Stochastic Environmental Research and Risk Assessment (2020) 34:2159–2177
https://doi.org/10.1007/s00477-020-01871-4

ORIGINAL PAPER

Performance of multisite stochastic precipitation models for a tropical monsoon region

Tue M. Vu[1] · Ashok K. Mishra[1]

Accepted: 31 August 2020 / Published online: 10 September 2020

Abstract
Stochastic weather generator (SWG) produces synthetic time series of weather data based on the statistical characteristics of observed weather for a given location. Although SWG models are extensively evaluated and applied in different hydro-climate related studies, they often ignore the spatial correlation between weather patterns observed at multiple locations. This can limit the value of some spatial impact assessments such as flood modeling, agricultural crop modeling, water resources management and urban infrastructure design. To address such limitations, multisite SWG models are implemented to preserve the spatial characteristics of weather variables. In this study, we compared the performance of three multisite stochastic precipitation models, which includes modified Wilks model (modWilks), RainSim V3 (RSIM) and perturbed K-Nearest Neighbor (pKNN) models. The performances of these models are investigated for a study area located in the tropical monsoon climate region over Central Highland, Vietnam. The models are evaluated based on their performance for simulating precipitation occurrence and amount statistics on a wet day, extreme cumulative wet/dry days, transition and joint probability of wet/dry state, cross-correlation across all sites as well as the behavior of precipitation amount in relation to neighboring station state. The performance of model depends on the type of the precipitation characteristics, for example, the RSIM model performed well in term of the mean precipitation intensity. Overall, the pKNN model outperformed other models in term of temporal statistics, spatial characteristics, as well as extreme events measured based on Intensity–Duration–Frequency (IDF) curves.

Keywords Multisite stochastic precipitation model · Perturbed KNN · Modified Wilks · RainSim · Tropical monsoon region · Vietnam

1 Introduction

Climate change likely to alter the seasonal precipitation and evaporation pattern resulting in more flood and drought events in different parts of the world (Konapala et al. 2020), for example the extreme rainfall pattern has recently increased over the USA (Vu and Mishra 2019). In order to develop appropriate strategies to mitigate such climate extreme events, long term data sets are needed to provide risk-based information for infrastructure development. However, the long-term data sets are often very limited in most part of the world, especially in developing countries (Mishra and Coulibaly 2009). Stochastic weather generators (SWG) are able to generate long-term synthetic weather data for applications in water resources planning, urban infrastructure design and environmental models. Different types of SWG models are available, which include, Weather GENerator (WGEN; Richardson 1981); CLImate GENerator (CLIGEN; Nicks and Gander 1994); (Long Ashton Research Station Weather Generator (LARS-WG; Semenov and Barrow 1997); Climate Data Generator (ClimGen; Stockle et al. 1999); WeatherMan (Pickering et al. 1994). Based on these models, the synthetic data are simulated for the individual station, which can be a major limitation in hydrologic modeling as well as for creating input data sets for rainfall-runoff models, where the statistical characteristics of multiple

✉ Tue M. Vu
tuev@g.clemson.edu

Ashok K. Mishra
ashokm@g.clemson.edu

[1] Glenn Department of Civil Engineering, Clemson University, Clemson, SC 29631, USA

(neighboring) stations need to be preserved. To overcome this limitation, a number of multisite SWG models were developed to accommodate the space–time dependency associated with multiple weather stations. These multisite SWG models are either conditioned on exogenous atmospheric forcing predictors (Bardossy and Plate 1992; Wilson et al. 1992; Bogardi et al. 1993; Rajagopalan and Lall 1999; Beersma and Buishand 2003) or self-dependence on variable itself (Wilks 1998; Qian et al. 2002; Brissette et al. 2007; Mhanna and Bauwens 2012; Chen et al. 2014).

The stochastic rainfall models can be classified into parametric, non-parametric and hybrid (or semi-parametric) approaches (Baigorria and Jones 2010). Parametric approach (Richardson 1981) splits the stochastic rainfall generation into two parts: precipitation occurrence (majority are modulated by Markov chain for two-states wet/dry days) and precipitation amount on wet days (fitting precipitation on wet days based on distributions such as gamma, skewed normal, Weibull or exponential). Whereas, the non-parametric approach resamples the precipitation based on nearest neighbor (Rajagopalan and Lall 1999) or Poisson cluster model (Burton et al. 2008; Kim et al. 2017). Hybrid stochastic rainfall model combines the advantages associated with these two approaches, for instance, the precipitation occurrence process is simulated based on the parametric approach (Markov chain) and the precipitation amount is based on the nearest neighbor concept (Apipattanavis et al. 2007; Steinschneider and Brown 2013).

One of the earliest and most popular multisite rainfall generation models was developed by Wilks (1998) (in brief: Wilks model) by extending the well-known single site stochastic rainfall model WGEN by Richardson (1981). The extension was developed by integrating an individual single-site model through a temporally independent but spatially correlated sequence of random numbers. Wilks model utilized the Markov chain two states first-order process for precipitation occurrence and a mixed exponential distribution to fit wet day rainfall amount. Subsequently, Wilks (1998) proposed to statistically downscale precipitation to multiple sites with large scale climate predictors. In a follow-up study, Qian et al. (2002) modified the Wilks model by formulating the precipitation occurrence process with two state first-order Markov chain conditioning on daily circulation patterns (such as Mean Sea Level Pressure) for six stations in Portuguese Guadiana River basin. The conditional simulation apparently improved the model performance in comparison to unconditional models by better capturing the observed statistics. Brissette et al. (2007) further improved the Wilks model by implementing a new algorithm for generating precipitation amounts at eight weather stations located in Quebec, Canada. Mhanna and Bauwens (2012) found the inefficiency of using random numbers to replace the non-positive definite pair-wise correlation matrix in the original Wilks model. To overcome this limitation, they suggested replacing negative eigenvalues by zeros or small positive values that further improved the performance of the multisite dependencies. Chen et al. (2014) developed a multisite SWG model MulGETS for daily precipitation and temperature for weather stations located in Canada, which incorporates the single site information with temporally independent but spatially correlated random numbers. The wet day precipitation amount was fitted with a multi gamma distribution (a combination of several gamma distributions) to address the spatial correlation of precipitation amounts.

Although stochastic models based on the parametric approach are able to simulate the event-based nature of the precipitation process, it may not be suitable for simulating statistics related to persistence and extreme events (Gregory et al. 1992). Several alternative models were developed to overcome such limitations associated with parametric models, for example, the Neyman–Scott Rectangular Pulse model (NSRP; Rodriguez-Iturbe et al. (1987)). The basic Neyman–Scott model (Neyman and Scott 1958) follows a Poisson process of storm origins which are associated with rectangular pulses (rain cells), with heights corresponding to rainfall intensity and widths to cell duration (Cowpertwait et al. 2002). Space–Time NSRP model was developed by Cowpertwait et al. (1995, 2002) based on the extension of the NSRP model into two-dimensional space. Subsequently, Burton et al. (2008) incorporated this stochastic rainfall modeling into the RainSim V3 model which significantly improved the daily statistics and annual extremes simulations. RainSim models are useful for statistical downscaling of Global Climate Models (GCMs) outputs. For example, Forsythe et al. (2014) combined RainSim with a Climate Research Unit Daily Weather Generator (CRU-WG) to downscale the GCMs simulation for a semi-arid mountain climate. Based on Burton et al. (2008), Bordoy and Burlando (2014) further improved the Spatio-temporal NSRP model to stochastically downscale climate model in orographic complex regions.

K-nearest neighbor (KNN) is another popular non-parametric SWG model that does not assume any prior distribution of weather variables. This approach is a bootstrapping technique and it can include additional covariates for simulation weather variables as well as for downscaling processes. Rajagopalan and Lall (1999) applied KNN to model daily precipitation and other climate variables for Salt Lake City, Utah. In another study, Buishand and Brandsma (2001) applied the KNN model to generate multisite daily precipitation and temperature for 25 stations along the Rhine river basin in Germany. It was observed that the KNN model is able to preserve the dependence characteristic between daily precipitation amounts at different sites, and it was used to statistically downscale the GCMs outputs for climate change studies (Yates et al. 2003; Gangopadhyay et al. 2005; Muluye 2011; Devak et al. 2015). Apipattanavis et al. (2007) proposed a semiparametric multivariate SWG model by integrating the three-state (dry, wet, extreme wet), first-order Markov chain to simulate the precipitation state and used a KNN model for simulating wet day and extreme wet day rainfall intensity. The proposed hybrid SWG model applied to generate daily weather data for Pergamino, Argentina, which indicated that the (dry/wet) spell statistics were better preserved by a modified model as well as an appropriate shift in the probability distribution function of the seasonal precipitation. Although KNN based SWG models were extensively applied in precipitation simulation, it has a key limitation due to the application of bootstrapping technique that shuffles historical days to form new sequences (Rajagopalan and Lall 1999; Buishand and Brandsma 2001), and the simulated values may not exceed historical values. This can be a limitation for non-stationary analysis of climate variables as well as modeling extreme events. Sharif and Burn (2007) suggested a modified approach that allows the nearest neighbor resampling procedure with perturbation of the historical data by adding a random component to the individual resampled data points. Recently, Agilan and Umamahesh (2019) introduced the modified version of the KNN approach by incorporating time covariate based non-stationary Generalized Pareto Distribution (GPD) which improved the simulation of extreme precipitation events for selected rainfall stations located in India and Vietnam.

The literature review suggested a wide range of application of stochastic rainfall modeling in different sectors (flood control, drought management, crop yield simulation, climate change impact and adaptation) (Peck et al. 2012; Thorndahl et al. 2017; Shrestha et al. 2017; Vu et al. 2018; Nguyen 2019; Aryal et al. 2019; Dao et al. 2020). For example, Peck et al. (2012) applied a single site SWG using the KNN model to construct the IDF curves for the city of London in Ontario, Canada. A similar rainfall resampling approach for constructing a single station's IDF curve was also reported in Thorndahl et al. (2017) and Shrestha et al. (2017). Although single-site rainfall models can generate IDF curves for a specific location, they often neglect the spatial dependency between the neighboring gauging stations, which can further limit the application of IDF curves for infrastructure designs. Given the advantages and disadvantages of multisite stochastic rainfall generation models, a limited number of studies compared their performance in different parts of the world. Vallam and Qin (2016) compared four multisite rainfall simulation [Multisite Stochastic Weather Generators (MulGETS), Multisite Rainfall Simulations (MRS), RainSim, KNN] for an urban tropical region based on 5 stations located in Singapore and concluded that the selected models did not perform well. In a recent study, Vu et al. (2018) compared five stochastic rainfall models (CLIGEN, ClimGen, LARS-WG, RainSim and WeatherMan) in three diverse climatic regions and concluded that the performance of these models varies with respect to climatic regimes. As thus, careful validation is required depending on the weather pattern as well as its applications in different water resources sectors.

In this study three representative models are selected, which includes Modified Wilks model (Mhanna and Bauwens 2012), RainSim v3 (Burton et al. 2008) and modified KNN model based on the original model by Rajagopalan and Lall (1999). These three models are based on different simulation mechanisms, for example, the first model is based on the parametric approach by formulating the rainfall occurrence process using first-order two-state Markov chain and rainfall intensity on wet days are derived by fitting 2-parameter Gamma distribution. The authors improved the performance of multisite dependencies by replacing the negative eigenvalues with small positive values. The second model applied different approaches using space–time NSRP to formulate the storm origins based on the Poisson process. The third model, KNN was further improved with the perturbation based on kernel bandwidth simulated by Least Square Cross Validation method. Due to the limited number of applications of multisite stochastic precipitation simulation for the tropical monsoon region with distinct wet/dry seasons, this study provides an opportunity to investigate the strengths and weaknesses of the multisite precipitation simulations models in a tropical monsoon region. This study is very relevant to flood and drought management as these extreme events are very frequent in such study domain. With this study, it is possible to identify strength of individual models for simulating different statistical properties of multisite precipitation data sets.

The overall objective of this study are: (a) to improve the traditional KNN model by incorporating a revised kernel bandwidth simulated by Least Square Cross Validation method to further improve their ability to preserve the spatial extreme rainfall characteristic; (b) to compare and evaluate the performance of pKNN with respect to the other popular stochastic rainfall generation model such as modWilks and RSIM for simulating space–time precipitation characteristics, (c) to capture the ability of multisite rainfall generation models in simulating the extreme events corresponding to thousand-year return periods, and (d) to generate and compare the capabilities of these stochastic models for generating IDF curves with respect to the traditional methods derived based on the Generalized

Extreme Value distribution fitted to the observed datasets. The study location is prone to flood and drought and is crucial for the socio-economy development of the country. The selected stochastic models are derived based on distinct model configurations and therefore they are well suited for comparative analysis. The paper is organized as follows. Section 2 illustrates the study area. Section 3 outlines the methodology and detail description of each multisite stochastic rainfall model used in the study. Detail evaluations, analyses and comparisons are presented in Sect. 4. The discussions and conclusions are summarized in the last sections.

2 Study area and data

The study area is located in the tropical monsoon region in Central Highland of Vietnam, which has a very complex terrain on a high topography with more than 1000 m above mean sea level. The total area of this region is about 60,000 km^2. Total annual rainfall varies between 1500 and 2400 mm with 80% of the precipitation occurs during the rainy season (May to October). Due to its steep slope topography and high rainfall amount during the rainy season, this area is prone to flash flood and drought events. Therefore, the spatio-temporal quantification of rainfall events using a multisite stochastic precipitation model can improve our understanding of the regional evolution of floods and droughts in the region. We have selected eight meteorological stations with a longer data period and fewer missing values. The data record includes daily precipitation, max, min and average temperature for 25 years (1981–2005) with less than 1% missing data. The location of eight stations overlay on the topography of Central Highland derived from SRTM (Shuttle Radar Topography Missions) 90 m resolution is provided in Fig. 1. Situated on a high mountainous area, the study region has quite a complex terrain, therefore the rainfall characteristics of selected stations depend on the distribution of mountain range that blocks the seasonal monsoon wind that flows from the northeast and southwest direction. Daily precipitation data has been used in the modWilks and RSIM models for stochastic modeling. The average daily temperature is an added covariate in the generation stochastic precipitation model in the pKNN approach.

3 Methodology

3.1 Modified Wilks model (modWilks)

Wilks (1998) model is a multisite stochastic rainfall model that integrated two-state first-order Markov chain for precipitation occurrences and a mixed exponential distribution for wet day rainfall amount. Wilks model overcomes the limitation of the single-site rainfall generation model by accounting the spatial correlation information between the stations. The model was extended by applying serially independent and spatially correlated random numbers to single site models that have a direct effect on the synthetic series with a realistic spatial correlation. Mhanna and Bauwens (2012) highlighted the limitation of the Wilks model that may result due to non-positive definiteness of the covariance matrices to preserve the spatial correlation between stations, and they further improved the model by applying analytical solutions. During the occurrence of wet/dry day sequence, the correlation matrix between occurrence series of station pairs i and j depend on the joint probability that both station pairs are dry $\pi_{00}(i,j)$ and on the marginal probability of single station is dry $\pi_0(i)$ and $\pi_0(j)$ as shown in Eq. (1):

$$\xi(i,j) = \frac{\pi_{00}(i,j) - \pi_0(i)\pi_0(j)}{\sigma(i)\sigma(j)} \tag{1}$$

where σ denotes the standard deviation as a function of the marginal probability for a given station. However, during the construction of the empirically derived curves for each pair of stations, Mhanna and Bauwens (2012) noticed that there is no significant correlation between random number $\omega(i,j)$ and the marginal probabilities $\pi_0(i)$ & $\pi_0(j)$. Therefore the authors applied the gamma coefficient γ (Eq. 2) instead of ξ to fully exploit the dependence of correlation matrix on joint probability:

$$\gamma(k,l) = \frac{\varphi(i,j) - 1}{\varphi(i,j) + 1} \tag{2}$$

where φ is odds-ratio and dependent only on joint probability as computed in Eq. (3):

$$\varphi(i,j) = \frac{\pi_{00}(i,j)\pi_{11}(i,j)}{\pi_{10}(i,j)\pi_{01}(i,j)} \tag{3}$$

.

For wet day rainfall amount generation, Mhanna and Bauwens (2012) applied the '*rank correlation concept*' to specify the correlations between rainfall series using a fractile correlation (also called rank correlation), which was further converted into the product-moment correlation to generate the required correlations of random numbers. For rainfall amount generation, Mhanna and Bauwens (2012) used the gamma distribution instead of mixed exponential distribution as used in the Wilks model. By doing so, for each individual station and each month, the non-zero precipitation time series are fitted with 2-parameter gamma distribution. The shape and scale parameter are estimated by the maximum likelihood method. Gamma

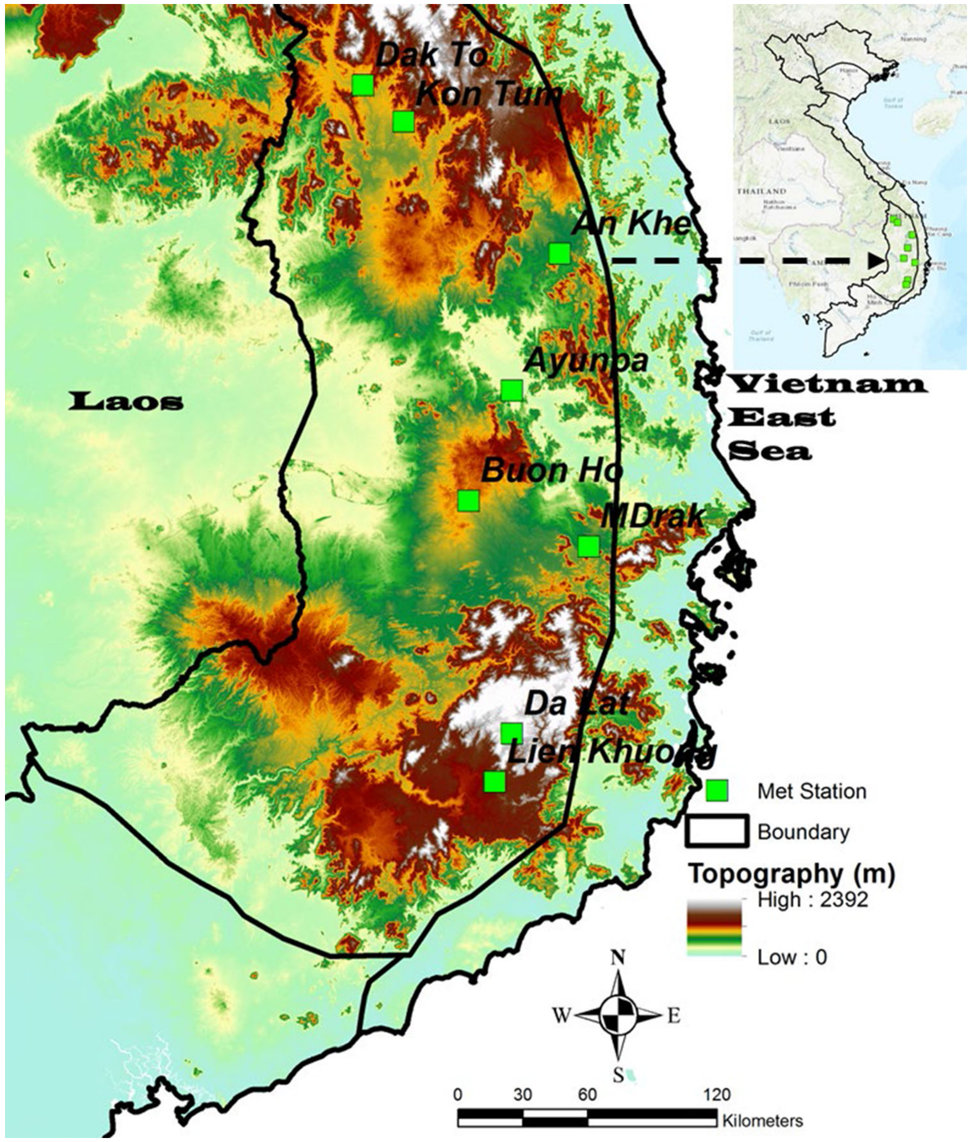


Fig. 1 Study area and rainfall station locations

distribution is also found to match the precipitation probability in the tropical climate region. More information on model formulation is available in Mhanna and Bauwens (2012).

3.2 RainSim V3 (RSIM)

RainSim Version 3 (*V3*) model was developed by Burton et al. (2008) based on the Spatial–Temporal Neyman–Scott Rectangular Pulse model (STNSRP) (Cowpertwait 1995, 2002). This model is capable of simulating rainfall for both single-site and spatially distributed regions of up to 200 km in diameter. The procedure applied to develop the NSRP model (Cowperwait 2002) is briefed discussed below:

1. The NSRP model applies a uniform Poisson process to simulate the storm origin with the occurrence rate represented by a parameter λ (Fig. 9(1) in the "Appendix").
2. At each storm origin, many rain cells are created independently based on a uniform Poisson process for different time intervals using parameter υ and β(Fig. 9(2)). For the STNSRP model, the rain cell generation at single-site model is replaced by a uniform Poisson process in space with density ρ and parameter γ. More information in this process can be found in Burton et al. (2008).
3. For each rain cell generated, a uniform rainfall rate is formulated with parameter duration η and intensity ξ (Fig. 9(3)).

4. Finally, the storm rainfall intensity is the summation of all intensities of all active rain cells at that time frame (Fig. 9(4)).

The overall procedure is provided in Fig. 9 (in the "Appendix") and a description of all seven parameters $\{\lambda, \upsilon, \beta, \rho, \gamma, \eta, \xi\}$ for the STNRSP model can be found in Table 1 in Burton et al. (2008).

The model has been calibrated for each calendar month using the log-parameter Shuffled Complex Evolution algorithm (Duan et al. 1993) and the scheme is built-in the Rainsim V3 package. Statistics include: mean, variance, skewness, lag-autocovariance, lag-autocorrelation, dry period probability, probability of dry–dry (or wet–wet), transition probabilities and the third order central moment, cross-covariances and correlations. The third order moment property (Cowpertwait 1998) is particularly important for applications where extreme rainfall events are important, such as flood risk assessment (e.g. Cowpertwait, 1995,1998,2002). ST-NSRP model can simulate both daily and hourly time scale. However, due to the lack of availability of sub-daily data sets, the hourly time scale is not considered. Detailed information can be found at Burton et al. (2008).

3.3 Perturbed K-nearest neighbor (pKNN model)

The perturbed non-parametric K nearest neighbor model was developed by modifying the original KNN model (Rajagopalan and Lall 1999). Instead of using the Euclidean distance approach for neighboring selection in the original model, the Mahalanobis distance approach was chosen as it does not require the explicit weighting of variables nor to standardize the variables (Yates et al. 2003). Additional information on KNN models and model development can be found in Rajagopalan and Lall (1999), Yates et al. (2003), Sharif and Burn (2007) and Gangopadhyay et al. (2005).

In the following section, we provide the steps used for constructing the perturbed KNN model. Assume that the data library has M stations, and it contains L variables with daily data for N years. The following steps are applied to search for the optimal subsequent day ($t + 1$) based on the array of weather variables on current day t.

(1) Compute the regional mean based on the L variables for M stations for the starting day t. In this study, the L variables consist of precipitation P and daily average temperature Tm. It is possible to include additional variable "s", such as wind speed, solar radiation or conditioning on the large-scale climate index (e.g., ENSO index) (Rajagopalan and Lall 1999; Yates et al. 2003). As the regional mean is computed from this step, a homogenous region (e.g., river basin) may be used to conform the spatial characteristics. This step results in a feature vector with only L members (mean P and Tm) for the selected day t.

(2) The next step is to identify the moving window length W. The moving average window W days were centered on the day t. Therefore, W/2 days prior to and after the central location t are selected. Based on the sensitivity analysis and literature review (Yates et al. 2003; Sharif and Burn 2007 and Gangopadhyay et al. 2005), we selected an optimal window length of W = 14 days (7 days lagged and 7 days lead). A total number of $(W + 1) \times N - 1$ pairs of vectors P and Tm for M stations are chosen as the potential candidates for the subsequent day $t + 1$.

(3) Compute the regional mean (P and Tm) of M stations based on the 374 pairs of the current day's weather in step 2. Correspondingly, we have 374 regional mean vectors.

(4) Compute a $L \times L$ covariant matrix S_t for the day t using the set of vector: $[(W + 1) \times N - 1] \times L$

(5) Randomly choose the weather on the first-day t (e.g. Jan 1st) consisting of all L variables, M stations from the set of all Jan 1st for N years. Using this as an initial step, the iteration was carried out to select the successive day $t + 1$.

(6) Subsequently, we calculated the Mahalanobis distance D_i (Eq. 4) between the 374 regional mean vector candidates of the current day weather vector $\bar{X}_t$ and the regional mean vector of the day t: $\bar{X}_i$:

$$D_i = \sqrt{(\bar{X}_t - \bar{X}_i)^T S_t^{-1} (\bar{X}_t - \bar{X}_i)} \tag{4}$$

In which: the $i = 1$ to $[(W + 1) \times N - 1]$. T is the transpose of the vector.

(7) Arrange the 374 D_i in increasing order, then select the first K shortest distances (Eq. 5). K is a variable, but a heuristic choice is normally chosen with referenced to Lall and Sharma (1996); Rajagopalan and Lall (1999); Yates et al. (2003):

$$K = \sqrt{(W + 1) \times N - 1} \tag{5}$$

Thus, K = 19 in this case.

(8) Compute a discrete kernel $p_{j=1\ldots k}$ in Eq. (6) to assign the probability at which the K-nearest neighbors are resampled:

$$p_{j=1\ldots k} = \frac{1/j}{\sum_{j=1}^{K} 1/j} \tag{6}$$

The resampling kernel $p_{j=1...k}$ is similar for any j and hence it can be computed and stored prior to the loop of simulation. The neighbor with the shortest distance gets the highest weight and the farthest gets the least. Detail information on the discrete kernel construction is discussed in Lall and Sharma (1996).

(9) The successive day $t + 1$ is selected from the suit of K nearest neighbor from the previous step. A random number r is selected ($0 < r < 1$) and if $p_1 < r < p_k$ then the successive day is selected for which r is nearest to p_j. If $r \leq p_1$ then the day corresponding to D_1 is chosen, else if $r \geq p_k$ the day corresponding to D_k is selected. Since this is a bootstrap approach, the simulations from the KNN model typically do not produce values that are not seen in the historical data (Rajagopalan and Lall 1999; Yates et al. 2003; Apipattanavis et al. 2007). In this study, we modified Sharif and Burn (2007) approach to perturb the successive $t + 1$ day value by adding a random component as described in the subsequent step.

(10) Stochastic weather generator for the successive day is achieved by applying a perturbation to the weather variables created from step 9 with a set of conditional standard deviation $\sigma_{M,L}$ and the kernel bandwidth λ specified by Sharma et al. (1998). The $\sigma_{M,L}$ is computed for each station and each weather variable obtained from the suit of step (3) and the kernel bandwidth λ can be simulated by using an Unbiased or Least Square Cross-Validation (LSCV) method (Bowman 1984; Rudemo 1982). The detailed procedure on estimating the bandwidth using the LSCV approach is available in Sharma et al. (1998).

The new value of weather variable L for day $t + 1$ for station M is computed by Eq. (7):

$$y^M_{L,t+1} = x^M_{L,t+1} + \lambda\sigma^M_L z_{t+1} \tag{7}$$

where $y^M_{L,t+1}$ is the new weather variable L for day $t + 1$ for station M perturbed from the $x^M_{L,t+1}$, which is the new value created from step 9; z_{t+1} is a random variate for day $t + 1$ obtained from a normal distribution with zero mean and unit variance. The new value for perturbed precipitation might lead to negative rainfall due to the bounded precipitation values. Therefore, we perturb the KNN output on day $t + 1$ conditioned on following criteria: (a) all M stations at day $t + 1$ are wet (to enable the effects of perturbation on wet days), (b) If there exists any station that is dry on day $t + 1$ and the mean rainfall value across all M stations is higher than 99th percentile of 374 mean values of M stations (to enable the effective of perturbation for the extreme values). Otherwise, no perturbation is needed for the day $t + 1$ and the iteration is continued with normal KNN procedure in order to reduce the computational time and to avoid the negative rainfall values. The two criteria ensure that the substantial large amount of synthetic rainfall days are perturbed. This also serves our purpose of preserving spatial extreme rainfall characteristics.

Repeat steps 6 through 10 based on a number of simulations needed for generating synthetic time series data. In order to simulate for multi-years, the iteration starts from step 5 instead. The perturbed KNN approach presented here simulate the regional mean precipitation values of 0 (mm/day) with exactly similar to original KNN. It does add significant contribution to perturb the non-zero regional precipitation for the study area, and therefore, substantially solve the limitation of original KNN by not creating any new values based on the historical data sets.

3.4 Statistical tests and goodness of fits

The stochastic models were used to generate 1250 data-years. The statistical characteristics of synthetically generated data were compared with the observed data in order to evaluate the performance of stochastic models (Semenov and Barrow 1997). The comparison was carried out at eight locations over Central Highland. Space–time comparisons were made between synthetic model outputs and observation in terms of precipitation occurrence, wet day intensity and cross-correlation. The wet day is defined as the day with total precipitation is higher or equal to 1 mm. The comparison was made using statistical measures based on following rainfall characteristics:

- Mean and Standard deviation of the number of wet days (rain day occurrence).
- Mean statistics of Maximum Consecutive Wet and Dry days.
- Mean, Standard deviation and Skewness of daily rainfall data.
- Transition occurrence Probability of 2 states: dry to wet (*P01*) and wet to wet (*P11*).

In term of spatial analyses, the following statistical test was carried out.

- The Joint Probability that station pairs are both wet (*JP11*) and both dry (*JP00*).
- Cross-correlation of precipitation occurrence and amount between stations.

In addition to cross-correlation between station pairs, we also computed the conditional log-odds ratio, which reflects the spatial correlation between rainfall occurrences at each pair of stations, which provides a measure of such dependency for binary variables. According to Mehrotra et al. (2006), the log-odds ratio is defined as in Eq. (8):

$$lr_{i,j} = \log\frac{p00_{i,j}p11_{i,j}}{p10_{i,j}p01_{i,j}} \quad (8)$$

where $lr_{i,j}$ is the log-odds ratio between i and j pair of stations; $p00_{i,j}, p01_{i,j}, p10_{i,j}, p11_{i,j}$ are the joint probabilities of wet/dry for station pairs with *0* denotes dry and *1* denotes wet day. The high values of the log-odds ratio are indicative of better spatial dependence between the variables.

Finally, we also evaluated the abilities of stochastic rainfall models in constructing the Intensity–Duration–Frequency curves (IDF) based on the annual maximum daily precipitation (AMR). In this study, Generalized Extreme Value distribution was used to derive frequency components embedded in IDF curves. This analysis can be helpful to compare the simulation of extreme rainfall events over the 1000 years of simulation.

3.5 Construction of IDF curves

The GEV model best approximates the distribution of the annual maximum over a finite range of return period intensities (Veneziano and Yoon 2013). The cumulative function of GEV distribution is in Eq. (9):

$$F(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \quad (9)$$

where μ, σ, ξ are shape, location and scale parameters respectively. For $\xi = 0$, Eq. (9) reduces to the Gumbel distribution (Extreme Value type I). Common estimation procedures for μ, σ, ξ are the method of moments (Madsen et al. 1997), method of L-moment (Hosking 1990), method of probability-weighted moments—PWM (Hosking et al. 1985) and maximum likelihood method—ML (Smith 1985). Even though the PWM or L-moment are more popular than ML in application to hydrologic extremes (because of computational simplicity and good performances over small samples), these techniques have the disadvantage of not being able to readily incorporate covariates (Katz et al. 2002). Therefore in this study, we applied the ML estimation approach due to its simplicity to find μ, σ, ξ suggested by Katz et al. (2002) after their detailed discussion on each of the estimation methods. For the model outputs, we computed the return period of the simulated annual maximum daily rainfall from stochastical models using Weibull's formula in Eq. (10):

$$T = \frac{n+1}{m} \quad (10)$$

in which n is the total number of values to be plotted and m is the rank value (sorted from largest to smallest). The reverse of the return period T is the exceedance probability of the mth largest value.

4 Results

4.1 Simulation of precipitation occurrence

Precipitation Occurrence (state *1* of whether the day is wet or state *0* if dry) is represented by wet day statistics (mean & standard deviation) and the maximum consecutive Wet/Dry days of the month. Comparisons were made by using scatter plots between three model outputs and observed data based on monthly average statistics.

Figure 2a, b illustrates the scatter plots between the observed mean and standard deviation of the number of wet days per month and the models respectively. The correlation coefficients R and root mean square errors *RMSE* between stochastic models and observation for the analyses are provided in Table 1. Figure 2a indicates that the studied region has distinct wet/dry months with the mean number of wet days per month range from 0 day (completely dry month) to 25 days (very wet month). The behavior of the three stochastic precipitation generation models are slightly difference from observation. The Markov chain-based model (modWilks) and non-parametric pKNN performed well in generating the mean occurrence of wet days at R = 1 (rounded to 2 decimals). The precipitation occurrence result obtained by the pKNN model is slightly better than modWilks at higher tail (catching the month with more wet days) during the rainy season, hence leading to smaller RMSE for the mean statistics (0.59 versus 0.82). The RSIM model is able to

Table 1 Goodness of fits for SWG model simulation versus observation: precipitation occurrence statistics

Model\Stat	Mean		Std		Max Wet days		Max Dry days	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE
ModWilks	1.00	0.82	0.93	0.79	0.97	0.58	0.97	2.18
RSIM	0.93	2.71	0.83	0.90	0.63	1.04	0.95	2.67
pKNN	1.00	0.59	0.90	0.79	0.92	0.51	0.99	1.36

All correlation values in this table are having p value < 0.05

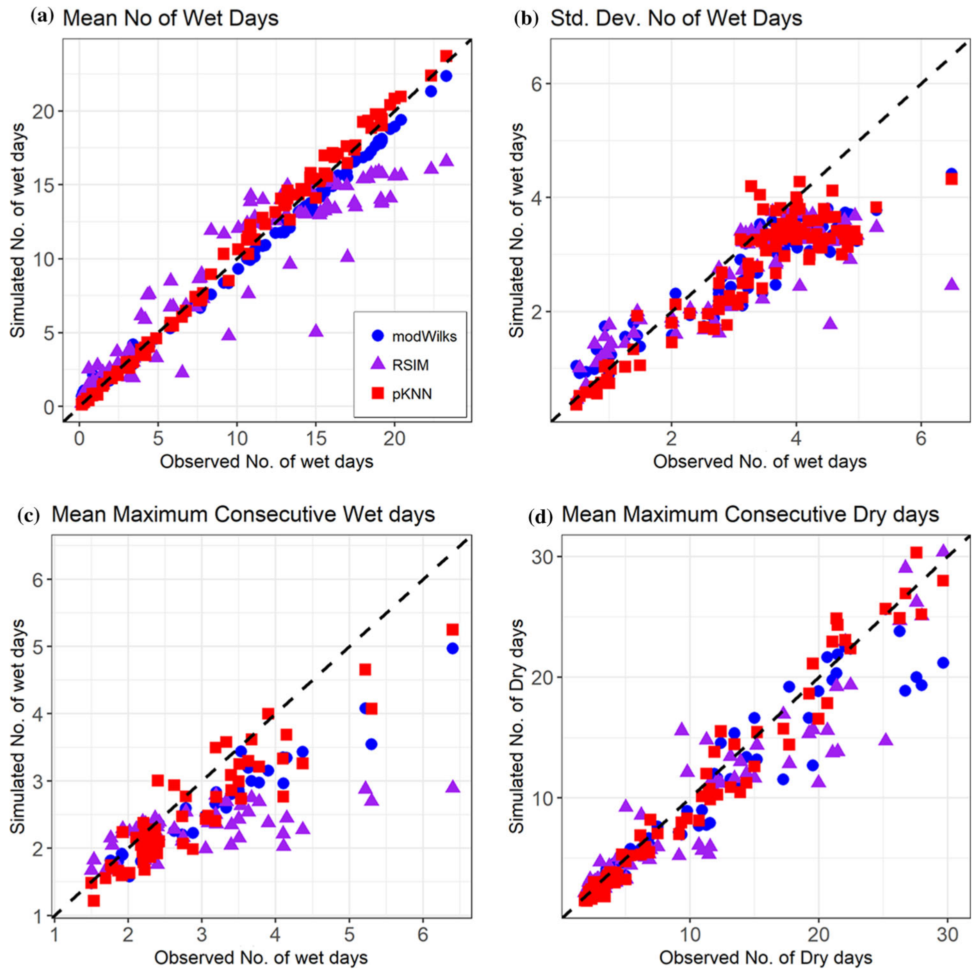


Fig. 2 Scatter plots between observed and synthetic precipitation occurrence statistics over all stations and all months: **a** Mean number of wet days; **b** Standard deviation number of wet days; **c** Mean Maximum Consecutive wet days and **d** Mean Maximum Consecutive dry days

generate the mean number of wet days for the dry season (less than 5 wet days a month). However, this model underestimates the mean number of wet days statistics for wet season (months having more than 15 wet days), leading to smaller correlation and higher RMSE compared to the other two models. This has similar finding with Vallam and Qin (2016) for multi-site synthetic rainfall generation over the nearby country Singapore. However, RSIM still performs well with R-value higher than 0.9 at 95% confidence interval. In a similar manner to mean statistics, modWilks and pKNN perform well for the standard deviations of the number of wet days per month with both $R > 0.9$ and RMSE = 0.79 (Table 1). RSIM relatively has a lower correlation index (R = 0.83) and slightly higher RMSE (at 0.9). All three models are able to simulate well the standard deviation of wet days with high correlation values ($R > 0.8$), it is indicated by Mehrotra et al. (2006) that the spatial and long-term temporal dependence for multi-stations are preserved.

Apart from the statistics of the mean and standard deviation, the maximum consecutive wet and dry day plays an important role in the prolonged extreme conditions that often result in drought and flood events. Hence, this information needs to be validated while using synthetic data from the models. The monthly average of maximum consecutive wet and dry days scatter plots are displayed in Fig. 2c, d, respectively, the R and RMSE are tabulated in Table 1. The max consecutive wet days range from 1 to 6 days whilst the similar statistics for dry days range from 1 to 30 days, which significantly portraits the distinct wet/dry season for the Central Highland region in Vietnam, especially during dry season where some of the whole months are completely dry. All models are able to simulate well the mean of max consecutive dry days statistics with correlation higher than 0.9. RSIM unexpectedly captures well some of the months with 28–30 dry days (Fig. 2d). However, for the mean of max consecutive wet days statistics, only modWilks and pKNN are able to retain the significant correlations (R equals to 0.97 and 0.92, respectively), while RSIM underestimates this characteristic (R = 0.63). This has similar pattern compared to the mean number of wet days statistics that we analysed priory. Overall, the pKNN model performed best with the lowest RMSE values for both categories of extreme wet/dry periods and maintaining statistically significant correlation statistics, followed by modWilks model.

To summarize, for precipitation occurrence simulation at an individual site, pKNN is the best model among the three with all statistical tests of the wet day occurrence and max consecutive wet/dry days. The Markov chain-based model modWilks also reproduced well the statistics for the occurrence of wet days, whereas, RSIM model comparatively performed less satisfaction compared to other models.

4.2 Simulation of Precipitation Amount

The stochastic process associated with the pKNN model is able to perform well to reproduce the precipitation occurrence, however, it is important to identify which model accurately generates the precipitation amount. Precipitation daily intensity (amount of precipitation on daily basis) statistics simulated from three stochastic models for eight stations are displayed in scatter plot in Fig. 3 and Table 2. The statistical test based on the mean value of precipitation intensity (for both wet/dry days per month and wet days only), standard deviation and skewness statistics are also examined to evaluate the variability of the precipitation distribution for selected locations.

In comparison to precipitation occurrence, all models perform well in generating the precipitation intensity with correlation value close to 1 (rounded to 2 decimals; as seen in Table 2). However, pKNN slightly overestimated the precipitation amount (R = 0.99 and having largest RMSE value, Table 2) at high tail as portraited in Fig. 3a, this is particularly because the perturbation parameters that were introduced in Eq. (7), which takes into account the all wet days or extreme wet event. The gamma distribution fitted to the modWilks model is slightly better compared to the other models with the lowest value of RMSE (0.22) observed at all the eight stations. RSIM is also having a great match in generating daily precipitation intensity with perfect score for correlation and relatively small RMSE, compared to pKNN. However, in the statistic for the mean of wet day precipitation intensity, the pKNN model performs well based on the lowest RMSE (RMSE = 1.08) with a very high significant correlation coefficient value (R = 0.97) for the same category. Despite being a good model for generating daily precipitation intensity, the RSIM model comparatively did not perform well in

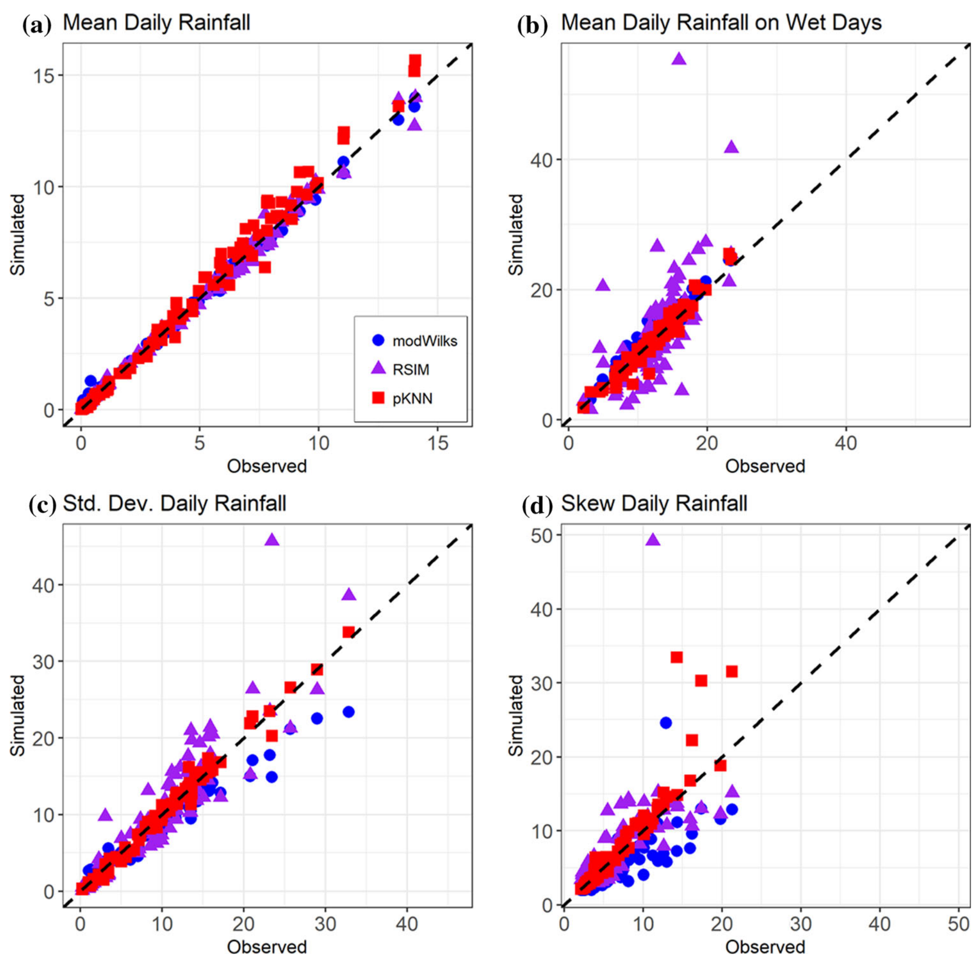


Fig. 3 Scatter plots between observed and synthetic precipitation amount statistics over all stations and all months: **a** Mean daily rainfall; **b** Mean daily rainfall on wet days; **c** Standard deviation of daily rainfall and **d** Skew coefficient of daily rainfall

Table 2 Goodness of fits for SWG model simulation versus observation: precipitation amount statistics

Model\Stat P_amt	Mean		Mean Wet days		Std		Skew	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE
modWilks	1.00	0.22	0.98	1.31	0.98	2.28	0.80	2.96
RSIM	1.00	0.29	0.64	6.11	0.90	3.50	0.62	4.53
pKNN	0.99	0.57	0.97	1.08	0.99	0.80	0.93	2.76

All correlation values in this table are having p value < 0.05

synthesizing this characteristic for wet days as observed based on the R-value (0.64). The scatter plot in Fig. 3b indicates the overestimation of RSIM when it doubles the mean rainfall on wet days for certain extreme events. This is because RSIM underestimates the mean number of wet days statistics for wet season when computing the precipitation occurrence statistic as depicted in Fig. 2a.

The modWilks slightly underestimates the standard deviation of daily precipitation at the higher tail, implicitly inferring the underestimation of low-frequency variability for precipitation. As depicted in Table 2, modWilks comparatively performed better than the RSIM model based on its smaller RMSE and higher R-value. Overall, the pKNN model proves to be the best model in preserving the standard deviation based on the goodness of fit statistics, followed by the modWilks model. The skewness coefficient was used to inspect the distribution of precipitation. The skewness coefficients between models and observations are slightly more scattered compared to the standard deviation. Based on the Goodness of fit test (Table 2), the pKNN model performed well in terms of skewness, followed by modWilks and RSIM.

4.3 Spatial characteristics

4.3.1 Multisite conditional probabilities of rainfall occurrence

The occurrence of rainfall for the current date is called conditional if the previous day's wet/dry state is known. The transition conditional probability of rainfall occurrence state from state dry to wet *p01* and state wet to wet *p11* are investigated in this study. The transition states were computed individually for all the stations and they are displayed using scatter plots (Fig. 4) and goodness of fit tests are tabulated in Table 3. In modWilks model, the transition probabilities were simulated using two-state first-order Markov chain processes, so it has a near-perfect correlation for *p01* and *p11* transition with observed data (R = 1.00) and low RMSE (RMSE = 0.01 and 0.04, respectively). The pKNN model has a comparatively lower correlation (R = 0.99 and 0.96) and higher RMSE values (RMSE = 0.05 and 0.09) due to the overestimation of the mid to high tail of *p01* and underestimation of the low to mid-tail *p11*. Although the RSIM model exhibits slightly lower correlation coefficient (R = 0.96/0.86 for p01/p11) and slightly higher RMSE for both transition states (RMSE = 0.05 and 0.11), it is considered successful in reproducing the two transition states probability when comparing with the study by Baigorria and Jones (2010) for GIST model with R = 0.981 and 0.836 for *p01* and *p11*, respectively. Baigorria and Jones (2010) also mentioned that due to the constrain to satisfy both the spatial correlations and first-order-in-time transition probability that the stochastic model might not preserve well this characteristic.

The above analysis provided the performance of stochastic models for the transition probabilities of precipitation occurrence based on the individual stations. In the following section, the transition probabilities are computed based on a number of n stations used in the network. Overall, there are $n(n-1)/2$ pair of stations are evaluated based on 12 months. It is necessary to evaluate the spatial relationship between all the stations using joint probability (*JP)* analysis. Figure 5 illustrates the scattered plots for the joint probability of the transition of precipitation occurrence for the eight stations based on the three stochastic models and observed information. The comparisons based on the joint distribution of simultaneous precipitation occurrence for 336 pairwise location-month combinations are computed and plotted for the state of (a) both stations are wet *JP11*, and (b) both stations are dry *JP00*. Similar to transition probability characteristic, the joint probability correlations (*JP11* and *JP00*) for modWilks and pKNN models are near perfect with correlation coefficient approximated to 1 (Table 3), and the RMSE values are closer to zero. RSIM has slightly lower R value (R = 0.95) and higher RMSE, however, it is in comparable favor with the study by Baigorria and Jones (2010) with their R-values for JP01 and JP11 of 0.98. Overall, it is observed that the performances of retaining spatial relationships of three models are great for both rainy and dry seasons.

Figure 6 illustrates the observed and synthetic conditional log-odds ratios for three models. This is a complementary index for the transition probability and joint probability characteristics. The log-odd ratio demonstrates

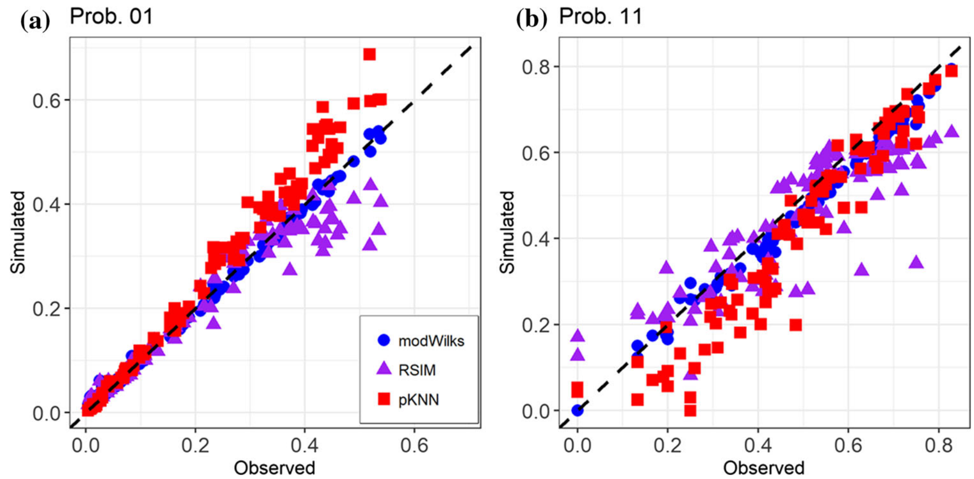


Fig. 4 Scatter plots between observed and synthetic transition occurrence probability over all stations and all months: **a** dry–wet **b** wet–wet

Table 3 Goodness of fits for SWG model simulation versus observation: transaction probability and joint probability

Model\Stat Pmarkov	p01		p11		JP11		JP00	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE
modWilks	1.00	0.01	1.00	0.04	1.00	0.02	1.00	0.03
RSIM	0.96	0.05	0.86	0.11	0.94	0.06	0.95	0.09
pKNN	0.99	0.05	0.96	0.09	0.99	0.03	1.00	0.02

All correlation values in this table are having p value < 0.05

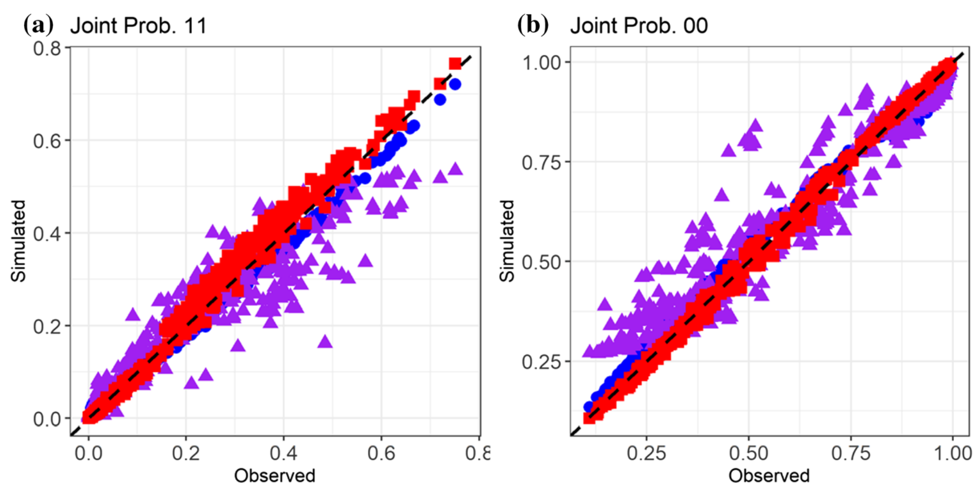


Fig. 5 Scatter plots between observed and synthetic transition occurrence of joint probability that station pairs are both **a** wet **b** dry, on a given day for all months computation considering 336 combinations of station pairs

the spatial correlation between the precipitation occurrence at each station pair, in which, the higher ratio indicates the better spatial dependence. For example, the neighbor stations within the same topology normally have higher log-odd ratio than the other locations. Each point on the scatter plot indicates the ratio evaluated for a pair of rain gauge stations (Mehrotra et al. 2006) for an individual month. It was observed that the log-odds ratio is captured less accurately (R = 0.7) by RSIM due to the widespread of points over the scatter plot. However, the modWilks model performed better compared to the RSIM model based on the scattered points located closer to the 1:1 line (R = 0.88). Overall, the pKNN model performed well based on the highest correlation coefficient (R = 0.96), which is able to capture the occurrence of precipitation simultaneously at all the stations.

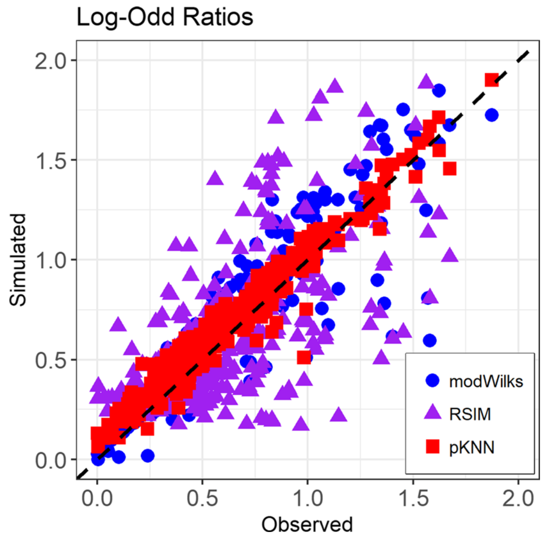


Fig. 6 Observed and model simulated Log-Odd ratios for all station pairs

4.3.2 Cross-correlation analysis

The cross-correlation analyses between multi-stations were computed for both precipitation occurrence and amount. These analyses clearly reveal the deficiency in both spatial precipitation occurrence and the amount for the models as compared to the joint probability study. Figure 7a, b demonstrates the cross-correlation of precipitation occurrence and amount, respectively for eight stations using 12 months, which resulted in 336 pairwise combinations. The goodness of fit tests (*R* and *RMSE*) for correlation analysis are provided in Table 4. In previous studies, the modWilks model performed well in simulating spatial cross-correlation patterns (Mhanna and Bauwens 2012; Wilks 1998). In our analysis, the modWilks model is to capture the cross-correlation of precipitation occurrence

Table 4 Goodness of fits for SWG model simulation versus observation: cross correlation characteristic

Model\Stat mon	Occ. Cross Cor		Amt. Cross Cor	
	R	RMSE	R	RMSE
modWilks	0.95	0.07	0.89	0.12
RSIM	0.77	0.12	0.77	0.14
pKNN	0.97	0.04	0.98	0.04

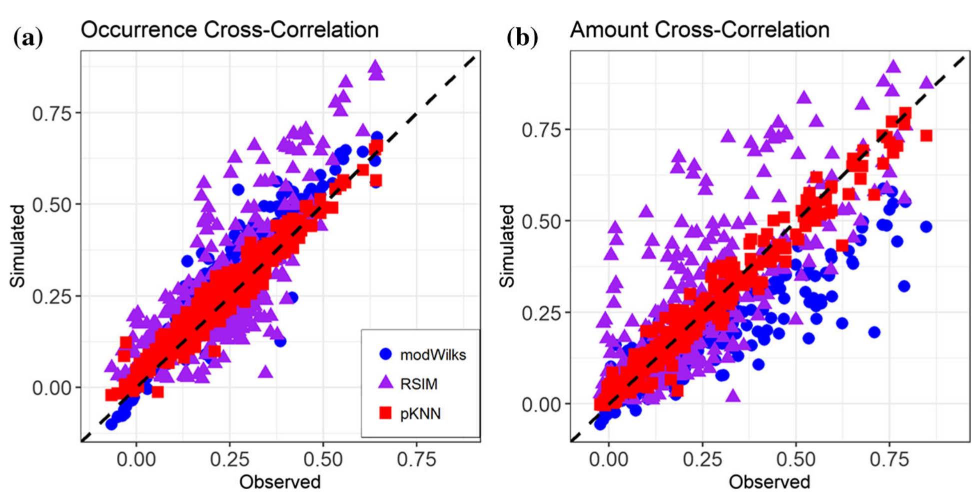
All correlation values in this table are having p value < 0.05

slightly better than precipitation amount. On the other hand, the modWilks model is comparatively better than the RSIM model based on the goodness of fit indices (Table 4). However, the pKNN model found to be the best model with significantly low RMSE values (0.04) and high correlation, especially for rainfall amount (R = 0.98). The original KNN model has been known to well preserve the cross-correlation structure of precipitation occurrence and amount between all the stations (Yates et al. 2003; Sharif and Burn 2007). We also compared (not shown) the cross-correlation performance between pKNN and original KNN models (without perturbation), it was observed that the original KNN performed slightly better than pKNN, due to the non-perturbation approach applied.

4.4 Intensity–duration–frequency analysis

The IDF curves play an important role in civil infrastructure systems design (Veneziano and Furcolo 2002; Vu and Mishra 2019). Therefore, it is an important aspect to investigate the performance of stochastic models in reproducing the extreme rainfall pattern, especially for the purposes of generating rainfall-induced flood. Previous studies did not investigate the performance of the stochastic models for constructing IDF curves. In this

Fig. 7 Scatter plots between observed and synthetic cross correlation for all station pairs and months **a** precipitation occurrences **b** precipitation amount



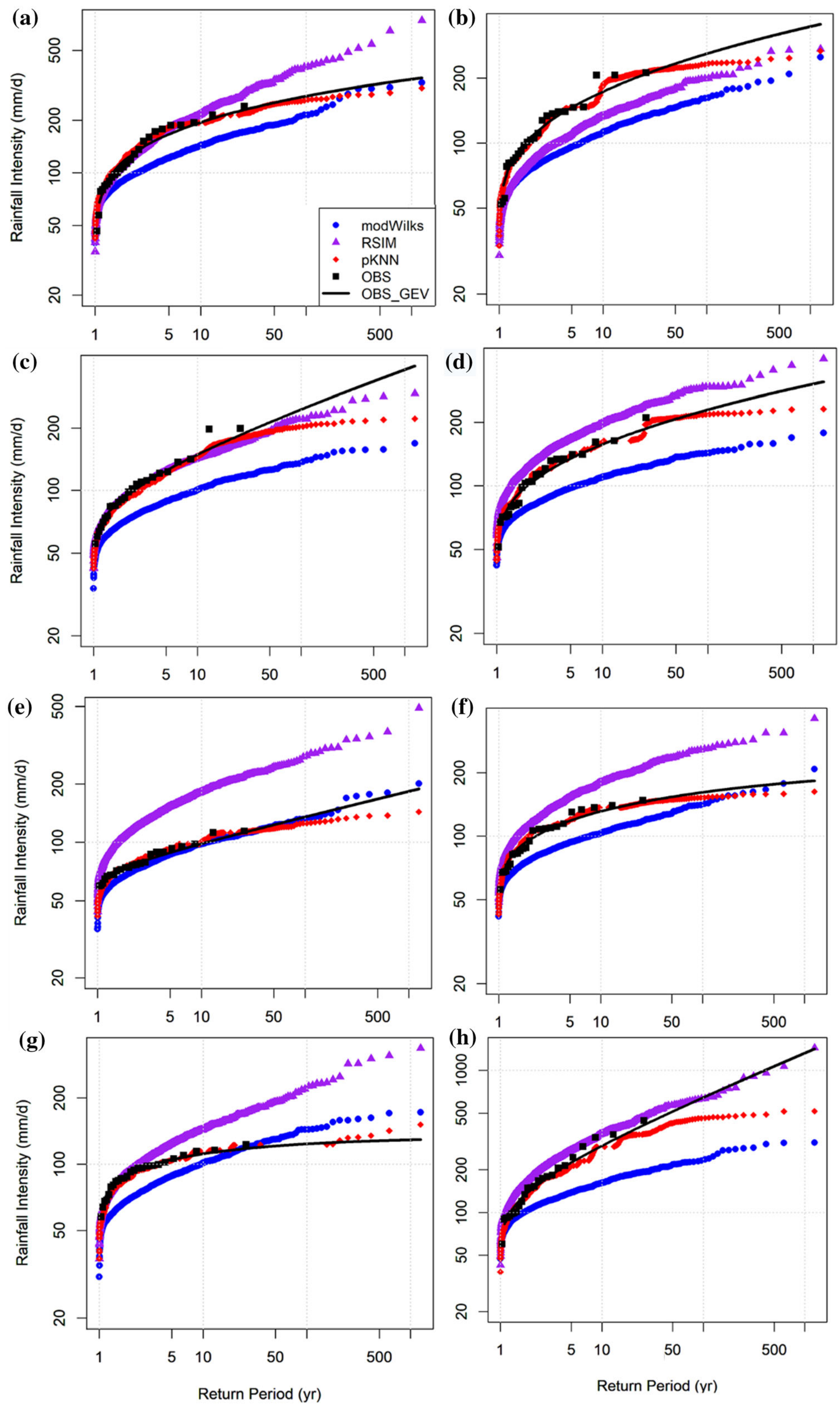


Fig. 8 Intensity Duration Frequency (IDF) curves derived based on the Annual Maximum Daily synthetic precipitation data sets for multisite locations: **a** An Khe, **b** Ayunpa, **c** Buon Ho, **d** Dakto, **e** Da Lat, **f** Kon Tum, **g** Lien Khuong, **h** M'Drak. The black squares represent IDF derived based on the 25 years of observed ('OBS') data. The return periods obtained from three stochastic model outputs are computed using Eq. (10) and plotted with different markers and colors

study, the IDF curves based on the return periods up to 1000 years are generated to investigate the capabilities of stochastic precipitation models for generating the extreme annual precipitation characteristics. The IDF curve for individual station is displayed in Fig. 8. The black squares represent IDF derived based on the 25 years of observed

('OBS') data. The return periods obtained from three stochastic model outputs are computed using Eq. (10) and plotted with different markers and colors.

The three models produced different extreme values distribution for 1000 years simulation and are compared with the GEV fitting to OBS data. For GEV fitting, a small sample size sometime makes the estimation of the GEV shape parameter rather erratic and sensitive to outliers (Veneziano et al. 2007). Although the sample size of 25 years is not considered sufficiently large, the models are partially able to capture the extreme precipitation intensity at 1000 years event. For example, pKNN is able to capture the IDF curve for Lien Khuong station (Fig. 8g) while RSIM can approximate the 1000 return period of M'Drak station (Fig. 8h). In six out of eight stations, RSIM overestimates the GEV fitting whilst modWilks shows underestimation. Only pKNN is able to capture well the GEV fitting at a return period of less than 50 years. This can be explained that because gamma distribution was applied to fit the daily precipitation into the modWilks model, this distribution type tends to underestimate heavy rainfall events (Wilson and Toumi 2005; Cho et al. 2004). On the other hand, due to the aggregation of rainfall intensities that RSIM usually overestimates precipitation (Vallam and Qin 2016). The perturbed KNN performs a nonparametric approach by shuffling the existing data, therefore it follows closely the observed data patterns, and is explained why it best matches the GEV distribution for the return period less than 50 years. It opens the way for further discussion by including the hybrid model in synthesizing the data (Chao et al. 2012; Furrer and Katz 2008).

5 Discussion

The performance of three stochastic rainfall models are compared based on different precipitation characteristics, such as precipitation occurrence, wet/dry spell and precipitation on wet day intensity, cross-correlation, transition probability, log-odds ratio, as well as extreme value analyses, are shown in Table 5. The choice of selecting the precipitation characteristic for multisite stochastic rainfall generation models are based on the previous literatures but not limited to Mehrotra et al. (2006), Vallam and Qin (2016), Baigorria and Jones (2010). The model performance was compared against precipitation characteristics based on the ranks in the order of highest (1) to lowest (3) (Vu et al. 2018). Thus, the model that has performed well in each metric will lead to the smallest summation of the rankings that can be considered as the best model. In this study, the choice of selecting the best multisite precipitation generation model is limited to the Central Highland region in Vietnam.

In our study, RSIM ranked lowest compared to other two models. Previous studies found RSIM may perform well for the rainfall stations located in specific precipitation regime. For example, Forsythe et al. (2014) suggested that RSIM can be used to successfully produce synthetic rainfall time-series in a semi-arid mountainous climate in Pakistan. On the other hand, Vallam and Qin (2016) indicated that RSIM usually overestimated the precipitation amount for wet seasons while comparing different models for several stations across Singapore. Their finding is comparable with our result that RSIM is not as good when simulating mean precipitation on wet days compared to pKNN and modWilks, even though that RSIM performs really well over mean daily precipitation (averaging for both wet and dry days), as thus it underestimates the wet day occurrences with threshold of 1 mm/day. Both modWilks and pKNN models are comparable in terms of statistical tests for precipitation occurrence and amount at individual stations. All models can simulate the spatial correlation between stations for standard deviation of wet days statistics, which showcases that the spatial and long-term dependence are preserved. In this study, the modWilks model seems to be a good model for simulating precipitation at both spatial and temporal scales. The addition of the gamma coefficient and rank correlation to preserve the correlation of random numbers applied to rainfall occurrence and amount clearly improved the Spatio-temporal correlation pattern. In addition, the common choice of gamma distribution has been applied to modWilks to fit rainfall amount is more appropriate for tropical monsoon rainfall compared to the mixed exponential distribution as recommended by Wilks (1998). The pKNN model outperforms other two models in terms of preserving the spatial characteristics such as cross-correlation and log-odds ratios. In addition to that pKNN model performed well in capturing the IDF curves while modWilks tend to underestimate heavy rainfall events and RSIM overestimates rainfall intensity. The improved performance shown by pKNN model can be attributed to the perturbation approach, that incorporate the kernel bandwidth estimation using LSCV method. In a nutshell, based on the aggregated ranking, the pKNN model is found to be the best model followed by modWilks and RSIM models. However, the limitation of the pKNN model is that it offers a simple modeling mechanism, but being a non-parametric model, it is computationally expensive in terms of simulation time (Mehrotra and Sharma 2009). Comparatively, modWilks and RSIM models are much faster in terms of computation time. Therefore, by performing comparative ranking analysis the stakeholders can choose the best model suitable for specific locations and weather patterns.

One of the key limitations that exist in SWG models is lack of simulating the extreme event (i.e. 5000–10,000-

Table 5 Overall assessment of the multisite SWG models using different rainfall characteristics

Rainfall characteristics	modWilks	RainSim	pKNN
Mean number of wet day	2	3	1
Standard deviation number of wet day	1	3	2
Mean maximum consecutive wet day	1	3	2
Mean maximum consecutive dry day	2	3	1
Mean daily rainfall amount	1	2	3
Mean daily rainfall amount (wet day only)	2	3	1
Standard deviation rainfall amount	2	3	1
Skewness rainfall amount	2	3	1
Transition occurrence probability from dry to wet p01	1	3	2
Transition occurrence probability from wet to wet p11	1	3	2
Joint Transition probability from wet to wet JP11	1	3	2
Joint Transition probability from dry to dry JP00	2	3	1
Cross correlation of rainfall occurrence	2	3	1
Cross correlation of rainfall amount	2	3	1
Log-odds Ratio	2	3	1
Extreme value analysis: IDF curves	2	3	1
Overall Performances	26	47	23

The best model rank 1 and the second best rank 2 and so on

year return period). In this study, similar observation was made when simulating the high tail distribution of the annual extreme event. This may be due to limited data length used in the study or the limitation of several stochastic models in simulating the full range of precipitation from extreme dry to extreme wet scale. To overcome such limitations, Chao et al. (2012) and Furrer and Katz (2008) performed simulation based on the entire range of precipitation using a hybrid approach that integrate gamma/exponential distribution for the lower to medium range whilst generalized Pareto distribution could be applied to the higher precipitation range. Our study can be further improved to address such limitation by developing the hybrid perturbed KNN modeling, which has the ability to synthesize the extreme precipitation events as well as capture the extreme events at higher return periods (e.g., 10,000-year) for revising the IDF curves.

6 Conclusions

We evaluated the performance of three multisite stochastic precipitation generator models (i.e., modified Wilks, perturbed KNN and RainSim) in a tropical monsoon climate region located in Central Highland, Vietnam. These three models consist of a parametric Markovian model (modified Wilks) and two non-parametric (perturbed K Nearest Neighbor and Spatial–Temporal Neyman Scott Rectangular Pulse RainSim) models. In this study, we modified the traditional KNN model by revising the kernel bandwidth used in the model development. The modified KNN model improved the spatial coherence structure of rainfall stations across the study region. The models were evaluated based on their performance to capture different precipitation characteristics. The challenging task is to find an optimal model that is able to generate the multisite precipitation characteristics which are useful for drought and flood management over the study region. The following conclusions can be drawn from this study:

1. *Precipitation occurrence* Three stochastic models are able to capture the mean and standard deviation for the number of wet days. However, pKNN and modWilks models have slightly better performance compared to RSIM. All the selected models performed well for simulating the maximum consecutive number of dry days per month. In terms of the maximum consecutive number of wet days, the RSIM model underestimated the observation, therefore, indicating a lower correlation coefficient compared to the other two models.
2. *Rainfall intensity* Even though all three models performed very well for simulating daily mean rainfall intensity, the performance of the RSIM model suffers when simulating daily mean rainfall intensity on the wet day only. This may be due to the relatively less performance of the RSIM model when simulating the precipitation occurrence. Based on the comparison for standard deviation and skewness for wet days intensity, the modWilks model performed reasonably well. Overall, the pKNN model found to be the best model

based on improved goodness of fits (correlation coefficient and RMSE).

3. *Transition probability analysis* The transition from wet to dry day and vice versa are evaluated based on comparison with occurrence probability and joint probability between stations. Due to the inclusion of Markov chain formulation, the parametric modWilks model found to be the perfect model with the absolute correlation. Both non-parametric models (pKNN and RSim) also performed well based on goodness of fit analysis.
4. The most important feature of the multisite rainfall model is the preservation of spatial correlation between multiple stations. A cross-correlation analysis was conducted based on precipitation occurrence, rainfall intensity and log-odds ratios. The modWilks model overcame the limitation associated with the original Wilks model in preserving the spatial correlation for precipitation occurrence and rainfall intensity respectively. Overall, the pKNN model performed better compared to modWilks and RainSim model based on better goodness of fit statistics.
5. The IDF curves were derived and compared between observed data and the synthetic data sets. In this study, the GEV distribution was applied to conduct frequency analysis embedded in constructing IDF curves. It was observed that the RSIM model overestimates half of the stations' IDF curves while modWilks underestimates 7 out of 8 stations. Only the pKNN model is able to capture the IDF curves for the return periods of less than 50 years. This is a key feature of the pKNN model, and it is recommended to include more observed data sets to better capture the higher tail of IDF curves.

In this study, even though the pKNN model comparatively performed well for multisite precipitation simulation over the Central Highland area located in Vietnam, it is, however, computationally expensive, especially due to the addition of the perturbation component. RainSim model may not be a good choice for the tropical monsoon region in Central Highland area in Vietnam, however, it performed reasonably well as a multisite stochastic rainfall generation model for European regions, and it can be further explored in other climatic regimes.

Acknowledgements Authors would like to thank the Risk Engineering and Systems Analytics Center, Clemson University, the Palmetto supercomputer for providing computing resources.

Compliance with ethical standards

Conflict of interest We have no conflict of interest to report.

Appendix

See Fig. 9.

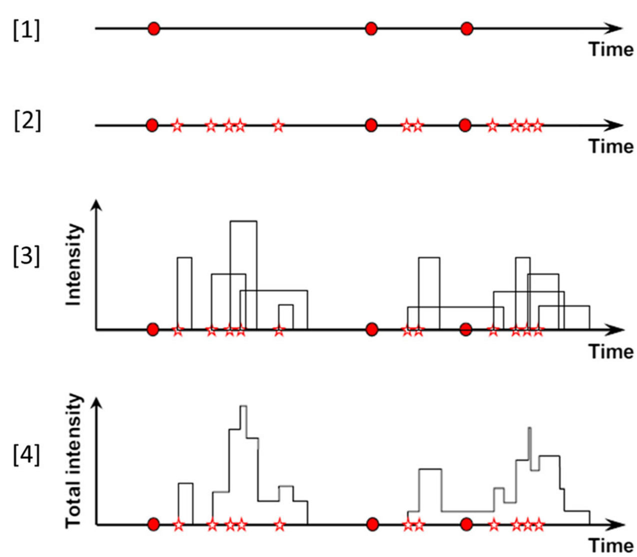


Fig. 9 Neyman Scott Rectangular Pulse model. (Adapted from Burton et al. 2008)

References

Agilan V, Umamahesh NV (2019) Rainfall generator for nonstationary extreme rainfall condition. J Hydrol Eng 24(9):04019027. https://doi.org/10.1061/(asce)he.1943-5584.0001821

Apipattanavis S, Podestá G, Rajagopalan B, Katz RW (2007) A semiparametric multivariate and multisite weather generator. Water Resour Res 43:W11401. https://doi.org/10.1029/2006WR005714

Aryal K, Thapa PK, Lamichhane D (2019) Revisiting agroforestry for building climate resilient communities: a case of package-based integrated agroforestry practices in Nepal. Emerg Sci J 3(5):303–311

Baigorria GA, Jones JW (2010) GiST: a stochastic model for generating spatially and temporally correlated daily rainfall data. J Clim 23:5990–6008. https://doi.org/10.1175/2010JCLI3537.1

Bardossy A, Plate EJ (1992) Space–time model for daily rainfall using atmospheric circulation patterns. Water Resour Res 285:1247–1259. https://doi.org/10.1029/91WR02589

Beersma JJ, Buishand TA (2003) Multisite simulation of daily precipitation and temperature conditional on the atmospheric circulation. Clim Res 25:121–133. https://doi.org/10.3354/cr025121

Bogardi I, Matyasovszky I, Bardossy A, Duckstein L (1993) Application of space–time stochastic model for daily precipitation using atmospheric circulation patterns. J Geophys Res 98(D6):16653–16667. https://doi.org/10.1029/93JD00919

Bordoy R, Burlando P (2014) Stochastic downscaling of precipitation to high-resolution scenarios in orographically complex regions: 1. Model evaluation. Water Resour Res 50:540–561. https://doi.org/10.1002/2012WR013289

Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71(2):353–360. https://doi.org/10.2307/2336252

Brissette FP, Khalili M, Leconte R (2007) Efficient stochastic generation of multisite synthetic precipitation data. J Hydrol 345:121–133. https://doi.org/10.1016/j.jhydrol.2007.06.035

Buishand TA, Brandsma T (2001) Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. Water Resour Res 37:2761–2776. https://doi.org/10.1029/2001WR000291

Burton A, Kilsby CG, Fowler HJ, Cowpertwait PSP, O'Connell PE (2008) RainSim: a spatial-temporal stochastic rainfall modelling system. Environ Model Softw 23:1356–1369. https://doi.org/10.1016/j.envsoft.2008.04.003

Chao L, Singh VP, Mishra AK (2012) Simulation of the entire range of daily precipitation using a hybrid probability distribution. Water Resour Res 48:W03521. https://doi.org/10.1029/2011WR011446

Chen J, Brissette FP, Zhang XJ (2014) A multisite stochastic weather generator for daily precipitation and temperature. Trans ASABE 57:1375–1391. https://doi.org/10.13031/trans.57.10685

Cho HK, Bowman KP, North GR (2004) A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. J Appl Meteorol 43:1586–1597. https://doi.org/10.1175/JAM2165.1

Cowpertwait PSP (1995) A generalized spatial-temporal model of rainfall based on a clustered point process. Proc R Soc Lond Ser A 450:163–175. https://doi.org/10.1098/rspa.1995.0077

Cowpertwait PSP (1998) A Poisson-cluster model of rainfall: high-order moments and extreme values. Proc R Soc Lond Ser A 454:885–898. https://doi.org/10.1098/rspa.1998.0191

Cowpertwait PSP, Kilsby CG, O'Connell PE (2002) A space–time Neyman–Scott model of rainfall: empirical analysis of extremes. Water Resour Res 38(8):1131. https://doi.org/10.1029/2001WR000709

Dao DC, Nguyen TS, Nguyen QM, Nguyen TT, Tran TD (2020) An analysis of shoreline changes using combined multitemporal remote sensing and digital evaluation model. Civ Eng J 6(1):1–10. https://doi.org/10.28991/cej-2020-03091448

Devak M, Dhanya CT, Gosain AK (2015) Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall. J Hydrol 525:286–301. https://doi.org/10.1016/j.jhydrol.2015.03.051

Duan Q, Gupta VK, Sorooshian S (1993) A shuffled complex evolution approach for effective and efficient optimization. J Optim Theory Appl 76(3):501–521. https://doi.org/10.1007/BF00939380

Forsythe N, Fowler HJ, Blenkinsop S, Burton A, Kilsby CG, Archer DR, Harpham C, Hashmi MZ (2014) Application of a stochastic weather generator to assess climate change impacts in a semi-arid climate: the Upper Indus Basin. J Hydrol 517:1019–1034. https://doi.org/10.1016/j.jhydrol.2014.06.031

Furrer EM, Katz RW (2008) Improving the simulation of extreme precipitation events by stochastic weather generators. Water Resour Res 44:W12439. https://doi.org/10.1029/2008WR007316

Gangopadhyay S, Clark M, Rajagopalan B (2005) Statistical downscaling using K-nearest neighbors. Water Resour Res 41:W02024. https://doi.org/10.1029/2004WR003444

Gregory JM, Wigley TML, Jones PD (1992) Determining and interpreting the order of a two-state Markov chain: application to models of daily precipitation. Water Resour Res 28:1443–1446. https://doi.org/10.1029/92WR00477

Hosking JRM (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc 52:105–124

Hosking JRM, Wallis JR, Wood EF (1985) Estimation of the generalized extreme value distribution by the method of probability weighted moments. Technometrics 27:251–261. https://doi.org/10.1080/00401706.1985.10488049

Katz RW, Parlange MB, Naveau P (2002) Statistics of extremes in hydrology. Adv Water Resour 25:1287–1304. https://doi.org/10.1016/S0309-1708(02)00056-8

Kim D, Cho H, Onof C, Choi M (2017) Let-It-Rain: a web application for stochastic point rainfall generation at ungaged basins and its applicability in runoff and flood modeling. Stoch Env Res Risk Assess 31:1023–1043

Konapala G, Mishra AK, Wada Y, Mann M (2020) Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. Nat Commun 11:3044. https://doi.org/10.1038/s41467-020-16757-w

Lall U, Sharma A (1996) A nearest neighbor bootstrap for resampling hydrological time series. Water Resour Res 32:679–693. https://doi.org/10.1029/95WR02966

Madsen H, Rasmussen PF, Rosbjerg D (1997) Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. Water Resour Res 33:747–758. https://doi.org/10.1029/96WR03848

Mehrotra R, Sharma A (2009) Evaluating spatio-temporal representations in daily rainfall sequences from three stochastic multisite weather generation approaches. Adv Water Resour 32:948–962. https://doi.org/10.1016/j.advwatres.2009.03.005

Mehrotra R, Srikanthan R, Sharma A (2006) A comparison of three stochastic multisite precipitation occurrence generators. J Hydrol 331:280–292. https://doi.org/10.1016/j.jhydrol.2006.05.016

Mhanna M, Bauwens W (2012) A stochastic space–time model for the generation of daily rainfall in the Gaza Strip. Int J Climatol 32:1098–1112. https://doi.org/10.1002/joc.2305

Mishra AK, Coulibaly P (2009) Developments in hydrometric network design: a review. Rev Geophys 47:RG2001. https://doi.org/10.1029/2007rg000243

Muluye GY (2011) Deriving meteorological variables from numerical weather prediction model output: a nearest neighbor approach. Water Resour Res 47:W07509. https://doi.org/10.1029/2010WR009750

Neyman J, Scott EL (1958) Statistical approach to problems of cosmology. J R Stat Soc B 20:1–29

Nguyen TT (2019) Evaluation of multi-precipitation products for multi-time scales and spatial distribution during 2007–2015. Civ Eng J 5(1):255–267. https://doi.org/10.28991/cej-2019-03091242

Nicks AD, Gander GA (1994) CLIGEN: a weather generator for climate inputs to water resources and other models. In: Watson DG, Zazueta FS, Harrison TV (eds.), Proceedings of fifth international conference on computer in agriculture. ASAE, St. Joseph, MI, pp 903–909

Peck A, Prodanovic P, Simonovic SPP (2012) Rainfall intensity duration frequency curves under climate change: city of London, Ontario, Canada. Can Water Res J 37(3):177–189. https://doi.org/10.4296/cwrj2011-935

Pickering NB, Hansen JW, Jones JW, Wells CM, Chan VK, Godwin DC (1994) WeatherMan: a utility for managing and generating daily weather data. Agron J 86:332–337. https://doi.org/10.2134/agronj1994.00021962008600020023x

Qian B, Corte-Real J, Xu H (2002) Multisite stochastic weather models for impact studies. Int J Climatol 22:1377–1397. https://doi.org/10.1002/joc.808

Rajagopalan B, Lall U (1999) A k–nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour Res 35(10):3089–3101. https://doi.org/10.1029/1999WR900028

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17:182–190. https://doi.org/10.1029/WR017i001p00182

Rodriguez-Iturbe I, Cox DR, Isham V (1987) Some models for rainfall based on stochastic point processes. Proc R Soc Lond A 410:269–288. https://doi.org/10.1098/rspa.1987.0039

Rudemo M (1982) Empirical choice of histograms and kernel density estimators. Scand J Stat 9:65–78

Semenov MA, Barrow EM (1997) Use of a stochastic weather generator in the development of climate change scenarios. Clim Change 35:397–414. https://doi.org/10.1023/A:1005342632279

Sharif M, Burn DH (2007) Improved K-nearest neighbor weather generating model. J Hydrol Eng 12:42–51. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(42)

Sharma A, Lall U, Tarboton DG (1998) Kernel bandwidth selection for a first order nonparametric streamflow simulation model. Stoch Hydrol Hydraul 12(1):33–52. https://doi.org/10.1007/s004770050008

Shrestha A, Babel MK, Weesakul S, Vojinovic Z (2017) Developing intensity–duration–frequency (IDF) curves under climate change uncertainty: the case of Bangkok, Thailand. Water 9:145. https://doi.org/10.3390/w90200145

Smith RL (1985) Maximum likelihood estimation in a class of non regular cases. Biometrika 72:67–92. https://doi.org/10.2307/2336336

Steinschneider S, Brown C (2013) A semiparametric multivariate, multisite weather generator with low-frequency variability for use in climate risk assessments. Water Res Resour 49(11):7205–7220. https://doi.org/10.1002/wrcr.20528

Stockle CO, Campbell GS, Nelson R (1999) ClimGen Manual. Biological Systems Engineering Department, Washington State University, Pullman

Thorndahl S, Andersen AK, Larsen AB (2017) Event-based stochastic point rainfall resampling for statistical replication and climate projection of historical rainfall series. Hydrol Earth Syst Sci 21:4433–4448. https://doi.org/10.5194/hess-21-4433-2017

Vallam P, Qin X (2016) Multisite rainfall simulation at tropical regions: a comparison of three types of generators. Meteorol. Appl. 23:425–437. https://doi.org/10.1002/met.1567

Veneziano D, Furcolo P (2002) Multifractality of rainfall and scaling of intensity–duration–frequency curves. Water Res Resour 38(12):42. https://doi.org/10.1029/2001wr000372

Veneziano D, Yoon S (2013) Rainfall extremes, excesses, and intensity–duration–frequency curves: a unified asymptotic framework and new nonasymptotic results based on multifractal measures. Water Res Resour 49(7):4320–4334. https://doi.org/10.1002/wrcr.20352

Veneziano D, Lepore C, Langousis A, Furcolo P (2007) Marginal methods of intensity–duration–frequency estimation in scaling and nonscaling rainfall. Water Res Resour 43(10):1–14. https://doi.org/10.1029/2007WR006040

Vu MT, Mishra AK (2019) Nonstationary frequency analysis of the recent extreme precipitation events in the United States. J Hydrol 575:999–1010. https://doi.org/10.1016/j.jhydrol.2019.05.090

Vu MT, Mishra AK, Konapala G, Liu D (2018) Evaluation of multiple stochastic rainfall generators in diverse climatic regions. Stoch Env Res Risk Assess 32(5):1337–1353. https://doi.org/10.1007/s00477-017-1458-0

Wilks DS (1998) Multisite generalization of a daily stochastic precipitation model. J Hydrol 210:178–191. https://doi.org/10.1016/S0022-1694(98)00186-3

Wilson PS, Toumi R (2005) A fundamental probability distribution for heavy rainfall. Geophys Res Lett 32:L14812. https://doi.org/10.1029/2005GL022465

Wilson LL, Lettenmaier DP, Skyllingstad E (1992) A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. J Geophys Res D3:2791–2809. https://doi.org/10.1029/91JD02155

Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K (2003) A technique for generating regional climate scenarios using a nearest neighbor algorithm. Water Resour Res 39(7):1199. https://doi.org/10.1029/2002WR001769