mhanna and Bauwens, 2011):

$$\left[p00_{k,m}\right]_l = P\left[(X_k)_t = 0 \cap P(X_m)_{t+1} = 0\right] \qquad (8)$$

where k and m are two generic station indicators, t is the day indicator, l is a time lag, $P[(X_k)_t = 0]$ is the probability that no precipitation occurs at station k, $[p00_{k,m}]_l$ is the probability that no precipitation occurs at both station k and m after a time lag of l days (Fig. 9a);

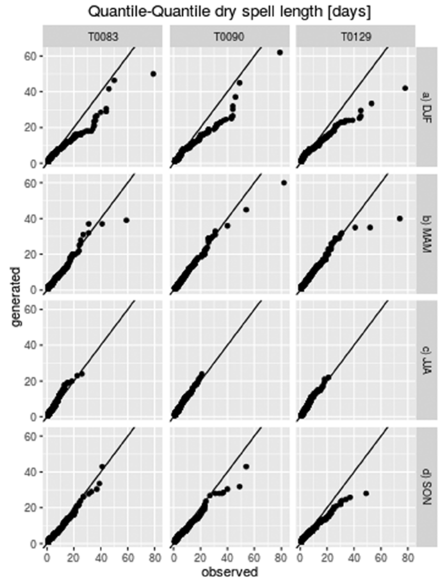

Fig. 7 - Q-Q plots for dry spell length [days] (generated vs observed) at stations T0083, T0090 and T0129 in the four seasons.
Fig. 7 - Diagrammi Q-Q per la durata dei periodi asciutti [giorni] (generate contro osservate) per le stazioni T0083, T0090 e T0129 nelle quattro stagioni.

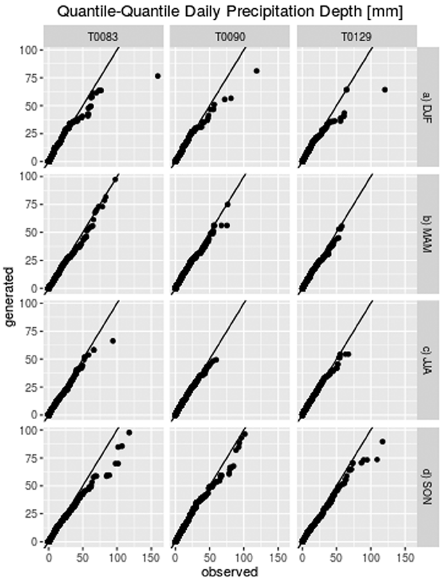

Fig. 6 - Q-Q plots for daily precipitation depth [mm] (generated vs observed) at stations T0083, T0090 and T0129 in the four seasons.
Fig. 6 - Diagrammi Q-Q per l'altezza di precipitazione [mm] (generate contro osservata) per le stazioni T0083, T0090 e T0129 nelle quattro stagioni.

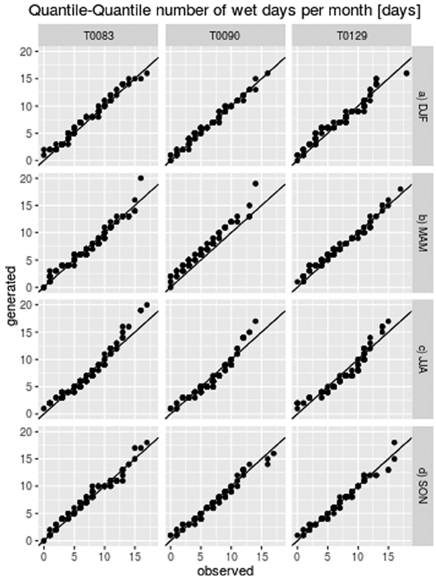

Fig. 8 - Q-Q plots for number of "wet" days per month (generated vs observed) at stations T0083,T0090 and T0129 in the four seasons.
Fig. 8 - Diagrammi Q-Q per il numero di giorni piovosi (generati contro osservati) per le stazioni T0083, T0090 e T0129 nelle quattro stagioni.

- probability that daily precipitation occurs at each station or at each pair of stations (Mhanna and Bauwens, 2011):

$$\left[p11_{k,m}\right]_l = P\left[(X_k)_t = 1 \cap P(X_m)_{t+1} = 1\right] \tag{9}$$

where, similarly, $P[(X_k)_t = 1]$ is the probability that precipitation occurs at station k, $[p11_{k,m}]_l$ is the probability that precipitation occurs at both stations k and m after a time lag of l days (Fig. 9b);

- continuity ratio (Wilks, 1998):

$$C_{k,m} = \frac{E\left\{(h_k)_t \mid \left[(h_k)_t > 0 \cap (h_m)_t = 0\right]\right\}}{E\left\{(h_k)_t \mid \left[(h_k)_t > 0 \cap (h_m)_t > 0\right]\right\}} \tag{10}$$

where $C_{k,m}$ is the ratio between the expected value of precipitation amount at the station k $(h_k)_t$, conditioned to the precipitation absence at station m on the same day, and the same expected value of precipitation amount at station k, but conditioned to precipitation occurrence at station m (Fig. 10).

The first three items are related to the validation of the single-site generation, while the others refer to the comprehensive multi-site generation. The model skill is assessed by testing the goodness of distribution fitting of the generated series to the measured ones by means of Kolmogorov – Smirnov (ks) tests. In order to represent an aggregated set of results, the p-values of ks tests are given as monthly boxplots for four variables: daily precipitation depth, number of days with precipitation, dry spell duration, and wet spell duration (Fig. 5). All values exceeding the significance level chosen as a benchmark (e.g. 0.05) assess that the hypothesis that the two samples come from the same distribution cannot be excluded. For every one of the four variables, every element of the statistics (e.g., one of the outliers in the boxplot) represents one station for the specific month. It can be seen that the cases which fall below the acceptance level are very few.

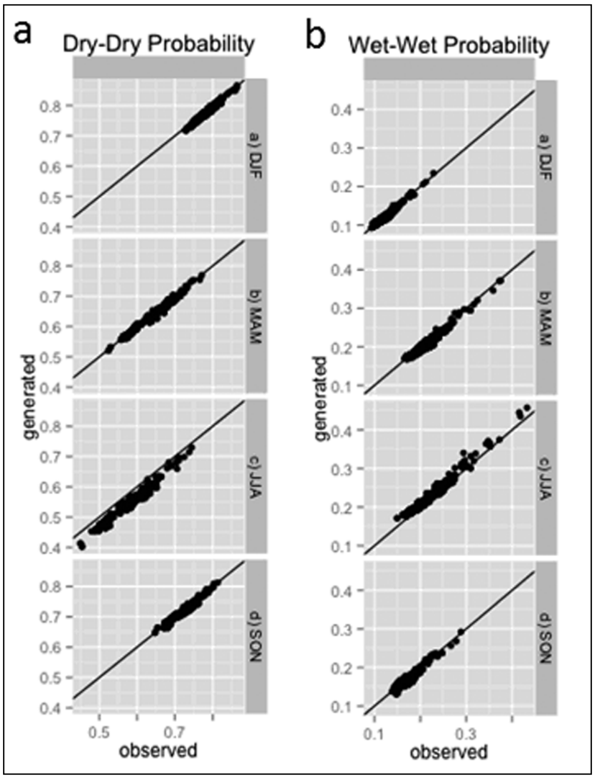

Fig. 9 - Probability that no precipitation occurs (a) or precipitation occurs (b) in each couple of stations (generated vs observed) at stations T0083, T0090 and T0129 in the four seasons.

Fig. 9 - Probabilità di non occorrenza di precipitazioni (a) o di occorrenza di precipitazioni (b) in ogni coppia di stazioni (generate contro osservate) alle stazioni T0083, T0090 e T0129 nelle quattro stagioni.

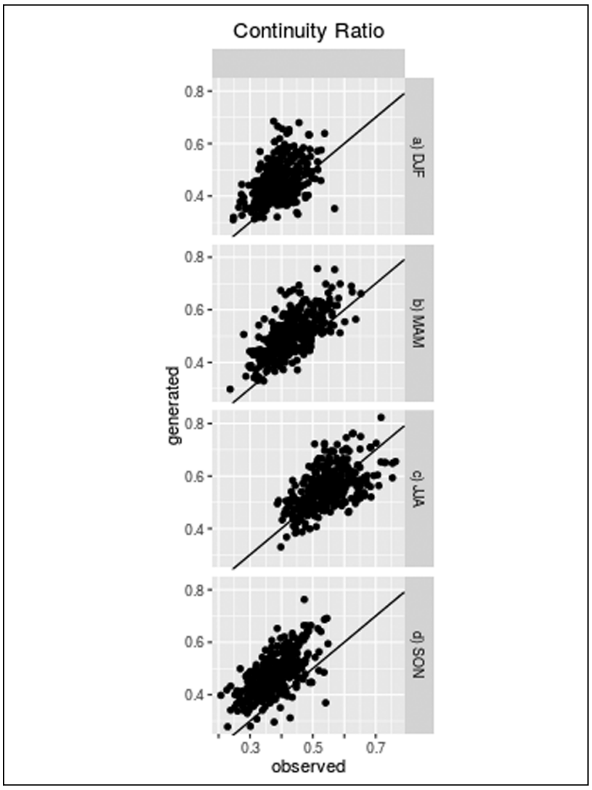

Fig. 10 - Continuity Ratio for each couple of stations (generated vs observed) at T0083, T0090 and T0129 in the four seasons.

Fig. 10 - Rapporto di continuità per ogni coppia di stazioni (generati contro osservati) alle stazioni T0083, T0090 e T0129 nelle quattro stagioni.

Results can be visually assessed by comparing observed and simulated values for each index. In this case, we chose three stations chosen at random: Cles (T0083, 652 m); Mezzolombardo

(T0090, 204 m); Trento Laste (T0129, 312 m) – seep map at Fig. 2.

Fig. 6 illustrates the seasonal comparison of observed and generated distribution probability through a Q-Q plot representation for three rain gauges (T0083,T0090 and T0129). Most of the points lie on the bisect showing a good agreement between observed and generated distributions. An outlier occurs in DJF, where the maximum generated value is underestimated compared to observations. This is a consequence of having chosen the sample-based probability distribution.

The length of the dry spells within each season is represented in Fig. 7 for the same stations of Fig. 6. Most dry spells are no longer than 20 days and are well modelled by the precipitation generator. However, some biases occur in winter - and, to a lesser extent, in autumn - where dry spell durations are underestimated, whereas the fit is good in spring and very good in summer, when drought problems can be far more significant. Similarly, the number of monthly days with precipitation is represented in Fig. 8.

A comparison of spatial coherence between observed and generated series is given by Fig. 9 and Fig. 10. In particular, Fig. 9 shows the scatterplot (generated vs observed) of the probability that precipitation simultaneously occurs or not for each couple of stations [$p00_{k,m}$] or [$p11_{k,m}$], respectively, for the three stations T0083, T0090, and T0129 – seep map at Fig. 2. The fit is quite satisfactory; this means that generated occurrence frequencies and correlations are well reproduced. However, in JJA the probability [$p00_{k,m}$] is slightly underestimated, but this is coherent with the bias shown in the Q-Q plots (Fig. 8) of the number of wet days for JJA. To test the dependence between precipitation amount and occurrence, the observed and generated continuity ratio values (Wilks, 1998, Brissette *et al.*, 2007, Zheng *et al.*, 2010, Breinl *et al.*, 2013) are represented in Fig. 10. The continuity ratio ranges from 0.2 to 0.6 in DJF and SON, in MAM and JJA may have higher values, reaching about 0.8 in JJA.

The continuity ratio is a measure of spatial intermittency and results show a good dependence between precipitation amount and occurrence among the stations of the dataset; in particular, this metric is a ratio between two conditional expected values of precipitation at a site as a function of precipitation occurrence at another site; it is strongly affected by the choice of precipitation probability distribution and some errors could cause significant errors in its estimation. Nevertheless, similarly to the findings of Wilks (1998) and Brissette *et al.* (2007), the application of RGENERATEPREC produces continuity ratios mostly centred around the bisect of the plot, showing a good correlation between precipitation amount and occurrence among the stations of the dataset.

4. CONCLUSIONS

The goal of this paper was to present new algorithms for stochastic generation of daily temperature and precipitation fields, implemented within a flexible, open source, analytical and statistical environment like R, in the form of new libraries. Software packages have been made freely available on CRAN repository (R Core Team, 2014). RMAWGEN carries out generations of weather series through Vector Auto-regressive Models; the latter work generally well for time-continuous variables, but present some critical issues for intermittent weather variables like precipitation. The problem can be tackled by considering the additional package RGENERATEPREC, developed to extend the methods applied in RMAWGEN to other models, where stochastic weather generations follow different algorithms.

The use of PCA Gaussianization is implemented to preserve the complete dependence structure, not only the correlation. It is reminded that the marginal Gaussianization of observed weather variables at a monthly scale means that RMAWGEN approach is independent of the statistical distribution of variables, and random weather time series can be regenerated with the same empirical probability distribution of the observed ones or with another distribution, assigned a priori.

RMAWGEN and RGENERATEPREC can be used separately or also jointly, if the conservation of the statistical links between their output is wanted. This is the case of an eventual use of the generated fields for model applications that require the spatial consistence of both temperature and precipitation fields, useful, for example, when generated series have to be spatially interpolated (to avoid inter-station meteorological inconsistencies in the same area), but also when the physical (internal) consistency in both temperature and precipitation is important for a station, e.g. for water balance models or for agro-ecological simulations, which may require the simultaneous occurrence of particular conditions of rainfall and temperature.

ACKNOWLEDGMENTS

This study was carried out in the frame of the projects ACE-SAP, ENVIROCHANGE and INDICLIMA, co-funded by the Autonomous

Province of Trento (Italy). The authors thank Dr. Fabio Zottele (Fondazione Edmund Mach) for his help and suggestions.

REFERENCES

Adenomon M., Oyejola B., 2013. Impact of agriculture and industrialization on gdp in Nigeria: Evidence from var and svar models. International journal of Analysis and Applications 1(1), 40-78.

Akaike H., May 1981. Likelihood of a model and information criteria. Journal of Econometrics 16(1), 3-14.

Bàrdossy A., Pegram G., 2009. Copula based multisite model for daily precipitation simulation. Hydrol. Earth Syst. Sci. Discuss. 6(3), 4485-4534.

Bivand R. S., Pebesma E. J., Gomez-Rubio V., 2008. Applied spatial data analysis with R. Springer, NY.

Breinl K., Turkington T., Stowasser M., 2013. Stochastic generation of multi-site daily precipitation for applications in risk management. Journal of Hydrology 498(0), 23-35.

Brissette F. P., Khalili M., Leconte R., 2007. Efficient stochastic generation of multi-site synthetic precipitation data. Journal of Hydrology 345(3-4), 121-133.

Chambers J. M., Hastie T. J. (eds), 1992. Statistical models in S. Wadsworth & Brooks/Cole, Pacific Grove, California.

Chen J., Brissette F., Leconte R., 2012. Weagets – a matlab-based daily scale weather generator for generating precipitation and temperature. Procedia Environmental Sciences 13(0), 2222-2235, 18th Biennial ISEM Conference on Ecological Modelling for Global Change and Coupled Human and Natural System.

Chen J., Brissette F.P., 2014. Comparison of five stochastic weather generators in simulating daily precipitation and temperature for the loess plateau of china. International Journal of Climatology, 34(10):3089-3105.

Chen S., Gopinath R., 2000. Gaussianization. In: Proc. NIPS'00, Denver, 2000:423-429.

Cordano E., 2015a. RGENERATE: Tools To Generate Vector Time Series. R package version 1.3.

Cordano E., 2015b. RGENERATEPREC: Tools To Generate Daily-Precipitation Time Series. R package version 1.0. URL http://CRAN.R-project.org/package=RGENERATEPREC

Cordano E., Eccel E., 2011. RMAWGEN: Multi-site Auto-regressive Weather GENerator. R package version 1.3.0.

Di Piazza A., Eccel E., 2012: Analisi di serie giornaliere di temperatura e precipitazione in Trentino nel periodo 1958-2010 (in Italian). Trento, 88 pp.

Eccel E., Cau P., Ranzi R., 2012. Data reconstruction and homogenization for reducing uncertainties in high-resolution climate analysis in alpine regions. Theoretical and Applied Climatology, 110(3):345-358. DOI 10.1007/s00704-012-0624-z

Erdogmus D., Jenssen R., Rao Y., Principe J., 2006. Gaussianization: An efficient multivarate density estimation technique for statistical signal processing. The Journal of VLSI Signal Processing 45, 67-83, 10.1007/s11265-006-9772-7.

Furrer E.M., Katz R.W., 2008. Improving the simulation of extreme precipitation events by stochastic weather generators. Water Resources Research 44(12), W12439.

Hamilton J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, New Jersey, United States.

Hannan E., Quinn B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B (Methodological) 41(2), 190-195.

Kahle D., Wickham H., 2013. ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3. URL http://.R-project.org/package=ggmap

Khalili M., Brissette F., Leconte R., 2009. Stochastic multi-site generation of daily weather data. Stochastic Environmental Research and Risk Assessment 23, 837-849, 10.1007/s00477-008-0275-x.

Kleiber W., Katz R.W., Rajagopalan B., 2012. Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. 2012(48), W01523 Water Resources Research. doi: 10.1029/2011WR011105.

Kleiber W., Katz R.W., Rajagopalan B., 2013. Daily minimum and maximum temperature simulation over complex terrain. The Annals of Applied Statistics, 7(1):588-612 doi: 10.1214/12-AOAS602.

Kojadinovic I., Yan J., 2010. Modeling multivariate distributions with continuous margins using the copula R package. Journal of Statistical Software 34(9), 1-20. URL

Laparra V., Camps-Valls G., Malo J., 2011. Iterative gaussianization: from ICA to random rotations. IEEE Transactions on Neural Networks.

Laparra V., Muñoz-Marí J., Camps-Valls G., Malo J., 2009. Pca gaussianization for one-class remote sensing image classification. URL: http://www.uv.es/lapeva/papers/SPIE09_one_class.pdf

Li C., Singh V.P., Mishra A.K., 2012. Simulation of the entire range of daily precipitation using a hybrid probability distribution. Water Resources Research 48(3), W03521

Liu J., Williams J.R., Wang X., Yang H., 2009. Using modawec to generate daily weather data for the epic model. Environmental Modelling & Software 24(5), 655-664.

Loecher M., Berlin School of Economics and Law, 2012. RgoogleMaps: Overlays on Google map tiles in R. R package version 1.2.0. URL http://.R-project.org/package=RgoogleMaps

Lütkepohl H., 2007. New Introduction to Multiple Time Series Analysis, 2nd Edition. Springer-Verlag, Berlin Hedelberg, Germany.

Luguterah A., Nasiru S., Anzagra L., 2013. Dynamic relationship between production growth rates of three major cereals in Ghana. Mathematical Theory and Modeling 3(8), 68-75.

Mahdi E., Ian McLeod A., 2012. Improved multivariate portmanteau test. Journal of Time Series Analysis 33(2), 211-222.

McCullagh P., Nelder J.A., 1992. Generalized Linear Models. Chapman and Hall, London.

Mearns L., Easterling W., Hays C., Marx D., 2001. Comparison of agricultural impacts of climate change calculated from high and low resolution climate change scenarios: Part i. the uncertainty due to spatial scale. Climatic Change 51, 131-172, DOI: 10.1023/A:1012297314857.

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. A stochastic space-time model for the generation of daily rainfall in the gaza strip. International Journal of Climatology, 32(7):1098-1112.

Parlange M.B., Katz R.W., 2000. An extended version of the Richardson model for simulating daily weather variables. J. Appl. Meteor. 39(5), 610-622, doi: 10.1175/1520-0450-39.5.610.

Peña D., Rodríguez J., 2002. A powerful portmanteau test of lack of fit for time series. Journal of the American Statistical Association 97(458), 601-610.

Pfaff B., 2008a. Analysis of Integrated and Cointegrated Time Series with R, 2nd Edition. Springer, New York, ISBN 0-387-27960-1.

Pfaff B., 2008b. Var, svar and svec models: Implementation within R package vars. Journal of Statistical Software 27(4):1-32.

Qian B., Corte-Real J., Xu H., 2002. Multisite stochastic weather models for impact studies. International Journal of Climatology 22(11), 1377-1397.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Racsko P., Szeidl L., Semenov M., 1991. A serial approach to local stochastic weather models. Ecological Modelling 57(1-2), 27-41.

Richardson C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. Wat. Resour. Res. 17, 182-190.

Rocca A., Bashanova O., Ginaldi F., Danuso F., 2012. Implementation and validation of Climak 3 weather generator. Italian Journal of Agrometeorology, 17(2):23-36.

Schwarz G., 1978. Estimating the dimension of a model. Ann. Statist. 6(2), 461-464.

Semenov M., Stratonovitch P., 2010. Use of multi-model ensembles from global climate models for assessment of climate change impacts. Climate Research 41(1), 1-14.

Semenov M.A., Barrow E.M., 1997. Use of a stochastic weather generator in the development of climate change scenarios. Climatic Change 35, 397-414, doi: 10.1023/A:1005342632279.

Serinaldi F., 2009. A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. J. Geophys. Res. 114(D10), D10103.

Shahin M.A., Ali M.A., Ali A.S., 2014. Vector autoregression (var) modeling and forecasting of temperature, humidity, and cloud coverage. In: Computational Intelligence Techniques in Earth and Environmental Sciences. Springer, pp. 29-51.

Thompson C.S., Thomson P.J., Zheng X., 2007. Fitting a multisite daily rainfall model to New Zealand data. Journal of Hydrology 340(1-2), 25-39.

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.

Yan J., 2007. Enjoy the joy of copulas: With a package copula. Journal of Statistical Software 21(4), 1-21.

Zheng X., Renwick J., Clark A., 2010. Simulation of multisite precipitation using an extended chain-dependent process. Water Resources Research 46: W01504.