Non-Gaussian spatiotemporal simulation of multisite daily precipitation: downscaling framework

M. A. Ben Alaya[1], T.B.M.J. Ouarda[3, 2] and F. Chebana[2]

[1]*Pacific Climate Impacts Consortium, University of Victoria,*
PO Box 1700 Stn CSC, Victoria, BC V8W2Y2, Canada

[2]*INRS-ETE, 490 rue de la Couronne, Québec (QC),*
Canada G1K 9A9

[3]*Masdar Institute of science and technology*
P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** mohamedalibenalaya@uvic.ca

2017-01-23

Abstract:

Probabilistic regression approaches for downscaling daily precipitation are very useful. They provide the whole conditional distribution at each forecast step to better represent the temporal variability. The question addressed in this paper is: How to simulate spatiotemporal characteristics of multisite daily precipitation from probabilistic regression models? Recent publications point out the complexity of multisite properties of daily precipitation and highlight the need for using a non-Gaussian flexible tool. This work proposes a reasonable compromise between simplicity and flexibility avoiding model misspecification. A suitable nonparametric bootstrapping (NB) technique is adopted. A downscaling model which merges a vector generalized linear model (VGLM as a probabilistic regression tool) and the proposed bootstrapping technique is introduced to simulate realistic multisite precipitation series. The model is applied to data sets from the southern part of the province of Quebec, Canada. It is shown that the model is capable of reproducing both at-site properties and the spatial structure of daily precipitations. Results indicate the superiority of the proposed NB technique, over a multivariate autoregressive Gaussian framework (i.e. Gaussian copula).

Keywords: Statistical downscaling, Vector generalized linear model, Multisite daily precipitation, Copula, Multivariate autoregressive Gaussian field, Binary entropy, Non parametric bootstrapping.

Introduction

Atmosphere–ocean general circulation models (AOGCMs) are very useful for assessing the evolution of the earth's climate system. However, the spatial resolution of AOGCMs is too coarse for regional and local climate studies. The above limitation has led to the development of downscaling techniques. These techniques include dynamical downscaling which includes a set of physically based limited area models (Eum et al. 2012), and statistical downscaling which identifies a statistical link between large scale atmospheric variables (predictors) and local variables (predictands) (Benestad et al. 2008). Among a number of weather variables, precipitation poses the largest challenges from a downscaling perspective because of its spatio-temporal intermittence, its highly skewed distribution and its complex stochastic dependencies. In several hydro-climatic studies, precipitation is shown to be the most dominating weather variable to explicitly affect water resources systems, since it plays an important role in the dynamics of the hydrological cycle. Precipitation data is generally collected at various sites, and downscaling techniques are required to adequately reproduce the observed temporal variability and to maintain the consistency of the spatiotemporal properties of precipitation at several sites. Properly reproducing the temporal variability in downscaling applications is very important in order to adequately represent extreme events. Furthermore, maintaining realistic relationships between multisite precipitations is particularly important for a number of applications such as hydrological modelling. Indeed streamflows depend strongly on the spatial distribution of precipitation in a watershed (Lindström et al. 1997).

Several statistical downscaling techniques have been developed in the literature. These methods can be divided into three main approaches: stochastic weather generators (Wilks and Wilby 1999), weather typing (Conway et al. 1996) and regression methods (Hessami et al. 2008, Jeong et al. 2012). Classical regression methods are commonly used because of their ease of implementation and their low computational requirement but they have several inadequacies. First and most importantly, they generally provide only the mean or the central part of the predictands and thus they underrepresent the temporal variability (Cawley et al. 2007). Second, they do not adequately reproduce various aspects of the spatial and temporal dependence of the variables (Harpham and Wilby 2005).

In this regard, probabilistic regression approaches have provided useful contributions in downscaling applications to accurately reproduce the observed temporal variability. Probabilistic regression approaches include: the Bayesian formulation (Fasbender and Ouarda 2010), quantile regression (Bremnes 2004, Friederichs and Hense 2007, Cannon 2011) and regression models where outputs are parameters of the conditional distribution such us the vector form of the generalized linear model (VGLM), the vector form of the generalized additive model (VGAM) (Yee and Wild 1996, Yee and Stephenson 2007) and the conditional density estimation network (CDEN) (Williams 1998, Li et al. 2013). Probabilistic regression approaches enable the definition of a complete dynamic univariate distribution function. In the case of VGLM, VGAM and CDEN, the output of the model is a vector of parameters of a distribution which depends on the predictor values. In addition to the location parameter (namely the mean), the scale and shape parameters can vary according to the updated values of atmospheric predictors and thus allowing for a better control and fit of the dispersion, skewness and kurtosis. Therefore, simulation of downscaled time series with a realistic temporal variability is achieved by drawing random numbers from the modeled conditional distribution at each forecast step (Williams 1998, Haylock et al. 2006). In this respect, the problem that arises is how to extend probabilistic regression approaches to multisite downscaling tasks.

Operationally, the multi-site replicates of the field predictands are readily obtained in the simulation stage. Generally, generating from a probabilistic regression model can be achieved by drawing random numbers from the uniform distribution and then applying the inverse cumulative distribution function of the parent distribution obtained from the probabilistic regression model. We must keep in mind that, the parameters of the parent distribution change at each forecast step based on the updated values of large-scale atmospheric predictors. To obtain spatially correlated simulations using probabilistic regression models, we need to simulate uniform random variables that are correlated. Thus, generating from a multivariate distribution on the unit cube (i.e, with uniform margins) could solve the issue. Such a multivariate distribution is called a copula. Copula functions allow describing the dependence structure independently from the marginal distributions and thus, using different marginal distributions at the same time without any transformations. During the last decade, the application of copulas in hydrology and climatology has grown rapidly. An introduction to the copula theory is provided in Joe (1997) and Nelsen (2013). The reader is directed to Genest and Chebana (2015) and Salvadori and De Michele (2007) for a detailed review of the development and applications of copulas in hydrology including frequency analysis, simulation, and geostatistical interpolation (Bárdossy and Li 2008, Chebana and Ouarda 2011, Requena et al. 2015, Zhang et al. 2015). In recent years, copula functions have been widely used to describe the dependence structure of climate variables and extremes (AghaKouchak 2014, Guerfi et al. 2015, Hobæk Haff et al. 2015, Mao et al. 2015, Vernieuwe et al. 2015).

To extend the probabilistic regression approach to multisite and multivariable downscaling, Ben Alaya et al. (2014) proposed a Gaussian copula procedure. Nevertheless, this approach does not take into account cross-correlations lagged in time and thus it cannot reproduce the short term autocorrelation properties of downscaled series such us the lag-1 cross-correlation. To solve this issue Ben Alaya et al. (2015) employed a multivariate autoregressive field as an extension to the Gaussian copula to account for the lag-1 cross-correlation. On the other hand, a careful examination of the dependence structure in hydrometeorological processes using copula reveals that the meta-Gaussian framework is very restrictive and cannot account for features like asymmetry and heavy tails and thus cannot realistically simulate the multisite dependency structure of daily precipitation (El Adlouni et al. 2008, Bárdossy and Pegram 2009, Lee et al. 2013).

To exploit this knowledge for precipitation simulation, Li et al. (2013) and Serinaldi (2009) considered copulas to introduce non-Gaussian temporal structures at a single site. Bargaoui and Bárdossy (2015) employed a bivariate copula to model short duration extreme precipitation. For multisite precipitation simulation, Bárdossy and Pegram (2009) and AghaKouchak et al. (2010) introduced non-Gaussian spatial tail dependency structures by simulating precipitation from a v-transformed normal copula proposed by Bárdossy (2006). Other theoretical models of copula can also be used to reproduce this spatial tail dependency such as metaelliptical copulas (Fang et al. 2002) or using vine copula (Gräler 2014).

In the case of precipitation simulation it would be useful to implement a spatiotemporal flexible copula that allows simultaneously modelling both temporal and spatial dependency. To our best knowledge, such a copula has not been exploited in the hydrometeorological literature including for downscaling, except for the multivariate autoregressive meta-Gaussian copula. Nevertheless, in the statistical literature Smith (2014) employed a vine copula to achieve this end. In the last decade, vine copulas emerged as a new efficient technique in econometrics. Vine copula use pair copula building blocks offering a flexible way to capture the inherent dependency patterns of high dimensional data sets, with regard to their symmetries, strength of dependence and tail dependency. On the other hand, the full specification of a vine copula model is not straightforward, since it requires the choice of a tree structure of the vine copula, the copula families for each pair copula term and their corresponding parameters (Czado et al. 2013). In addition, the application for spatial and temporal structure dependency greatly increases the number of parameters which would unquestionably make the model less parsimonious and increase the associated uncertainty.

In order, to avoid any model misspecification, information about the data dependence structure can be reproduced in the simulation step by resampling using the data ranks (Vinod and López-de-Lacalle 2009, Vaz de Melo Mendes and Leal 2010, Srivastav and Simonovic 2014). Indeed the data ranks are the statistics retaining the greatest amount of information about the data dependence structures (Oakes 1982, Genest and Plante 2003, Song and Singh 2010). In this respect, the aim of the present paper is to propose a new approach to maximize the amount of information about the dependence structure that is preserved in the simulation step from a probabilistic regression downscaling model. Hence, instead of using a flexible copula, a simple non-parametric bootstrapping technique is employed. The procedure consists in generating uniform random series between 0 and 1 and then sorting them according to their observed ranks. The resulting multisite precipitation downscaling model involves a new hybrid procedure merging a parametric probabilistic regression model (the VGLM) and a non-parametric bootstrapping (NB) technique. The introduced bootstrapping technique represents a fair compromise between simplicity and flexibility to generate realistic multisite properties of precipitation from a probabilistic regression model.

Since traditional multisite resampling techniques are closely related to observed data, they suffer from the inability to generate values that are more extreme than those observed. In this respect, the main advantage of the proposed non parametric resampling approach compared to traditional non-parametric techniques, is its ability to mimic only the observed ranks without affecting the univariate marginal properties. Indeed the proposed VGLM-NB model takes advantage of the probabilistic regression component to allow univariate margins to be dynamic and thus varying in the future according to the large scale atmospheric predictors. This attractive characteristic helps to preserve the dependence structure without tying the simulations too close to observed data.

The paper is structured as follows: after this introduction, the proposed hybrid multisite VGLM-NB model is described. An application to a case study of daily data sets from the province of Quebec is carried out. The model validation is done using statistical characteristics such as mean, standard deviation, dependence structure (both spatial and temporal), precipitation indices and an entropy-based congregation measure. Obtained results are compared to those corresponding to a VGLM-MAR which is a VGLM combined with multivariate autoregressive (MAR) Gaussian field. Finally discussions and conclusions are given.

2. Study area and data

Observed daily precipitations from nine Environment Canada weather stations located in the province of Quebec (Canada) are used in this study (see Figure 1). The list of stations is presented in Table 1. Predictor variables are obtained from the reanalysis product NCEP/NCAR interpolated on the CGCM3 Gaussian grid (3.75 ° latitude and longitude). Six grids covering the predictand stations area are selected (see Figure 1), and 25 NCEP predictors are available for each grid (see Table 2). Thus, a total of 150 daily predictors are available for the downscaling process. To reduce the number of predictors, a principal component analysis (PCA) is performed. The first principal components that preserve more than 97% of the total variance are selected. The data sets cover the period between January, 1st 1961 and December, 31st 2000. This record period is divided into two periods for the calibration (1961-1980) and the validation (1981-2000).

3. Methodology

In this section, the proposed VGLM-NB model is presented. The corresponding probabilistic framework is presented with a description of the conditional Bernoulli-Generalized Pareto regression model and the proposed nonparametric bootstrapping technique.

3.1. Vector generalized linear model

The precipitation amount distribution, at a daily time scale, tends to be strongly skewed, and is commonly assumed to be gamma distributed (Stephenson et al. 1999, Giorgi et al. 2001, Yang et al. 2005). In a regression perspective, the generalized linear model (GLM) extends classical regression to handle the normality assumption of the model output. Here the output may follow a range of distributions that allow the variance to depend on the mean such us the exponential distribution family and particularly the Gamma distribution (Coe and Stern 1982, Stern and Coe 1984, Chandler and Wheater 2002). Nevertheless, recent findings suggest that the gamma distribution can be unsuitable for modeling precipitation extremes since it is very restrictive and cannot account for features like heavy tails. Therefore, to treat this issue, other options have been proposed in the literature, particularly the generalized Pareto (GP) and the reverse Weibull (WEI) distributions (Ashkar and Ouarda 1996, Serinaldi and Kilsby 2014). However, due to the fact that the variance does not depend on the mean, these two distributions cannot be used in a GLM. Vector generalized linear models (VGLMs) have been developed to handle this inadequacy (Yee and Stephenson 2007). Instead of the conditional mean only, VGLM provides the entire response distribution by employing a linear regression model where the outputs are vectors of parameters of the selected conditional distribution (Kleiber et al. 2012). Moreover, in downscaling applications, VGLM has a particular advantage since it allows reproducing a realistic temporal variability of the downscaled results by drawing values from the obtained conditional distribution at each forecast step.

The structure of the proposed model allows considering a suitable distribution for each station. Among several options proposed in the literature, Gamma, mixed exponential, GP and reverse WEI are the most commonly used and are therefore considered in the current work to represent the precipitation amount on wet days (days with positive values of precipitation amounts, when precipitation falls). However, for the sake of simplicity, only one distribution that provides a good overall fit for all stations is selected. In our study, the examination of the Q-Q plots presented in Figure 2 reveals that all these distributions fit fairly well the precipitation amounts. However, the GP distribution is chosen since it is more successful in reproducing the upper tails. The expression of the zero adjusted GP distribution is given by:

$$f(y) = 1 - \left(1 + \beta \frac{y}{\alpha}\right)^{-1/\beta} \quad ; \quad y > 0 \tag{1}$$

where y is the precipitation amount, $\alpha\,(\alpha > 0)$ and β (where $1+\beta\, y/\alpha > 0$) are respectively the scale and the shape parameters of the zero-adjusted GP model.

Therefore, a mixed Bernoulli–GP distribution with a vector of parameters $p = (\rho, \alpha, \beta)$ is considered to represent the whole precipitation distribution that includes both occurrences and amounts in a single distribution. The vector of parameters includes the probability of precipitation ρ which is the parameter of the Bernoulli process, and the scale $\alpha\,(\alpha > 0)$ and shape β (where $1+\beta\, y/\alpha > 0$ and y represents the precipitation values) are parameters of the zero adjusted GP distribution. Hence, the proposed precipitation model can be considered as a mixture of Dirac mass on zero (representing the probability on zero) and GP distribution for precipitation amounts (representing positive values of precipitation amounts). Using the VGLM, these parameters are considered to vary for a given day t according to the value of large-scale atmospheric predictors $x(t)$. However, only the shape parameter β is fixed to guarantee the convergence of the maximum likelihood estimates. For the parameter of the probability of precipitation occurrences we adopt a logistic regression which is expressed as:

$$\rho(t) = \frac{1}{1+\exp\left[-a^T x(t)\right]} \tag{2}$$

where a is the coefficient of the logistic model. The scale parameters $\alpha(t)$ are modeled using an exponential link written as:

$$\alpha(t) = \exp\left[b^T x(t)\right] \tag{3}$$

where b is the coefficient of the model. Thus, the conditional Bernoulli-GP density function for the precipitation $y(t)$ on a day t is expressed as:

$$f_t[y(t) \mid x(t)] = \begin{cases} 1 - \dfrac{1}{1+\exp\left[-a^T x(t)\right]} & \text{if} \quad y(t) = 0 \\ \dfrac{1}{1+\exp\left[-a^T x(t)\right]} \times \left[1 - \left(1 + \beta \dfrac{y(t)}{\exp\left[b^T x(t)\right]}\right)^{-1/\beta}\right] & \text{if} \quad y(t) > 0 \end{cases} \tag{4}$$

The coefficients a, b and β are obtained following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Haylock et al. 2006, Cawley et al. 2007, Cannon 2008):

$$\mathscr{L} = \sum_{t=1}^{T} \log\left\{ f_t \left[y(t) \mid x(t) \right] \right\} \tag{5}$$

via the simplex search method of Lagarias et al. (1999). This is a direct search method that does not use numerical or analytic gradients.

Now, consider a calibration period of length T and precipitation series at several sites $j = 1, 2, \ldots, m$. While in the current case study $m = 9$ sites, the proposed methodology is very general and can also be conducted using large number of sites. The proposed VGLM regression can be trained separately for each precipitation variables y_j at the site j, and thus to obtain the estimated parameters $\hat{p}_j(t)$ and the conditional distributions $\hat{f}_{tj}(y_j \mid x(t))$ for each day $t = 1, 2, \ldots, T$. Figure 3a shows the steps involved for estimating the parameters of the VGLM models.

3.2. Non parametric bootstrapping technique

These dynamic marginal distributions obtained from the VGLM models can be coupled with a random field with uniform margins. Thus, in simulation, generation of the multi-site replicates of the precipitation field is readily achieved by generating properly associated multivariate variants between 0 and 1 with uniform margins, which are back-transformed to synthetic field predictands by applying the inverse cumulative distribution function. To address this point, hidden multivariate variants $u(t) = \left[u_1(t), \ldots, u_d(t) \right]$ uniformly distributed between 0 and 1 are extracted where $u_j(t)$ for $j = 1, \ldots, m$ are obtained from the following equation:

$$u_j(t) = \hat{F}_{tj}(y_j(t)) \tag{6}$$

where $\hat{F}_{tj}$ is the cumulative distribution function at time t for site j obtained from the VGLM model. Figure 3b shows the steps involved in obtaining the hidden multivariate variants over the calibration period. First, the VGLM can be evaluated during the calibration period separately for each station. This will allow obtaining the entire conditional distribution for each day from the calibration period. Then the obtained conditional CDFs can be applied to their corresponding predictand values to express precipitation as a probability of non-exceedances ranging from 0 to 1. In order to map $u_j(t)$ onto the full range of the uniform distribution between 0 and 1, the cumulative probabilities $F_{tj}(y_j(t))$ are randomly drawn from a uniform distribution on $[0, 1-\rho(t)]$ for dry days. The resulting data matrix $u(t)$ represents values between 0 and 1 that contain the unexplained information by the VGLM model including spatial dependence structures and long term and short term temporal structures.

The question that should be addressed in this step is: "how to extract information about the data dependence structure from the data matrix $u(t)$, and how to preserve this information in the simulation step?”. This information is contained in the ranks matrix **R** of the data matrix $u(t)$ (Oakes 1982, Genest and Plante 2003, Song and Singh 2010). Hence, if the ranks of the data matrix $u(t)$ are preserved in the simulation, the data dependence structure will be preserved as well. Recall that copula functions allow modelling the data ranks in order to model the data dependence structure. Thus, the rank matrix **R** can be modeled using a multivariate copula. In the case of precipitation simulation it would be useful to simulate from a flexible multivariate copula model that preserves both temporal and spatial dependence structures. However achieving such flexibility may require an increasing number of parameters which would makes the copula model less parsimonious and increases the associated uncertainty without ensuring that the ranks of the data will be preserved. In this respect, to avoid any model misspecification, the rank matrix $\mathbf{R}$ can be used in the simulation to preserve a great amount of information about the data dependence structure. The idea consists in generating multivariate random variables from the uniform distribution with the same dimension as the matrix $\mathbf{R}$, and then ordering each column according to the corresponding column in $\mathbf{R}$.

Finally the synthetic precipitation series during the validation period can be obtained from the VGLM-NB model using the following three steps.

(i) Randomly generate multivariate random variables from the uniform distribution with same dimension as the matrix $\mathbf{R}$ during the validation period.

(ii) Sort each column of the obtained matrix in step (i) according to the corresponding column in $\mathbf{R}$.

(iii) Apply the inverse cumulative Bernoulli-GP distribution expressed in Equation (3) for each site j and for each forecast day t from the validation period to the sorted matrix obtained in step (ii).

Let us now consider the univariate variant $u_j(t)$ at a site j and the same variant $u_j(t+h)$ lagged by h days. Since the rank column R_j on this site j is preserved, the ranks matrix $\mathbf{R}_j^h$ of the data matrix $[u_j(t), u_j(t+h)]$ will be preserved as well. This implies that the proposed approaches can be expected to preserve the temporal correlation at individual sites during the simulation. The proposed NB approach is similar to a copula, since both are based on the generation of uniformly distribution random variables that are correlated, except that copula allows modelling the ranks matrix whereas the proposed approach mimics the data ranks rather than modeling them.

As discussed by Serinaldi and Kilsby (2014), taking into account the spatial correlation and the short term autocorrelation in a probabilistic regression model can be introduced in two ways: (i) by introducing the precipitation at previous time steps as an additional covariate, or (ii) by using a random field with uniform marginals and a suitable spatio-temporal structure. The first way implies a sequential simulation; it can be used for cases involving a small number of sites (Serinaldi 2009, Kleiber et al. 2012). In the second way, multisite characteristics and temporal autocorrelation are introduced in the simulation stage using correlated random numbers with uniform marginal distributions. This second way is adopted in the current work. This technique avoids a sequential simulation conditioned on the simulation of the precipitation at the previous time steps and can be adapted for a large number of sites. In the proposed approach the probabilistic regression component uses a single discrete-continuous distribution and thus avoids the split between occurrence process (the transition between wet and dry days) and precipitation amount process (positive precipitation values in wet days). In this way, the number of the random field substrates to be used in the simulation stage is reduced from two (one for the occurrence process and one for the amount process) to one, thus making the model more parsimonious.

3.3. Quality assessment of downscaled precipitation

To assess the performance of the proposed VGLM-NB model, we compare it to VGLM-MAR which is a downscaling model using the same mixed Bernoulli-Generalized Pareto distribution and extended to multisite tasks using a first order multivariate autoregressive random field framework (Ben Alaya et al. 2015).

3.3.1. Quality assessment of univariate characteristics

Two approaches are considered for the quality assessment of univariate characteristics of the VGLM-NB model. The first approach is based on a direct comparison between the estimated and observed values using statistical criteria, while the second approach is based on calculating climate indices. In the two validation approaches, the VGLM-NB model results are compared to those obtained using the VGLM-MAR.

In the first validation approach, four statistical criteria are used for model validation. These criteria are:

$$ME = \frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right) \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right)^2} \tag{8}$$

$$D = \sigma^2(y_{obs}) - \sigma^2(y_{est}) \tag{9}$$

$$FAR = \frac{a}{b} \quad (10)$$

where n denotes the number of observations, y_{obs_t} refers to the observed value, y_{est_t} is the estimated value, t denotes the day, σ is the standard deviation, a the number of false alerts for observed dry days, and b is the total number of observed dry days.

The first criterion is the mean error (ME) which is a measure of accuracy. The second criterion is the root mean square error (RMSE) which is given by an inverse measure of the accuracy and must be minimized, and the third criterion D measures the difference between observed and modeled variances, this criterion evaluates the performance of the model in reproducing the observed variability. The last criterion, the false alarm rate (FAR), is the fraction of false alerts associated with observed dry days and must be minimized.

In a second validation approach, a set of several precipitation indices that reflect precipitation variability on a seasonal and monthly basis are considered. Five indices related to precipitation amounts are considered: the mean precipitation of wet days (MPWD), the 90th percentile of daily precipitation (Pmax90), the maximum 1-day precipitation (PX1D), the maximum 3-day precipitation (PX3D), and the maximum 5-day precipitation (PX5D). In addition, three other indices are considered for precipitation occurrences: the maximum number of consecutive wet days (WRUN), the maximum number of consecutive dry days (DRUN) and the number of wet days (NWD). All indices are calculated on a monthly time scale, whereas the P90max is calculated on a seasonal time scale.

3.3.2. Quality assessment of multisite characteristics

Multisite characteristics are verified using scatter plots of observed and modeled lag-0 and lag-1 cross-correlations and log odds ratios (LOR). Lag-0 cross correlations correspond to cross correlations between all pairs of data (not lagged in time) whereas Lag-1 cross correlations correspond to cross correlations between all pairs of data lagged by 1 day.

A log-odds ratio between a pair of stations *i* and *j* is expressed as:

$$LOR_{i,j} = \ln\left[\frac{p00_{i,j}\, p11_{i,j}}{p10_{i,j}\, p01_{i,j}}\right], \tag{11}$$

Where $p00_{i,j}, p11_{i,j}, p10_{i,j}, p01_{i,j}$ are the joint probabilities of no rain at either one of the two stations, rain at both stations, rain at station *i* and no rain at station *j*, and finally no rain at station *i* and rain at station *j*, respectively. The log odds ratio provides a measure of the spatial correlation between precipitation occurrences at each pair of stations where higher values indicate better defined spatial dependence (Mehrotra et al. 2004, Mehrotra and Sharma 2006).

The dynamics of flood events are strongly related to the simultaneous occurrence of extreme precipitation at several sites. A pairwise correlation is often used for the specification of multisite precipitation models (this is the case of the VGLM-MAR). On the other hand multisite properties of extreme precipitation could be related to higher-order correlations than a traditional pairwise correlation (Serinaldi et al. 2014). In this respect, a diagnostic based on higher order correlations between extreme precipitations is necessary but often ignored. To this end, Bárdossy and Pegram (2009) introduced the binary entropy as a measure of dependence in a given triplet. This measure overcomes a pairwise validation in order to look effectively at the high-order dependence properties. The entropy theory was first formulated by (Shannon 1948) to provide a measure of information contained in a set of data. To calculate the binary entropy, we first fix a given quantile threshold to divide each precipitation series into binary sets by allocating 0 to the lower partition defined by the threshold and 1 otherwise. At each day, the eight possible states of a given binary triple can be defined using the set $\{i, j, k\}$ for $i, j, k = 0,1$. Then, the eight binary probabilities $p(i, j, k)$, for $i, j, k = 0,1$ can be calculated over all days from the validation period. For example, $p(1,1,1)$ represents the probability that all three binary sets on a given day are simultaneously equal to 1, and $p(0,0,0)$ that they are all equal to 0. The binary entropy H can be computed as

$$H = -\sum_{i,j,k=0}^{1} p(i, j, k) \ln(p(i, j, k)). \tag{12}$$

Hence, the lower the entropy is, the stronger will be the association between the variables at a given threshold.

4. Results

The VGLM-NB model was trained for the calibration period (1960-1980), using precipitation data series from the nine stations and the 40 predictors obtained by the PCA. Once the parameters of the conditional Bernoulli-GA distribution ($\rho_j(t), \alpha_j(t)$ and $\beta_j(t)$) have been estimated for each day t and for each site j over the calibration period, all the obtained conditional marginal distributions were used to obtain the hidden variables $u(t)$

and then to calculate the rank data matrix **R**. Finally, for each of the nine sites, 1000 daily precipitations series were generated during the validation period (1981-2000) using VGLM-NB described in Section 3 and the VGLM-MAR for comparisons. We assume that 1000 simulations are sufficiently enough to provide stable estimates of precipitation characteristics. Figure 4 shows an example of one precipitation simulation obtained using the VGLM-NB model at Cedars station during the year 1981. Based on the simulated series, VGLM-NB seems to be able to preserve at site properties of the natural process of both precipitation amounts and precipitation occurrences.

For the evaluation of the univariate characteristics of VGLM-NB and VGLM-MAR using statistical criteria, the RMSE and ME where calculated using the conditional means of 1000 realisations, whereas the differences between observed and modeled variances where calculated using the mean variance values of the 1000 simulations. Table 3 shows values of the obtained criteria. Generally, the two compared models give similar results in terms of RMSE, ME and D. This result is expected since both VGLM-NB and VGLM-MAR have the same probabilistic regression component. For precipitation occurrences, in terms of FAR results show that VGLM-NB has fewer FAR over all stations. This result shows that, although both VGLM-NB and VGLM-MAR are trained using the same probabilistic regression component (the Bernoulli-generalized Pareto regression model), the non-parametric bootstrapping technique leads to better at-site results than the MAR approach. In addition, by the evaluation of univariate characteristics using precipitation indices, the RMSE values of these indices (presented in Table 4) show that VGLM-NB performs better than VGLM-MAR for all indices, except for the 90th percentile of daily precipitation. This result demonstrates that the VGLM-NB is more able to represent precipitation variability on a monthly basis than the VGLM-MAR. To evaluate the ability of both VGLM-NB and VGLM-MAR to simulate short term autocorrelation, Figure 5 shows observed and modeled lag-1 autocorrelation for precipitation series at the nine stations during the validation period. It can be seen from Figure 5 that VGLM-NB preserves more adequately the lag-1 autocorrelation at a single site.

To evaluate the ability of the models to simulate spatially realistic precipitation fields, Figure 6 compares the distribution of observed and downscaled daily average precipitations over the 9 stations for VGLM-NB, VGLM-MAR and univariate VGLM without multisite extension. The comparison with the univariate VGLM is beneficial to identify the real gain contributed by the two multisite components of VGLM-NB and VGLM-MAR. The observed and modeled CDFs are presented in Figure 6.a and the Q-Q plots for quantiles corresponding to non-exceeded probabilities ranging between 0.01 and 0.99 with a step of 0.01 in Figure 6.b. Results indicate that the performance of VGLM-NB in reproducing the distribution of daily average precipitation is satisfactory compared to VGLM and VGLM-MAR. Both VGLM and VGLM-MAR underestimate the higher precipitation amounts and overestimates the lower precipitation amounts. Although VGLM-NB slightly overestimates observed quantiles, it tends to fairly well reproduce low and high values. This overestimation may be explained by the fact that VGLM-NB supposes that the rank matrix of the variants $u(t)$ remain the same during the validation period.

Figure 7 shows scatterplots between observed and modeled lag-0 and lag-1 cross-correlations for all station pairs considering only wet days during the validation period. Lag-0 cross-correlation is presented in Figure 7.a and lag-1 cross-correlation in Figure

7.b. The correlation values for each model are obtained using the mean of the correlation values calculated from the 1000 realisations. For lag-0 cross-correlation, the points correspond to all 36 combinations of pairs of stations, while for lag-1 cross-correlation points correspond to all 81 combinations because lag-1 cross-correlations are generally not symmetric. Figure 7.a shows that observed values of lag-0 cross-correlation range between -0.02 and 0.65. VGLM-NB gives better preservation of lag-0 cross-correlation than both VGLM-MAR and traditional VGLM. Because VGLM is not a multisite model, it gives the poorest performances and generally underestimates lag-0 cross-correlations. Figure 5b indicates that, for the lag-1 cross-correlation, observed values range between -0.1 and 0.28. For VGLM-NB the performance in reproducing lag-1 cross correlation is less good than the on corresponding to lag-0 cross correlation. However, this performance seems to be always better than the two other models.

To further evaluate the multisite performance, Figure 8.a presents observed and modeled log odds ratios for the VGLM-NB, VGLM-MAR and univariate VGLM at all stations. Results indicate that the VGLM-NB model provides very close correspondence with observed log odds ratios and gives better results than the two other models. VGLM-MAR outperforms the univariate VGLM but its results are less accurate than VGLM-NB, especially when the observed correlations are high.

Figure 9 shows scatter plots of observed and modeled binary entropy for precipitation occurrences (Figure 9a) and at three quantile thresholds: 0.75 (Figure 9.b), 0.90 (Figure 9.c) and 0.975 (Figure 9.d). Points correspond to all combinations of stations triplets. It can be seen from Figure 9.a that simulated precipitation occurrences using both VGLM and VGLM-MAR data exhibit higher binary entropy values than observed data. Similar results were found for binary entropy corresponding to the quantile thresholds 0.75, 0.90 and 0.95. This result indicates that the Gaussian dependence structure is not enough to capture the stronger association of extreme precipitation. It is clear that the VGLM-NB is closer to the data across the range of the binary entropy *H* than the VGLM-MAR model, indicating that non-parametric bootstrapping simulation is an improvement over the multivariate autoregressive Gaussian framework. In reality, this result is expected, since the VGLM-MAR captures the spatial structure by modeling a combination of bivariate relationships using the Gaussian copula. Improving the capture of spatial structure using parametric models requires the application of high-dimensional copulas such us a vine copula.

5. Discussions

Unlike the VGLM-MAR, an attractive characteristic of the proposed VGLM-NB is that pairwise correlations are not used for the model definition. Indeed, the employed non-parametric bootstrapping technique does not model dependency structures but mimics the observed data ranks to preserve the unexplained multisite properties by the VGLM. As it is the case for most resampling methods (Ouarda et al. 1997, Buishand and Brandsma 1999, Buishand and Brandsma 2001, Mehrotra and Sharma 2009, Lee et al. 2012), this approach is data driven, non-parametric and thus avoiding any model misspecification when preserving multisite properties. However, while resampling models suffer from the inability to generate values that are more extreme than those observed, the probabilistic regression component of the proposed hybrid model allows overcoming this drawback. Indeed, regression methods and resampling techniques can be combined to take advantage of their strengths for downscaling tasks. For this purpose, a widely used approach consists in using resampling or randomisation methods to address the inability of the traditional regression component to preserve the temporal variability and multisite properties (Jeong et al. 2012, Jeong et al. 2013, Khalili et al. 2013). These hybrid approaches are based on a static noise observed during the calibration of the regression component. Therefore, the part of the variability which is explained by the randomization component does not depend on the predictors, and thus, it is supposed to be constant in a changing climate. For this reason, this traditional hybrid structure may not represent local change in the temporal variability in a climate change simulation. Hence, the hybrid structure employed here to describe the VGLM-NB (as well as the VGLM-MAR), allows the temporal variability to be reproduced in the regression component (using the VGLM component) and thus it may change in the future according to the large scale atmospheric predictors.

Although the proposed non parametric approach allows preserving the multisite dependence structure at gauged sites, this dependence structure is still unknown. In regionalization applications where simulations at ungaged locations are required it is imperative to know the structure of the spatial dependence. In such a situation, a spatial model is required and thus modelling the data ranks through copulas would be more advantageous. Another limitation of the proposed approach is that the data rank matrix of the hidden variants $u(t)$ is supposed to be the same (i.e. stationary) in the future. In this respect, allowing the dependence to be dynamic requires also a parametric modelling.

It should be mentioned that a very important point that has not been considered in this work is the selection of predictor variables. The selection of predictor variables in the development of statistical downscaling models requires comprehensive considerations. In the case of precipitation, the best description of the conditional distribution may require the use of different subsets of predictor variables for precipitation amounts and precipitation occurrences. Predictor variables must be physically sensible, realistically modeled by the AOGCM, and able to fully reflect the climate change signal. In the current work, NCEP/NCAR data are used for calibration and validation in order to assess the potential of the proposed approach, although the final objective is to use AOGCM outputs. Even if NCEP data are complete and physically consistent they are still subject to model biases (Hofer et al. 2012). NCEP variables which are not assimilated (such as precipitation), but generated by the parameterizations based on dynamical model can significantly deviate from real weather. The use of such variables for the calibration and validation of empirical downscaling techniques may not be a good idea, since it may induce a significant deviation of the modeled relationships predictors/predictands from the reality which makes evaluation of downscaling techniques more difficult.

The downscaling problem as is tackled in this paper can be viewed as a regression problem, where we try to predict climate variables at small scale from climate variables at synoptic scale. However, due to the large literature that addresses the precipitation modelling in general, the downscaling issue may be viewed as an adjustment of existed precipitation models to account for large scale climate drivers (GCM precipitation, SLP, wind speed, etc.). Wilks (2010) suggested that these adjustments can be accomplished in two ways: (i) through imposed changes in the corresponding monthly statistics, (ii) or by controlling the precipitation model parameters by daily variations in simulated atmospheric circulation. In this context, the VGLM component of the proposed model focuses on the second way in the adjustment procedure. Indeed, through the VGLM component, large scale climate drivers are employed as exogenous variables to describe parameters of the mixed Bernoulli-GP distribution.

6. Conclusions

A VGLM-NB model is proposed in this paper for simultaneously downscaling AOGCM predictors to daily multisite precipitation. The VGLM-NB relies on a probabilistic modeling framework in order to predict the conditional Bernoulli-Generalized Pareto distribution of precipitation at a daily time scale. A non-parametric bootstrapping technique is proposed to preserve a realistic representation of relationships between sites at both time and space. This rank-based sampling method is easy to implement and does not model the dependency structures, but mimic the observed historical characteristics of multisite precipitation and thus avoids any model specification error. However, it should be mentioned that it cannot be used for simulations at ungagged locations. Indeed, in such a situation, modeling the data ranks through spatial copulas would be more appropriate.

The developed model was then applied to generate daily precipitation series at nine stations located in the southern part of the province of Quebec (Canada). Model evaluations suggest that the VGLM-NB model is capable of generating series with realistic spatial and temporal variability. The developed model can be easily applied to other variables such as temperature and wind speed making it a valuable tool not only for downscaling purposes but also for environmental and climatic modelling, where often non-normally distributed random variables are involved.

7. References

AghaKouchak, A. (2014). "Entropy–copula in hydrology and climatology." Journal of Hydrometeorology **15**(6): 2176-2189.

AghaKouchak, A., A. Bárdossy and E. Habib (2010). "Conditional simulation of remotely sensed rainfall data using a non-Gaussian v-transformed copula." Advances in Water Resources **33**(6): 624-634.

Ashkar, F. and T. B. Ouarda (1996). "On some methods of fitting the generalized Pareto distribution." Journal of Hydrology **177**(1): 117-141.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." Water Resour. Res. **42**(11): W11416.

Bárdossy, A. and J. Li (2008). "Geostatistical interpolation using copulas." Water Resour. Res. **44**(7): W07412.

Bárdossy, A. and G. G. S. Pegram (2009). "Copula based multisite model for daily precipitation simulation." Hydrology and Earth System Sciences **13**(12): 2299-2314.

Bargaoui, Z. K. and A. Bárdossy (2015). "Modeling short duration extreme precipitation patterns using copula and generalized maximum pseudo-likelihood estimation with censoring." Advances in Water Resources **84**: 1-13.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9).

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of Climate **28**(6): 2349-2364.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bremnes, J. B. (2004). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Buishand, T. A. and T. Brandsma (1999). "Dependence of precipitation on temperature at Florence and Livorno (Italy)." Climate Research **12**(1): 53-63.

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

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chandler, R. E. and H. S. Wheater (2002). "Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland." Water Resources Research **38**(10): 10-11-10-11.

Chebana, F. and T. B. Ouarda (2011). "Multivariate quantiles in hydrological frequency analysis." Environmetrics **22**(1): 63-78.

Coe, R. and R. Stern (1982). "Fitting models to daily rainfall data." Journal of Applied Meteorology **21**(7): 1024-1031.

Conway, D., R. Wilby and P. Jones (1996). "Precipitation and air flow indices over the British Isles." Climate Research **7**: 169-183.

Czado, C., E. C. Brechmann and L. Gruber (2013). Selection of vine copulas. Copulae in Mathematical and Quantitative Finance, Springer**:** 17-37.

El Adlouni, S., B. Bobée and T. Ouarda (2008). "On the tails of extreme event distributions in hydrology." Journal of Hydrology **355**(1): 16-33.

Eum, H.-I., P. Gachon, R. Laprise and T. Ouarda (2012). "Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme." Climate Dynamics **38**(7-8): 1433-1457.

Fang, H.-B., K.-T. Fang and S. Kotz (2002). "The meta-elliptical distributions with given marginals." Journal of Multivariate Analysis **82**(1): 1-16.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review **135**(6).

Genest, C. and F. Chebana (2015). "Copula modeling in hydrologic frequency analysis." In Handbook of Applied Hydrology (V.P. Singh, Editor) **McGraw-Hill, New York,** (in press).

Genest, C. and J. F. Plante (2003). "On Blest's measure of rank correlation." Canadian Journal of Statistics **31**(1): 35-52.

Giorgi, F., J. Christensen, M. Hulme, H. Von Storch, P. Whetton, R. Jones, L. Mearns, C. Fu, R. Arritt and B. Bates (2001). "Regional climate information-evaluation and projections." Climate Change 2001: The Scientific Basis. Contribution of Working Group to the Third Assessment Report of the Intergouvernmental Panel on Climate Change [Houghton, JT et al.(eds)]. Cambridge University Press, Cambridge, United Kongdom and New York, US.

Gräler, B. (2014). "Modelling skewed spatial random fields through the spatial vine copula." Spatial Statistics **10**: 87-102.

Guerfi, N., A. A. Assani, M. Mesfioui and C. Kinnard (2015). "Comparison of the temporal variability of winter daily extreme temperatures and precipitations in southern Quebec (Canada) using the Lombard and copula methods." International Journal of Climatology.

Harpham, C. and R. L. Wilby (2005). "Multi-site downscaling of heavy daily precipitation occurrence and amounts." Journal of Hydrology **312**(1): 235-255.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Hessami, M., P. Gachon, T. B. M. J. Ouarda and A. St-Hilaire (2008). "Automated regression-based statistical downscaling tool." Environmental Modelling & Software **23**(6): 813-834.

Hobæk Haff, I., A. Frigessi and D. Maraun (2015). "How well do regional climate models simulate the spatial dependence of precipitation? An application of pair-copula constructions." Journal of Geophysical Research: Atmospheres **120**(7): 2624-2646.

Hofer, M., B. Marzeion and T. Mölg (2012). "Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru)." Climate Dynamics **39**(7-8): 1969-1980.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2012). "Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada." Stochastic Environmental Research and Risk Assessment **26**(5): 633-653.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Joe, H. (1997). Multivariate models and multivariate dependence concepts, CRC Press.

Khalili, M., V. T. Van Nguyen and P. Gachon (2013). "A statistical approach to multi-site multivariate downscaling of daily extreme temperature series." International Journal of Climatology **33**(1): 15-32.

Kleiber, W., R. W. Katz and B. Rajagopalan (2012). "Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes." Water Resources Research **48**(1).

Lagarias, J. C., J. A. Reeds, M. H. Wright and P. E. Wright (1999). "Convergence properties of the Nelder-Mead simplex method in low dimensions." SIAM Journal on Optimization **9**(1): 112-147.

Lee, T., R. Modarres and T. Ouarda (2013). "Data-based analysis of bivariate copula tail dependence for drought duration and severity." Hydrological Processes **27**(10): 1454-1463.

Lee, T., T. B. Ouarda and C. Jeong (2012). "Nonparametric multivariate weather generator and an extreme value theory for bandwidth selection." Journal of Hydrology **452**: 161-171.

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

Li, C., V. P. Singh and A. K. Mishra (2013). "Monthly river flow simulation with a joint conditional density estimation network." Water Resources Research **49**(6): 3229-3242.

Lindström, G., B. Johansson, M. Persson, M. Gardelin and S. Bergström (1997). "Development and test of the distributed HBV-96 hydrological model." Journal of Hydrology **201**(1-4): 272-288.

Mao, G., S. Vogl, P. Laux, S. Wagner and H. Kunstmann (2015). "Stochastic bias correction of dynamically downscaled precipitation fields for Germany through Copula-based integration of gridded observation data." Hydrology and Earth System Sciences **19**(4): 1787-1806.

Mehrotra, R. and A. Sharma (2006). "A nonparametric stochastic downscaling framework for daily rainfall at multiple locations." Journal of Geophysical Research: Atmospheres (1984–2012) **111**(D15).

Mehrotra, R. and A. Sharma (2009). "Evaluating spatio-temporal representations in daily rainfall sequences from three stochastic multi-site weather generation approaches." Advances in Water Resources **32**(6): 948-962.

Mehrotra, R., A. Sharma and I. Cordery (2004). "Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence." Journal of Geophysical Research: Atmospheres (1984–2012) **109**(D14).

Nelsen, R. B. (2013). An introduction to copulas, Springer Science & Business Media.

Oakes, D. (1982). "A model for association in bivariate survival data." Journal of the Royal Statistical Society. Series B (Methodological): 414-422.

Ouarda, T. B. M. J., J. W. Labadie and D. G. Fontaine (1997). "Indexed sequential hydrologic modeling for hydropower capacity estimation." Journal of the American Water Resources Association **33**(6): 1337-1349.

Requena, A. I., I. Flores, L. Mediero and L. Garrote (2015). "Extension of observed flood series by combining a distributed hydro-meteorological model and a copula-based model." Stochastic Environmental Research and Risk Assessment: 1-16.

Salvadori, G. and C. De Michele (2007). "On the use of copulas in hydrology: theory and practice." Journal of Hydrologic Engineering **12**(4): 369-380.

Serinaldi, F. (2009). "A multisite daily rainfall generator driven by bivariate copula-based mixed distributions." Journal of Geophysical Research: Atmospheres (1984–2012) **114**(D10).

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

Serinaldi, F., A. Bárdossy and C. G. Kilsby (2014). "Upper tail dependence in rainfall extremes: would we know it if we saw it?" Stochastic Environmental Research and Risk Assessment **29**(4): 1211-1233.

Serinaldi, F. and C. G. Kilsby (2014). "Simulating daily rainfall fields over large areas for collective risk estimation." Journal of Hydrology **512**: 285-302.

Shannon, C. (1948). "A mathematical theory of communication." Bell Syst Tech J **27**(3): 379–423.

Smith, M. S. (2014). "Copula modelling of dependence in multivariate time series." International Journal of Forecasting.

Song, S. and V. P. Singh (2010). "Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data." Stochastic Environmental Research and Risk Assessment **24**(3): 425-444.

Srivastav, R. K. and S. P. Simonovic (2014). "Multi-site, multivariate weather generator using maximum entropy bootstrap." Climate Dynamics **44**(11-12): 3431-3448.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Stern, R. and R. Coe (1984). "A model fitting analysis of daily rainfall data." Journal of the Royal Statistical Society. Series A (General): 1-34.

Vaz de Melo Mendes, B. and R. P. C. Leal (2010). "Portfolio management with semi-parametric bootstrapping." Journal of Risk Management in Financial Institutions **3**(2): 174-183.

Vernieuwe, H., S. Vandenberghe, B. De Baets and N. E. Verhoest (2015). "A continuous rainfall model based on vine copulas." Hydrology and Earth System Sciences Discussions **12**(1): 489-524.

Vinod, H. D. and J. López-de-Lacalle (2009). "Maximum entropy bootstrap for time series: the meboot R package." Journal of Statistical Software **29**(5): 1-19.

Wilks, D. S. (2010). "Use of stochastic weathergenerators for precipitation downscaling." Wiley Interdisciplinary Reviews: Climate Change **1**(6): 898-907.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models." Progress in Physical Geography **23**(3): 329-357.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Yee, T. W. and A. G. Stephenson (2007). "Vector generalized linear and additive extreme value models." Extremes **10**(1-2): 1-19.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." Journal of the Royal Statistical Society. Series B (Methodological): 481-493.

Zhang, Q., M. Xiao and V. P. Singh (2015). "Uncertainty evaluation of copula analysis of hydrological droughts in the East River basin, China." Global and Planetary Change **129**: 1-9.

List of Tables

Table 1. List of the 9 stations used in this study.

No.	Site	Name of station	Latitude (°N)	Longitude (°W)
1	7031360	Chelsea	45.52	-75.78
2	7014290	Cedars	45.3	-74.05
3	7025440	Nicolet	46.25	-72.60
4	7022160	Drummondville	45.88	-72.48
5	7012071	Donnacona 2	46.68	-71.73
6	7066685	Roberval A	48.52	-72.27
7	7060400	Bagotville A	48.33	-71
8	7056480	Rimouski	48.45	-68.53
9	7047910	Seven Island A	50.22	-66.27

Table 2. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 3. Quality assessment of the estimated series for the validation period (1981–2000) for VGLM-NB and VGLM-MAR. Statistics are ME and RMSE, Differences between observed and modeled variances (D) and false alarm ratio FAR.

Number of station		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RMSE	VGLM-NB	**7.34**	**7.17**	7.29	5.53	**6.06**	**5.49**	**5.49**	5.49	6.47
	VGLM-MAR	7.37	7.22	**6.91**	**5.18**	6.28	5.60	5.60	**5.36**	**6.29**
ME	VGLM-NB	**0.02**	-0.29	-0.31	-0.46	**-1.03**	-0.30	-1.05	**-0.24**	**0.13**
	VGLM-MAR	0.43	**-0.27**	**-0.30**	**-0.41**	-1.04	-0.30	**-0.90**	-0.28	0.48
D	VGLM-NB	-19.55	**7.58**	**-1.05**	**5.13**	19.28	8.19	18.15	**2.81**	-9.52
	VGLM-MAR	**-17.41**	8.66	2.23	8.21	**18.34**	**7.84**	**17.45**	3.38	**-7.23**
FAR	VGLM-NB	**0.35**	**0.356**	**0.31**	**0.31**	**0.33**	**0.37**	**0.33**	**0.36**	**0.37**
	VGLM-MAR	0.39	0.37	0.34	0.33	0.35	0.41	0.37	0.41	0.41

Bold character means better result.

Table 4. RMSE of precipitation indices for the validation period (1981–2000) for both VGLM-NB and VGLM-MAR.

	Indices	VGLM-NB	VGLM-MAR
Precipitation amount	PX1D (mm)	**23.25**	33.40
	PX3D (mm)	**21.31**	35.85
	PX5D (mm)	**21.59**	34.63
	Pmax90 (mm)	3.71	**3.44**
	MWD (mm)	**1.47**	1.99
Precipitation occurrences	WRUN (days)	**1.96**	2.10
	DRUN (days)	**3.32**	4.41
	NWD (days)	**4.09**	4.65

Bold character means better result.

List of Figures

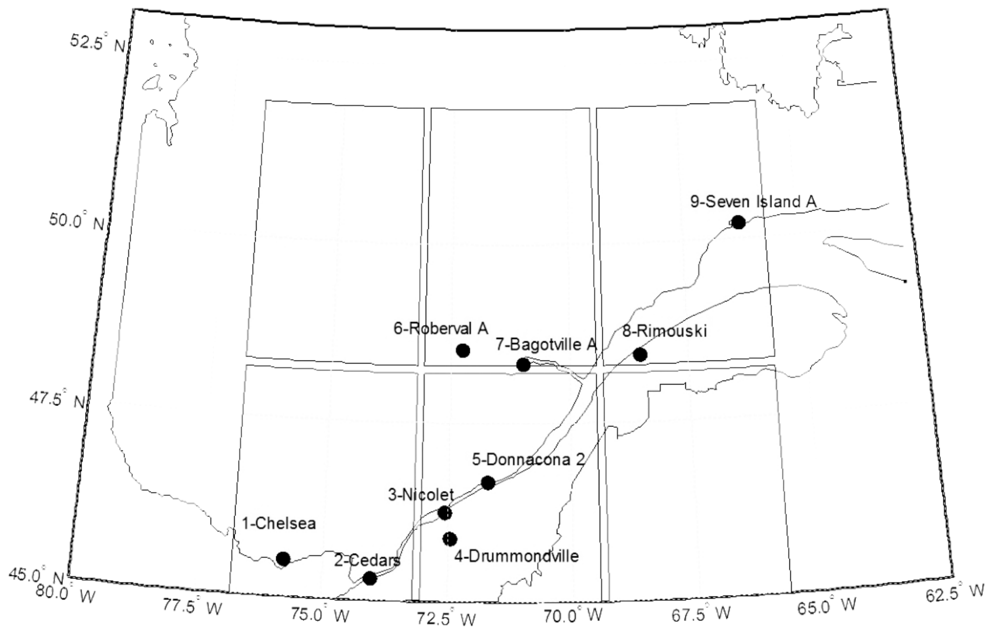

Figure 1. The locations of precipitation stations and CGCM3 grid.

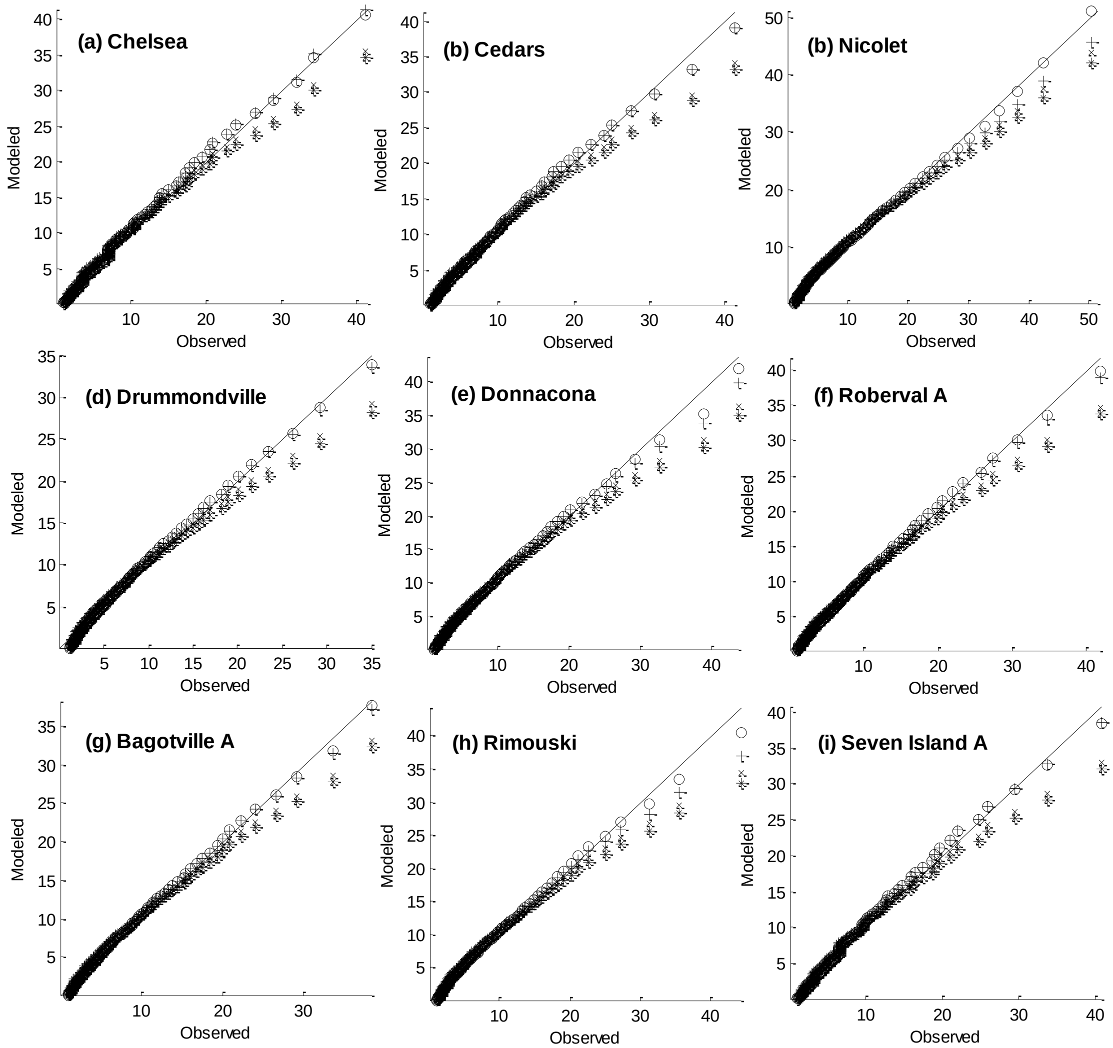

Figure 2. Q–Q plot of observed and modeled quantiles for Gamma distribution (stars), Reverse WEI distribution (x-mark), GP distribution (circles) and mixed Exponential distribution (plus).

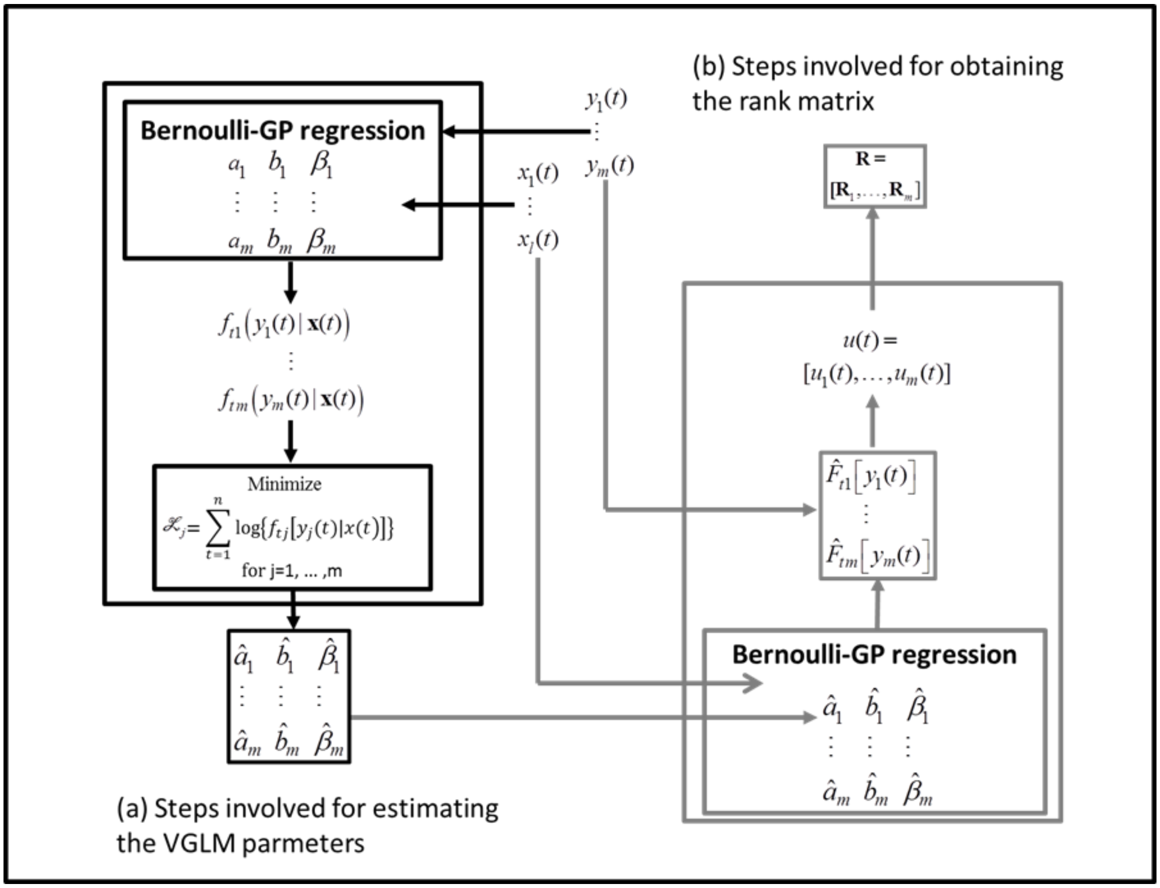

Figure 3. Steps involved for estimating the VGLM prameters (a) and obtaining the rank matrix (b).

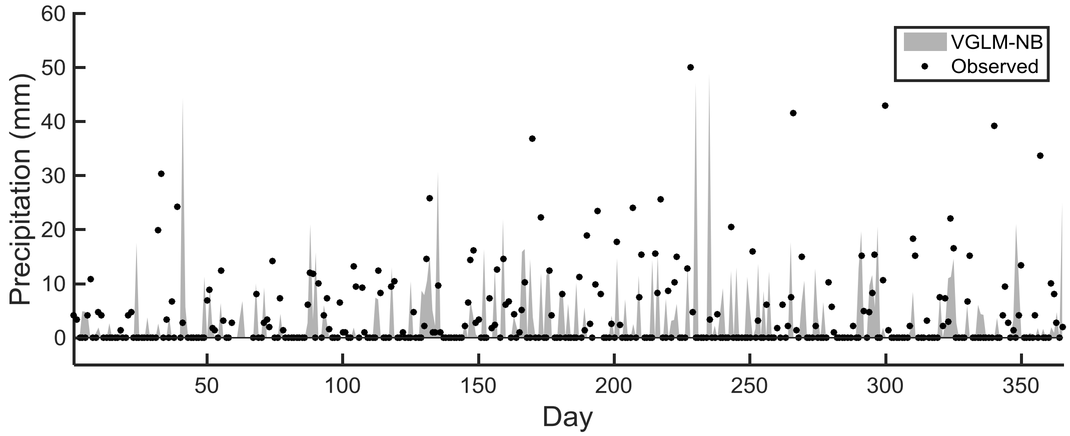

Figure 4. Example of one precipitation simulation using VGLM-NB at Cedars station during 1981.

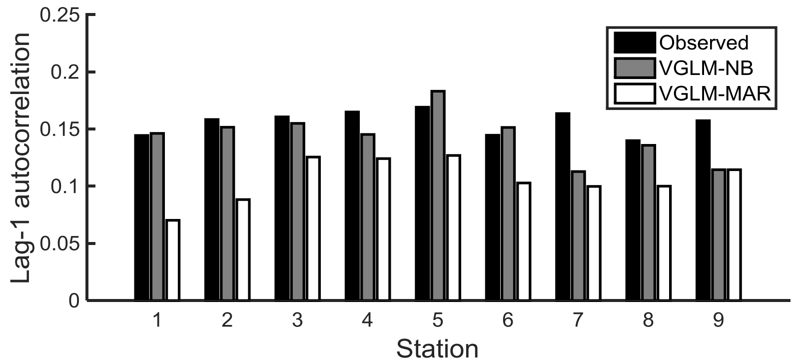

Figure 5. Observed and modeled lag-1 autocorrelation for precipitation series at the nine stations during the validation period.

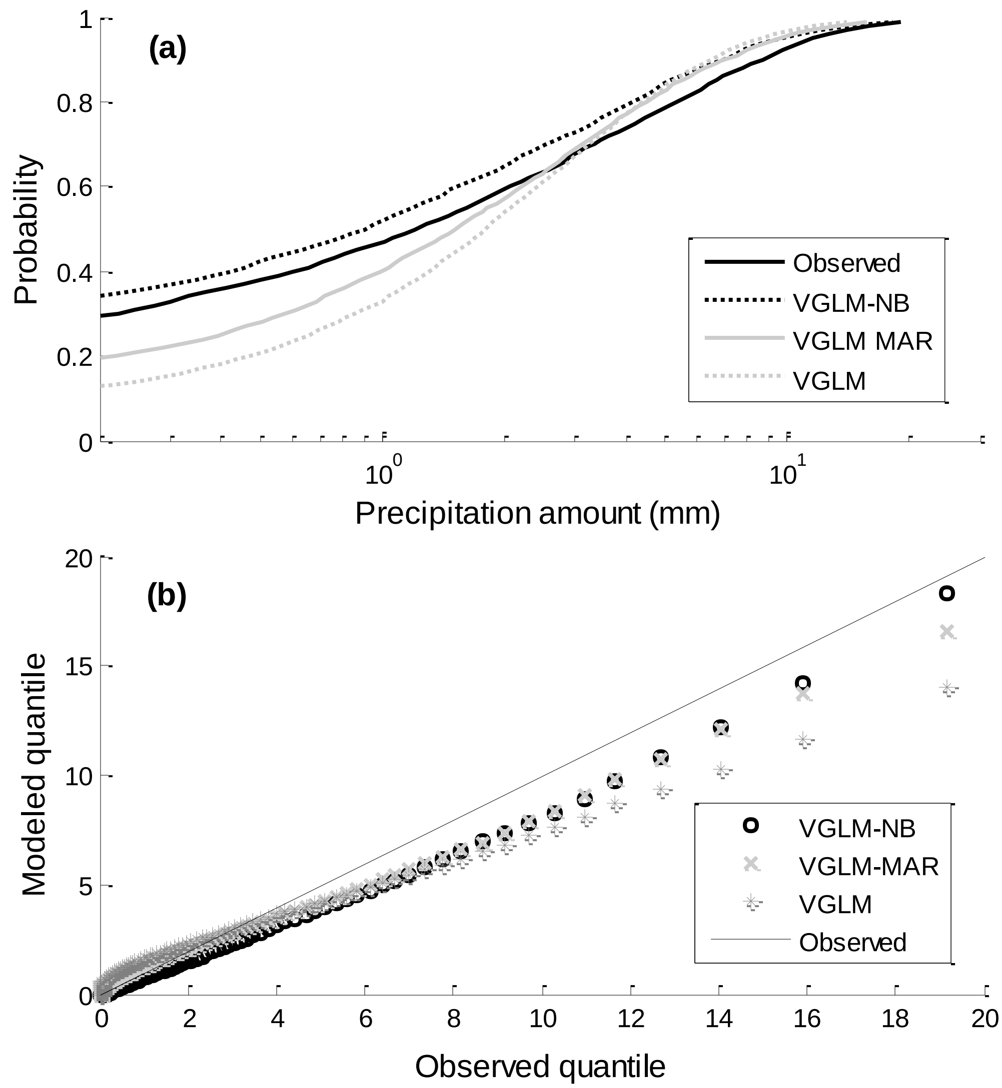

Figure 6. Observed and predicted daily average precipitation over the nine stations. The CDF is presented in (a) and the Q-Q plots in (b).

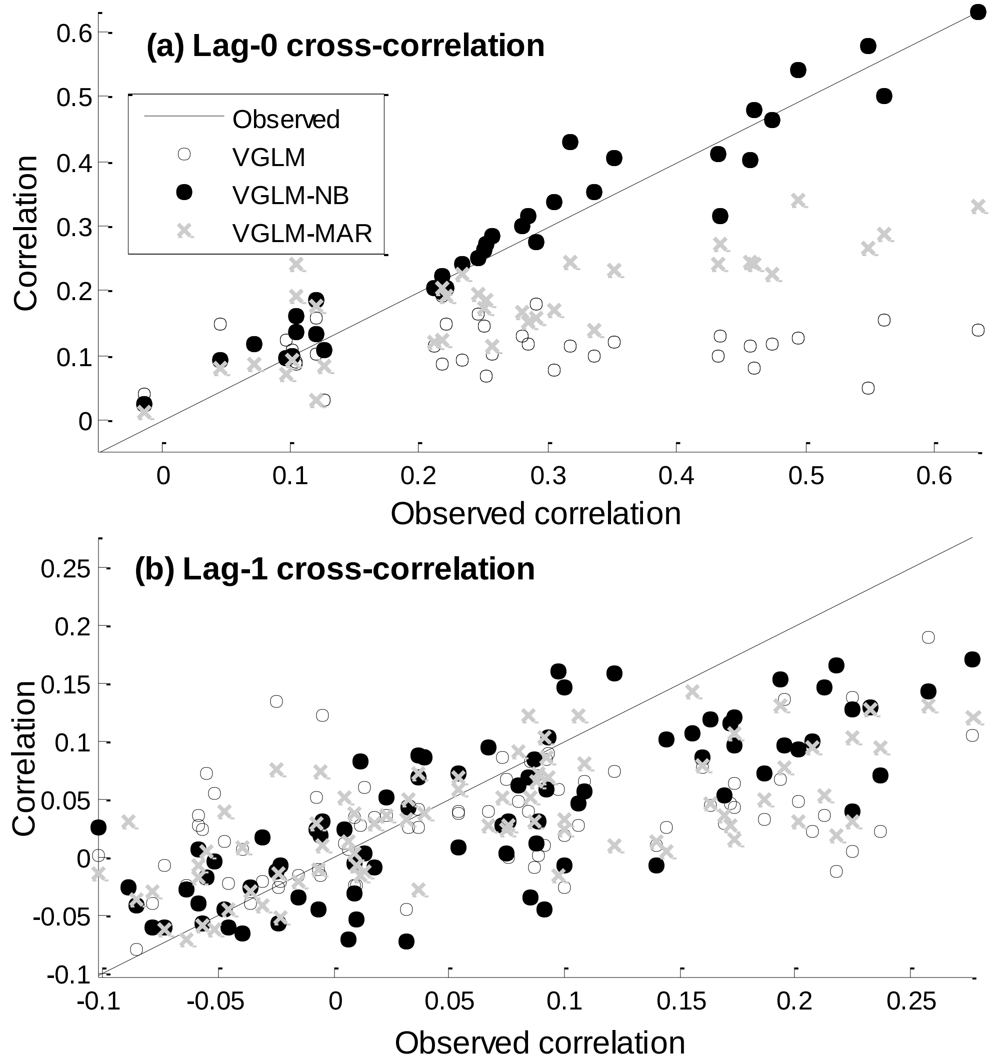

Figure 7. Scatter plots of observed and modeled lag-0 cross-correlation (a) and lag-1 cross-correlation during the validation period. Correlation values are obtained using the mean of the correlation values calculated from 100 simulations.

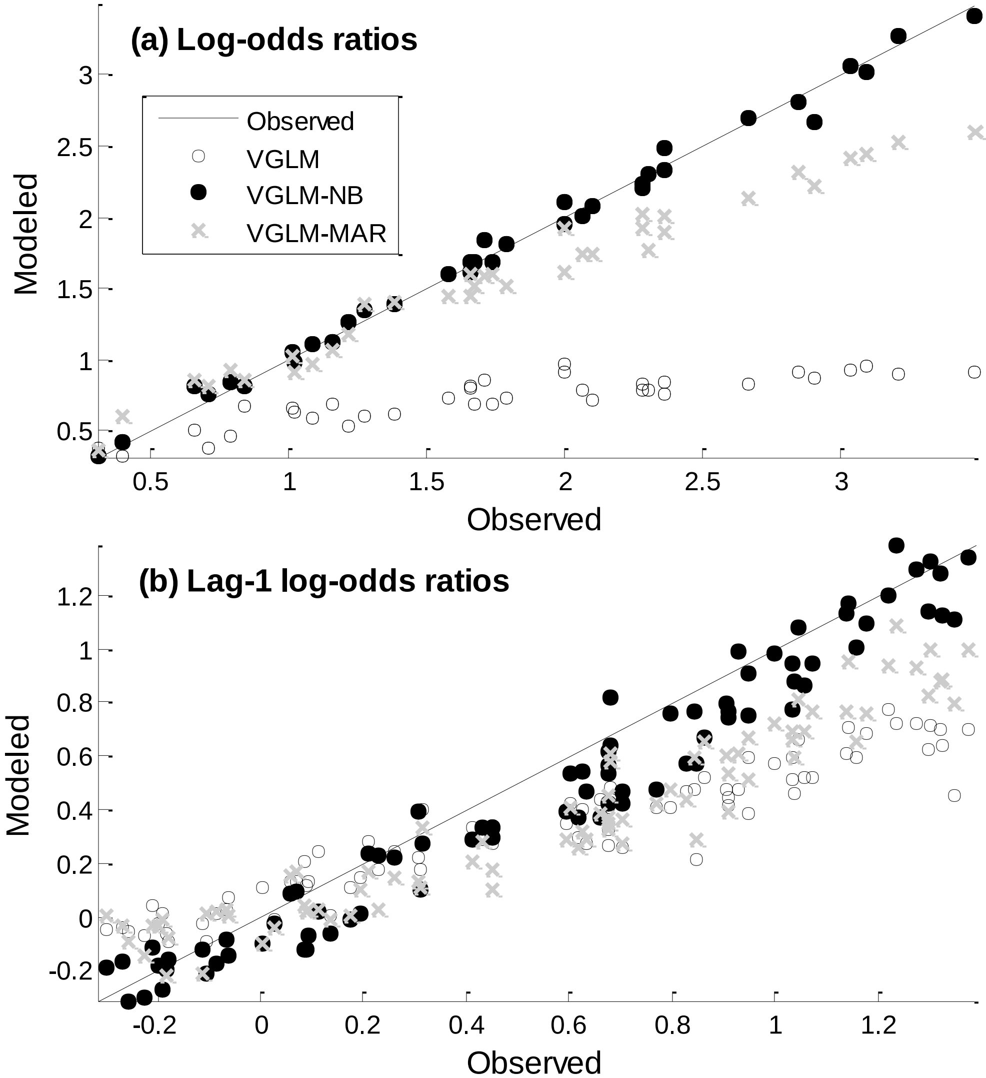

Figure 8. Scatter plots of observed and modeled log odds ratios (a) and lag-1 log odds ratios during the validation period. Values are obtained using the mean values from 100 simulations.

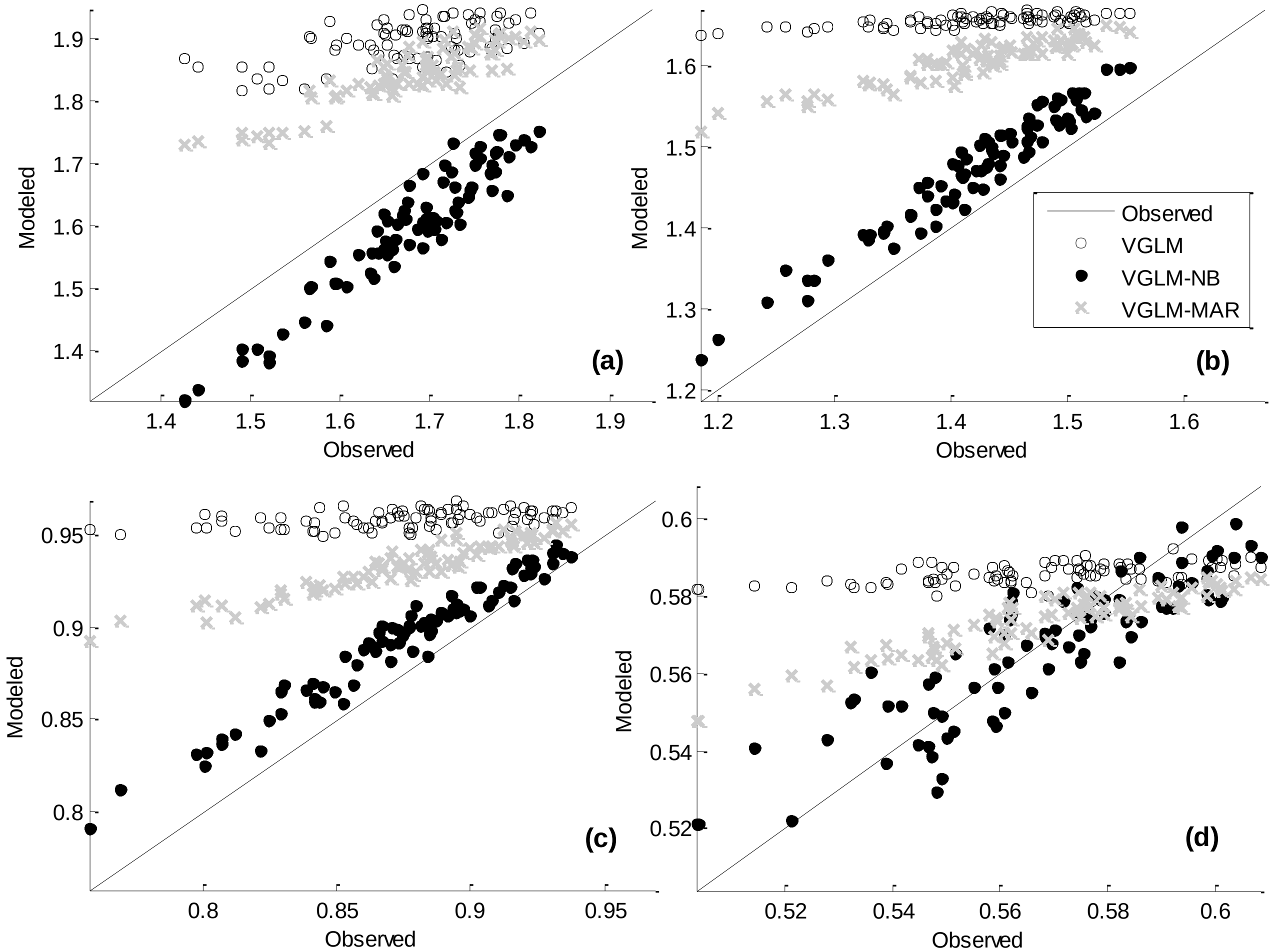

Figure 9. Scatter plots of observed and modeled binary entropy for precipitation occurrences (a), and at three quantile thresholds: 0.75 (b), 0.90 (c) and 0.95 (d). Points correspond to all combinations of triplets of stations.