Discrete k-nearest neighbor resampling for simulating multisite precipitation occurrence and model adaption to climate change

: Discrete KNNR for Multisite Occurrence (DKMO version1.0) - model development

Keywords: daily precipitation, discrete, k-nearest neighbor, Markov chain, multisite, occurrence

Taesam Lee[1] and Vijay P. Singh[2]

[1] Department of Civil Engineering, ERI, Gyeongsang National University,

501 Jinju-daero, Jinju, Gyeongnam, South Korea, 660-701

[2] Department of Biological and Agricultural Engineering & Zachry Department of Civil Engineering, Texas A&M University, 321 Scoates Hall, College Station, Texas, United States, 77843

20
Corresponding Author :

Taesam Lee, Ph.D.
Gyeongsang National University, Dept. of Civil Engineering
Tel)+82-55-772-1797, Fax)+82-55-772-1799
Email) tae3lee@gnu.ac.kr

Abstract

Stochastic weather simulation models are commonly employed in water resources management, agricultural applications, forest management, transportation management, and recreational activities. Stochastic simulation of multisite precipitation occurrence is a challenge because of its intermittent characteristics as well as spatial and temporal cross-correlation. This study proposes a novel simulation method for multisite precipitation occurrence employing a nonparametric technique, the discrete version of the k-nearest neighbor resampling (KNNR), and coupling it with Genetic Algorithm (GA). Its modification for the study of climatic change adaptation is also tested. The datasets simulated from both the DKNNR model and an existing traditional model were evaluated using a number of statistics, such as occurrence and transition probabilities as well as temporal and spatial cross-correlations. Results showed that the proposed DKNNR model with GA simulated multisite precipitation occurrence preserved the lagged crosscorrelation between sites while the existing conventional model was not able to reproduce lagged crosscorrelation between stations, so long stochastic simulation was required. Also, the GA mixing process provided a number of new patterns that were different from observations, which was not feasible with the sole DKNNR model. When climate change was considered, the model performed satisfactorily, but further improvement is required to more accurately simulate specific variations of the occurrence probability.

1. Introduction

Stochastic simulation of weather variables has been employed for water resources management, hydrological design, agricultural irrigation, forest management, transportation planning and evacuation, recreation activities, filling-in missing historical data, simulating data, extending observed records, and simulating different weather conditions. Stochastic simulation models play a key role in producing weather sequences, while preserving the statistical characteristics of observed data. A number of stochastic weather simulation models have been developed using parametric and nonparametric approaches (Lee, 2017; Lee et al., 2012; Wilby et al., 2003; Wilks, 1999; Wilks and Wilby, 1999).

Parametric approaches simulate statistical characteristics of observed weather data with a set of parameters that are determined by fitting (Jeong et al., 2012; Lee, 2016; Zheng and Katz, 2008), whereas in nonparametric approaches, historical analogs with current conditions are searched, following the weather simulation data (Buishand and Brandsma, 2001; Lee et al., 2012). Combinations of parametric and nonparametric approaches have also been proposed (Apipattanavis et al., 2007; Frost et al., 2011).

Among weather variables, precipitation possesses intermittency and zero values between precipitation events, which make it difficult to properly reproduce the events (Beersma and Buishand, 2003; Hughes et al., 1999; Katz and Zheng, 1999). To overcome the problem of intermittency and zero values, precipitation is simulated separately from other variables. The main method for reproducing intermittency has been the multiplication of precipitation occurrence and an amount as $Z=X\cdot Y$, where X is the occurrence (binary as either 0 or 1) and Y is the amount (Jeong et al., 2013; Lee and Park, 2017; Todorovic and Woolhiser, 1975). The spatial and temporal

68   dependence in the occurrence and amount of precipitation introduces further complexity in

69   multisite simulation.

70   Wilks (1998) presented a multisite simulation model for the occurrence process (i.e. X) using

71   the standard normal variable that is spatially dependent, representing the relation between the

72   occurrence variable and the standard normal variable with simulation data. Originally, the

73   occurrence of precipitation had been simulated with a discrete Markov Chain (MC) model (Katz,

74   1977). Compared to the MC model that requires a significant number of parameters for generating

75   multisite occurrence, the multisite occurrence model proposed by Wilks (1998) transforms the

76   standard normal variate and simulates the sequence with multivariate normal distribution, and then

77   back-transforms the multivariate normal sequence to the original domain. The model is able to

78   reproduce the contemporaneous multisite dependence structure and lagged dependence only for

79   the same site but it requires a complex simulation process to estimate parameters for each site and

80   is unable to preserve lagged dependence between sites. Also, a recent improvement has also been

81   made, but the weakness of the model in Wilks (1998) was not significantly improved (Evin et al.,

82   2018; Mehrotra et al., 2006; Srikanthan and Pegram, 2009).

83   Lee et al. (2010a) proposed a nonparametric-based stochastic simulation model for

84   hydrometeorological variables. Their model overcame the shortcomings of a previous

85   nonparametric simulation model (Lall and Sharma, 1996), called k-nearest neighbor resampling

86   (KNNR) but the simulated data do not produce patterns different from those of the observed data

87   (Brandsma and Buishand, 1998; Mehrotra et al., 2006; St-Hilaire et al., 2012). In addition to

88   KNNR, Lee et al. (2010a) used a meta-heuristic Genetic Algorithm (GA) that led to the

89   reproduction of similar populations by mixing the simulated datasets. Note that the reproduction

90   procedure of the GA allows to generate new patterns that are similar to observed patterns, but a

91   small number of totally new patterns are simulated from the mutation procedure of the GA.

92   While KNNR is employed to find historical analogues of multisite occurrence similar to the

93   current status of a simulation series, GA is applied to use its skill to generate a new descendant

94   from the historical parent chosen with the KNNR. In this procedure, the multisite occurrence of

95   precipitation can be simulated while preserving spatial and temporal correlations. Meta-heuristic

96   techniques, such as GA, have been popularly employed in a number of hydrometeorological

97   applications (Chau, 2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013).

98   Although a number of variants of KNNR-GA have been applied (Lee et al., 2012; Lee and Park,

99   2017), none of them can simulate multisite occurrence of precipitation whose characteristics are

100   binary and temporally and spatially related.

101   Therefore, this study proposes a stochastic simulation method for multisite occurrence of

102   precipitation with the KNNR-GA based nonparametric approach that (1) simulates multisite

103   occurrence with a simple and direct procedure without parameterization of all the required

104   occurrence probabilities; and (2) reproduces the complex temporal and spatial correlation between

105   stations as well as the basic occurrence probabilities. The proposed nonparametric model is

106   compared with the popular model proposed by Wilks (1998). Even though the multisite occurrence

107   data generated from the Wilks model preserves various statistical characteristics of the observed

108   data well, significant underestimation of lagged cross-correlation still exists. Furthermore, the

109   relation between standard normal variable and occurrence variable relies on long stochastic

110   simulation.

111   The paper is organized as follows. The next section presents the mathematical background

112   of existing multisite occurrence modeling and section discusses the modeling procedure. The

113   study area and data are reported in section 4. The model application is presented in section 5.

114   Results of the proposed model are discussed in section 6, and summary and conclusions are

115   presented in section 7.

2. Background

2.1. Single site occurrence modeling

118   Let X_t^s represent the occurrence of daily precipitation for a location s ($s=1,\ldots, S$) on day t

119   ($t=1,\ldots, n$; n is the number observed days) and let X_t^s be either zero for dry day or one for wet day.

120   The first order Markov chain model for X_t^s is defined with the assumption that the occurrence

121   probability of a wet day is fully defined by the previous day as

$$\Pr\{X_t^s = 1 \mid X_{t-1}^s = 0\} = p_{01}^s \tag{1}$$

$$\Pr\{X_t^s = 1 \mid X_{t-1}^s = 1\} = p_{11}^s \tag{2}$$

124   Also $p_{00}^s = 1 - p_{01}^s$ and $p_{10}^s = 1 - p_{11}^s$, since the summation of zero and one should be unity

125   with the same previous condition. This consists of a transition probability matrix (TPM) as

$$TPM^s = \begin{bmatrix} p_{00}^s & p_{01}^s \\ p_{10}^s & p_{11}^s \end{bmatrix} = \begin{bmatrix} 1 - p_{01}^s & p_{01}^s \\ 1 - p_{11}^s & p_{11}^s \end{bmatrix} \tag{3}$$

127   The marginal distributions of TPM (i.e. p_0 and p_1) can be expressed with TPM and its condition of

128    $p_0 + p_1 = 1$ as:

$$p_0^s = \frac{p_{01}^s}{1 + p_{01}^s - p_{11}^s} \tag{4}$$

130
$$p_1^s = \frac{1 - p_{11}^s}{1 + p_{01}^s - p_{11}^s} \tag{5}$$

131 Note that p_1 represents the probability of precipitation occurrence for a day, while p_0 does non-

132 occurrence. The lag-1 autocorrelation of precipitation occurrence is the combination of transition

133 probabilities as:

134
$$\rho_1(s,s) = p_{11}^s - p_{01}^s \tag{6}$$

135   The simulation can be done by comparing TPM with a uniform random number (u_t^s) as

136
$$X_t^s = \begin{cases} 1 & \text{if } u_t^s \leq p_{i1}^s \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

137 where p_{i1}^s is the selected probability from TPM regarding the previous condition i (i.e. either 0 or

138 1). Wilks (1998) suggested a different method using a standard normal random number $w_t^s \sim N[0,1]$

139 as

140
$$X_t^s = \begin{cases} 1 & \text{if } w_t^s \leq \Phi^{-1}(p_{i1}^s) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

141 where Φ^{-1} indicates the inverse of the standard normal cumulative function Φ.

2.2. Multisite occurrence modeling

143   Wilks (1998) suggested a multisite occurrence model using a standard normal random

144 number (here, denoted as MONR) that is spatially dependent but serially independent. The

145 correlation of the standard normal variate for a site pair of q and s can be expressed as:

146
$$\tau(q,s) = corr[w_t^q, w_t^s] \tag{9}$$

147   Also, the correlation of the original occurrence variate is

148     $$\rho(q,s) = corr[X_t^q, X_t^s] \tag{10}$$

149   Once the correlation of the standard normal variate is known, the simulation of multisite

150   precipitation occurrence is straightforward. Multivariate standard normal distribution is used with

151   a parameter set of $[\mathbf{0}, \mathbf{T}]$ where $\mathbf{0}$ is the zero vector (Sx1) and $\mathbf{T}$ is the correlation matrix with the

152   elements of $\tau(q,s)$ for $q \in \{1,...,S\}$ and $s \in \{1,...,S\}$.

153   Since direct estimation of $\tau(q,s)$ is not feasible, a simulation technique is used to estimate

154     $\tau(q,s)$ from $\rho(q,s)$. A long sequence of the occurrences is simulated with different values of

155     $\tau(q,s)$ and its corresponding correlation of the original domain $\rho(q,s)$ is estimated with the

156   simulated long sequence by the inverse standard normal cumulative function (i.e. Φ^{-1}). A curve

157   between $\tau(q,s)$ and $\rho(q,s)$ is derived from this long simulation with the MONR model and is

158   employed for parameter estimation for a real application.

159   ## 3. DKNNR

160   ### 3.1. DKNNR modeling procedure

161   In the current study, a novel multisite simulation model for discrete occurrence of precipitation

162   variable with k-nearest neighbor resampling (KNNR) technique (Lall and Sharma, 1996; Lee

163   and Ouarda, 2011; Lee et al., 2017) for a discrete case (denoted as Discrete KNNR; DKNNR)

164   is proposed by combining a mixture mechanism with Genetic Algorithm (GA). Provided the

165   number of nearest neighbors, k, is known, the discrete k-nearest neighbor resampling with

166   genetic algorithm is done as follows:

167   (1) Estimate the distance between the current (i.e. time index: c) multisite occurrence

168   X_c^s and the observed multisite occurrence x_i^s. Here, the distance is measured for

169         $i=1,\ldots, n$-1 as

170
$$D_i = \sum_{s=1}^{S} \left| X_c^s - x_i^s \right| \tag{11}$$

171   (2) Arrange the estimated distances from step (1) in ascending order, select the first k

172   distances (i.e., the smallest k values), and reserve the time indices of the smallest k

173   distances.

174   (3) Randomly select one of the stored k time indices with the weighting probability

175   given by

176
$$w_m = \frac{1/m}{\sum_{j=1}^{k} 1/j}, \qquad m = 1,\ldots, k \tag{12}$$

177   (4) Assume the selected time index from step (3) as p. Note that there are a number of

178   values that have the same distance as the selected D_p, since D_p is a natural number

179   between 0 and S. For example, if $S=2$ and $X_c^1=0$ and $X_c^2=1$, the two sequences have

180   the same $D=1$ as [$x_i^1=0$ and $x_i^2=0$] and [$x_i^1=1$ and $x_i^2=1$]. In this case, a random

181   selection procedure is required to take into account the cases with the same quantity.

182   One particular time index is randomly selected with equal probabilities among the

183   time indices of the same distances. Note that instead of the random selection, one

184   can always use the first one. In such a case, only one historical combination of

185   multisite occurrences will be selected.

186   (5) Assign the binary vector of the proceeding index of the selected time as

187         $\mathbf{x}_{p+1} = [x_{p+1}^s]_{s \in \{1, S\}}$. Here, p is the finally selected time index from step (4).

188   (6) Execute the following steps for GA mixing if GA mixing is subjectively selected.

189   Otherwise, skip this step.

190   (6-1) Reproduction: Select one additional time index using steps (1) through (4) and

191   denote this index as p^*. Obtain the corresponding precipitation occurrence

192   values, $\mathbf{x}_{p^*+1} = [x_{p^*+1}^s]_{s \in \{1,\dots,S\}}$. The subsequent two GA operators employ the two

193   selected vectors, $\mathbf{x}_{p+1}$ and $\mathbf{x}_{p^*+1}$. This reproduction process is a mating process

194   by finding another individual that has characteristics similar to those of the

195   current one $\mathbf{x}_{p+1}$. With this procedure, a vector similar to the current vector will

196   be mated and will produce a new descendant.

197   (6-2) Crossover: Replace each element x_{p+1}^s with $x_{p^*+1}^s$ at probability P_{cr}, i.e.,

198
$$X_{c+1}^s = \begin{cases} x_{p^*+1}^s & \text{if } \varepsilon < P_{cr} \\ x_{p+1}^s & \text{otherwise} \end{cases} \tag{13}$$

199   where ε is a uniform random number between 0 and 1. From this crossover, a

200   new occurrence vector whose elements are similar to the historical ones is generated.

201   (6-3) Mutation: Replace each element (i.e., each station, $s=1,\dots,S$) with one selected

202   from all the observations of this element for $i=1,\dots,n$ with probability P_m, i.e.,

203
$$X_{c+1}^s = \begin{cases} x_{\xi+1}^s & \text{if } \varepsilon < P_m \\ x_{p+1}^s & \text{otherwise} \end{cases} \tag{14}$$

204   where $x_{\xi+1}^s$ is selected from $[x_i^s]_{i\in\{1,...,n\}}$ with equal probability for $i=1,\ldots, n$

205   and ε is a uniform random number between 0 and 1. This mutation procedure

206   allows to generate a multisite occurrence combination that is totally different

207   from the historical records. Without this procedure, multisite occurrences

208   always similar to historical combinations are generated, which is not feasible

209   for a simulation purpose.

210   (7) Repeat steps (1)-(6) until the required data are generated.

211   The selection of the number of nearest neighbors (k) has been investigated by Lall and

212 Sharma (1996) and Lee and Ouarda (2011). A simple selection method was applied in the current

213 study as $k = \sqrt{n}$. For hydrometeorological stochastic simulation, this heuristic approach of the k

214 selection has been employed (Lall and Sharma, 1996; Lee and Ouarda, 2012; Lee et al., 2010b;

215 Prairie et al., 2006; Rajagopalan and Lall, 1999). One can use generalized cross-validation (GCV)

216 as shown in Sharma and Lall (1996) and Lee and Ouarda 2011 by treating this simulation as a

217 prediction problem. However, the current multisite occurrence simulation does not necessarily

218 require an accurate value prediction and not much difference in simulation using the simple

219 heuristic approach has been reported. Also, this heuristic approach of the k selection has been

220 popularly employed for hydrometeorological stochastic simulations (Lall and Sharma, 1996; Lee

221 and Ouarda, 2012; Lee et al., 2010b; Prairie et al., 2006; Rajagopalan and Lall, 1999).

222   In Appendix A, an example of the DKNNR simulation procedure is explained in detail.

223 **3.2. Adaptation to climate change**

224   The capability of model to take climate change into account is critical. For example, the

225 marginal distributions and transition probabilities in Eqs. (5) and (3) can change in future climate

226   scenarios. It is known that nonparametric simulation models have a difficulty to adapt to climate

227   change, since the models employ in general the current observation sequences. However, the

228   proposed model in the current study possesses the capability to adapt to the variations of

229   probabilities by tuning the crossover and mutation probabilities in P_{cr} (13) and P_m (14), adding the

230   condition when applied.

231   For example, the probability of P_{11} can be increased with the cross-over probability P_{cr} by

232   adding the condition to increase the probability of P_{11} as:

233
$$X_{c+1}^s = \begin{cases} x_{p*+1}^s & \text{if } \varepsilon < P_{cr} \ \& \ x_{p*+1}^s = 1 \ \& \ X_c^s = 1 \\ x_{p+1}^s & \text{otherwise} \end{cases} \tag{15}$$

234   It is obviously possible to increase the probability of P_1 by removing the condition of $X_c^s = 1$.

235   In addition, further adjustment can be made with the mutation process in Eq. (14) as

236
$$X_{c+1}^s = \begin{cases} x_{\xi+1}^s & \text{if } \varepsilon < P_m \text{ and } x_{\xi+1}^s = 1 \\ x_{p+1}^s & \text{otherwise} \end{cases} \tag{16}$$

237   This adjustment of adding the condition $x_{\xi+1}^s = 1$ can increase the marginal distribution as much as

238     $P_m \times P_1$. This has been tested in a case study.

4. Study area and data description

240   For testing the occurrence model, 12 weather stations were selected from Yeongnam province

241   which is located in the southeastern part of South Korea, as shown in Figure 1. Information on

242   longitude and latitude (fourth and fifth columns) as well as order index and the identification

243   number (first and second columns) of these stations operated by Korea Meteorological

244   Administration with the area name (third column) is shown in Table 1. The employed precipitation

245   dataset presents strong seasonality, since this area is dry from late fall to early autumn and humid

246   and rainy during the remaining seasons, especially in summer. The employed stations are not far

247   from each other, at most 100 km apart, and not much high mountains are located in the current

248   study area. Therefore, this region can be considered as a homogeneous region (Lee et al., 2007).

249   Figure 1 illustrates the locations of the selected weather stations. All the stations are inside

250   Yeongnam province which consists of two different regions as north and south Gyeongsang as

251   well as the self-governing cities of Busan, Daegu, and Ulsan. Most of the Yeongnam region is

252   drained to Nakdong River. To validate the proposed model appropriately, test sites must be highly

253   correlated with each other as well as have significant temporal relation. The stations inside the

254   Yeongnam area cover one of the most important watersheds, the Nakdong River basin, where the

255   Nakdong River passes through the entire basin and its hydrological assessments for agriculture

256   and climate change have a particular value in flood control and water resources management such

257   as floods and droughts.

258   It is important to analyze the impact of weather conditions for planning agricultural

259   operations and water resources management, especially during the summer season, because around

260   50-60 percent of the annual precipitation occurs during the summer season from June to

261   September. The length of daily precipitation data record ranges from 1976 to 2015 and the summer

262   season record was employed, since a large number of rainy days occur during summer and it is

263   important to preserve these characteristics. Also, the whole year dataset was tested and other

264   seasons were further applied but the correlation coefficient was relatively high and its correlation

265   matrix estimated was not a positive semi-definite matrix for the MONR model.

5. Application

To analyze the performance of the proposed DKNNR model, the occurrence of precipitation was simulated. The DKNNR simulation was compared with that of the MONR model. For each model, 100 series of daily occurrence with the same record length were simulated. The key statistics of observed data and each generated series, such as transition probabilities (P_{11}, P_{01}, and P_1) and cross-correlation (see Eq.(10)), were determined. The MONR model underestimated the lag-1 cross-correlation, as indicated by Wilks (1998). In the current study, this statistic was analyzed, since a synoptic scale weather system often results in lagged cross-correlation for daily precipitation data (Wilks, 1998). It was formulated as

$$\rho_1(q,s) = corr[X_{t-1}^q, X_t^s] \qquad (17)$$

Statistics from 100 generated series were evaluated by the root mean square error (RMSE) expressed as:

$$RMSE = \left(\frac{1}{N}\sum_{m=1}^{N}(\Gamma_m^G - \Gamma^h)^2\right)^{1/2} \qquad (18)$$

where N is the number of series (here 100), Γ_m^G is the statistic estimated from the mth generated series, while Γ^h is the statistic for the observed data. Note that lower RMSE indicates better performance, represented by the summarized error of a given statistic of generated series from the statistic of the observed data.

The 100 simulated statistic values were illustrated with boxplots to show their variability as shown in Figure 5 - Figure 7. The box of boxplot represents the interquartile range (IQR) ranging from 25 percentile to 75 percentile. The whiskers extend to up and down 1.5×IQR. Data beyond

286   the whiskers (1.5×IQR) are indicated by a plus sign (+). The horizontal line inside the box

287   represents the median of the data. The statistics of the observed data are denoted by a cross (x).

288   The closer a cross is to the horizontal line inside the box, the better the simulated data from a model

289   reproduces the statistical characteristics of the observed data.

6. Results

6.1.   GA mixing and its probability selection

292   The roles of crossover probability P_{cr} (Eq. (13)) and mutation probability P_m (Eq.(14)) were studied

293   by Lee et al. (2010b). In the current study, we further tested by selecting an appropriate parameter

294   set of these two parameters with the simulated data from the DKNNR model and the record length

295   of 100,000. RMSE (Eq. (18)) of the three transition and limiting probabilities (P_{11}, P_{01}, and P_1)

296   between the simulated data and the observed was used, since those probabilities are key statistics

297   that the simulated data must match the observed data and no parameterization of these probabilities

298   was made for the current DKNNR model. Results are shown in Figure 2 and Figure 3 for P_{cr} and

299   P_m, respectively. For P_{cr} in Figure 2, the probability of 0.02 shows the smallest RMSE in all

300   transition and limiting probabilities. The RMSE of P_m in Figure 3 shows a slight fluctuation along

301   with P_m. However, all three probabilities (P_{11}, P_{01}, and P_1) have relatively small RMSEs in P_m

302   =0.003. Therefore, the parameter set 0.02 and 0.003 was chosen for P_{cr} and P_m, respectively, and

303   employed in the current study. We also tested the simulation without the GA mixing procedure

304   (results not shown). The results showed that no better result could be found from the simulation

305   without GA mixing. The necessity of the GA mixing is further discussed in the following.

306   We further tested and discuss why the GA mixing is necessary in the proposed DKNNR

307   model as follows. For example, assume that three weather stations are considered and observed

308 data only has the occurrence cases of 000, 001,011,010, 011,100,111 among 2^3=8 possible cases.

309 In other words, no patterns for 110 and 101 is found in the observed data. Note that 0 is dry day

310 and 1 is rainy (or wet) day. The KNNR is a resampling process in that the simulation data is

311 resampled from observations. Therefore, no new patterns such as 110 and 101 can be found in the

312 simulated data.

313 This can be problematic for the simulation purpose in that one of the major simulation

314 purposes is to simulate sequences that might possibly happen in the future. The wet (1) or dry (0)

315 for multisite precipitation occurrence is decided by the spatial distribution of a precipitation

316 weather system. A humid air mass can be distributed randomly, relying on wind velocity and

317 direction as well as the surrounding air pressure. In general, any combinations of wet and dry

318 stations can be possible, especially when the simulation continues infinitely. Therefore, the

319 patterns of simulated data must be allowed to have any possible combinations, here 4096 even if

320 it has not been observed from the historical records. Also, its probability to have this new pattern

321 must not be high, since it has not been observed in the historical records and this can be taken into

322 account by low probability of the crossover and mutation.

323 This drawback of the KNNR model frequently happens in multisite occurrence as the

324 number of stations increases. Note that the number of patterns increases as 2^n where n is the number

325 of stations. If n=12, then 4096 cases must be observed. However, among 4096 cases, observed

326 cases are limited, since the number of data is limited. The GA process can mix two candidate

327 patterns to produce new patterns. For example, in the three station case, a new pattern 101 can be

328 produced from two observed occurrence candidates of 001 and 100 by the crossover of the first

329 value of 001 to the first value of 100 (i.e. 001 →101), which is not in the observed data.

330   Note that the data employed in the case study are 40 years and 122 days (summer months)

331   in each year. The total number of the observed data is 4880 and the number of possible cases is

332   4096. We checked the number of possible cases that were not found in the observed data. The

333   result shows that 3379 cases were not observed at all for the entire cases as shown in Figure 4.

334   We further investigated the number of new patterns that were generated with the

335   probabilities P_{cr}=0.02, P_m=0.001 by the proposed GA mixing. The generated data for 100

336   sequences from DKNNR with the GA mixing shows that the number 3379 was reduced to 1200,

337   which is not in the dataset among the 4096 possible patterns. Therefore, more than 2000 new

338   patterns were simulated with the GA mixing process. The KNNR model without the GA mixing

339   did not produce any new patterns in the 100 sequences with the same length of the historical data.

340

6.2.   Occurrence and transition probabilities

342   The data simulated from the proposed DKNNR model and the existing MONR model were

343   analyzed. The estimated transition probabilities (P_{11} and P_{01} in Eq. (3)) as well as the occurrence

344   probability (P_1 in Eq. (5)) are shown in Table 2 and Figure 5 - Figure 7 for the observed data and

345   the data generated from the DKNNR and MONR models. In Table 2, the observed statistic shows

346   that P_{11} is always higher than P_{01} and P_1 is between P_{11} and P_{01}. Site 6 shows the lowest P_{11} and

347   P_1 and site 12 shows the highest P_{11}.

348   As shown in Figure 5, the probability P_{11} of the observed data shows that sites 6, 7, 8, and 9

349   located in the northern part of the region exhibited lower consistency (i.e. consecutive rainy days)

350   than did the other sites, while sites 5 and 12 had higher probability of P_{11} than did other sites. Both

351   models preserved well the observed P_{11} statistic. It seems that the MONR model had a slightly

352 better performance, since this statistic is parameterized in the model as shown in section 2.2 and

353 that is the same for P01 and P1 as shown in Figure 6 and Figure 7. Note that the MONR model

354 employed the transition probabilities in simulating rainfall occurrence, while the DKNNR model

355 did not. The occurrence probability P_1 can be described with the combination of transition

356 probabilities as in Eq. (5). Even though the transition probabilities were not employed in

357 simulating rainfall occurrence, the DKNNR model preserved this statistic fairly well.

358   In the DKNNR modeling procedure, the simple distance measurement in Eq. (11) allows to

359 preserve transition probabilities in that the following multisite occurrence is resampled from the

360 historical data whose previous states of multisite occurrence (x_i^s) are similar to the current

361 simulation multisite occurrence (X_c^s). This summarized distance (D_i) is an essential tool in the

362 proposed DKNNR modeling. The condition of the current weather system is memorized and the

363 system is conditioned on simulating the following multisite occurrence with the distance

364 measurement like a precipitation weather system dynamically changes but often it impacts the

365 system of the following day.

366   As shown in Figure 6, the P_{01} probability showed a slightly different behavior such that sites

367 1, 2, and 3 located in the middle part of the Yeongnam province showed a higher probability than

368 did other sites. A slight underestimation was seen for sites 2 and 11 but it was not critical, since its

369 observed value with a cross mark was close to the upper IQR representing 75 percentile.

370   The behavior of P_1 was found to be the same as that of the P_{11} probability. It can be seen in

371 Figure 7 that no significant underestimation is seen for the DKNNR model (top panel). The P_1

372 statistic was fairly preserved by both DKNNR and MONR models. Note that the MONR model

373 parameterized the P_1 statistic through the transition probabilities as in Eq. (5), while the DKNNR

374 model did not. Although the DKNNR model used almost no parameters for simulation, the P_1

375 statistic was preserved fairly well.

6.3. Cross-correlation

377   Cross-correlation is a measure of the relationship between sites. The preservation of cross-

378 correlation is important for the simulation of precipitation occurrence and is required in the

379 regional analysis for water resources management or agricultural applications. Furthermore,

380 lagged cross-correlation is also essential as much as is cross-correlation (i.e. contemporaneous

381 correlation). For example, the amount of streamflow for a watershed from a certain precipitation

382 event is highly related with lagged cross-correlation.

383   Daily precipitation occurrence, in general, shows the strongest serial correlation at lag-1 and

384 its correlation decays as the lag gets longer. This is because a precipitation weather system moves

385 according to the surrounding pressure and wind direction that dynamically change within a day or

386 week. Therefore, we analyzed the lag-1 cross-correlation in the current study as the representative

387 lagged correlation structure.

388   The cross-correlation of observed data is shown in Table 3. High cross-correlation among

389 grouped sites, such as sites 6, 7, and 8 (northern part) and sites 3, 4, and 5, as well as 12 (southeast

390 coastal area, 0.68-0.87), was found. As expected, sites 5 and 12 had the highest cross-correlation

391 (0.87) due to proximity. The northern sites and coastal sites showed low cross-correlation. This

392 observed cross-correlation was well preserved in the data generated from both DKNNR and

393 MONR models, as shown in Figure 8 as well as Table 4 and Table 5. However, consistently slight

394 but significant underestimation of cross-correlation was seen for the data generated by the MONR

395 model (see the bottom panel of Figure 8). Note that the errorbars are extended to upper and lower

396   lines of the circles to 1.95×standard deviation. The difference of RMSE in Table 6 showed this

397   characteristic, as most of the values were positive, indicating that the proposed DKNNR model

398   performed better for cross-correlation.

399   The lag-1 cross-correlation of observed data, as shown in Table 7, ranged from 0.22-0.35.

400   The lag-1 cross-correlation for the same site (i.e. $\rho_1(q,s)$, $q=s$) was autocorrelation and was highly

401   related with P_{01} and P_{11} as in Eq. (6). All the lag-1 cross-correlations exhibited similar magnitudes

402   even for autocorrelation. This implies that the lag-1 cross-correlation among the selected sites was

403   as strong as the autocorrelation and as much as the transition probabilities P_{01} and P_{11}, thereof.

404   The observed lag-1 cross-correlations were well preserved in the data generated by the

405   DKNNR model, as shown in the top panel of Figure 9, while the MONR model showed significant

406   underestimation, as seen in the bottom panel of Figure 9. The difference of RMSE shown in Table

407   8 reflects this behavior. In the bottom panel of Figure 9, some of the lag-1 cross-correlations were

408   well preserved, that were aligned with the base line. From Table 8, the MONR model reproduced

409   the autocorrelations well with the shaded values. It is because the lag-1 autocorrelation was

410   indirectly parameterized with the transition probabilities of P_{11} and P_{01} as in Eq. (6). Other than

411   this autocorrelation, the lag-1 cross-correlation was not reproduced well with the MONR model.

412   This shortcoming was mentioned by Wilks (1998). Meanwhile, the proposed DKNNR model

413   preserved this statistic without any parameterization.

414   We further tested the performance measurements of MAE and Bias whose estimates showed

415   that MAE had no difference from RMSE. In addition, Bias of lag-1 correlation presented

416   significant negative values implying its underestimation for the simulated data of the MONR

417   model as shown in Table 9, while Table 10 of the DKNNR model showed a much smaller bias.

418   Also, the whole year data instead of the summer season data was tested for model fitting.

419   Note that all the results presented above were for the summer season data (June-September) as

420   mentioned in section 4 on data description. The lag-1 cross-correlation is shown in Figure 10 which

421   indicates that the same characteristic was observed as for the summer season, such that the

422   proposed DKNNR model preserved better the lagged cross-correlation than did the existing

423   MONR model. Other statistics, such as correlation matrix and transition probabilities, exhibited

424   the same results (not shown). Also, other seasons were tried but the estimated correlation matrix

425   was not a positive semi-definite matrix and its inverse cannot be made for multivariate normal

426   distribution in the MONR model. It was because the selected stations were close to each other

427   (around 50-100 km) and produced high cross-correlation, especially in the occurrence during dry

428   seasons. Special remedy for the existing MONR model should be applied, such as decreasing

429   cross-correlation by force, but further remedy was not applied in the current study since it was not

430   within the current scope and focus.

431   ### 6.4. Adaptation to climate change

432   Model adaptability to climate change in hydro-meteorological simulation models is a critical

433   factor, since one of the major applications of the models is to assess the impact of climate change.

434   Therefore, we tested the capability of the proposed model in the current study by adjusting the

435   probabilities of cross-over and mutation as in Eqs. (15) and (16). A number of variations can be

436   made with different conditions.

437   In Figure 11, the changes of transition and marginal probabilities are shown along with the

438   increase of crossover probability P_{cr} from 0.01 to 0.2 with the condition that that the candidate

439   value is one and the previous value is also one as in Eq. (15) for the selected 5 stations among the

440   12 stations (from station 1 to station 5, see Table 1 for details). The stations were limited in this

441   analysis due to computational time. In each case 100 series were simulated. The average value of

442   the simulated statistics is presented in the figure. It is obvious that the transition probability P_{11}

443   increased as intended along with the increase of P_{cr}. As expected from Eq. (5), P_1 presents that the

444   change of P_1 is highly related to P_{11}. However, the probability P_{01} fluctuated along with the

445   increase of P_{cr}. Elaborate work to adjust all the probabilities is however required.

446   The changes in transition and marginal probabilities are presented in Figure 12 with

447   increasing mutation probability P_m from 0.01 to 0.2 under the condition that the candidate value is

448   one so that the marginal probability P_1 increased. P_{01} also increased along with increasing P_1. The

449   change of P11 was not related with other probabilities. The combination of the adjustment of P_{cr}

450   and P_m with a certain condition to the previous state will allow the specific adaptation for

451   simulating future climatic scenarios.

452   As an example, assume that the occurrence probability (P1) of the control period is 0.26 (see

453   the dotted line with cross on the bottom panel of Figure 11 and Figure 12) and GCM output

454   indicates that the occurrence probability (P1) increases up to 0.27. This can be achieved with

455   increasing either the crossover probability to 0.1 or the mutation probability to 0.05. Note that the

456   crossover probabilities might affect the stations each other, while the mutation probabilities do not

457   Climate change, however, may refer to a larger phenomenon, which cannot be addressed

458   directly through modifying only the marginal and transition probabilities as in the current study.

459   Further model development on systematically varying temporal and spatial cross-correlations is

460   required to properly address climate change of the regional precipitation system.

7. Conclusions

461

462 In the current study, the discrete version of a nonparametric simulation model, based on

463 KNNR, is proposed to overcome the shortcomings of the existing MONR model, such as long

464 stochastic simulation for parameter estimation and underestimation of the lagged crosscorrelation

465 between sites as well as testing the adaptability for climatic change. Occurrence and transition

466 probabilities and cross-correlation as well as lag-1 cross-correlation are estimated for both models.

467 Better preservation of cross-correlation and lag-1 cross-correlation with the DKNNR model than

468 the MONR model is observed. For some cases (i.e., the whole year data and other seasons than the

469 summer season), the estimated cross-correlation matrix is not a positive semi-definite matrix so

470 the multivariate normal simulation is not applicable for the MONR model, because the tested sites

471 are close to each other with high cross-correlation.

472 Results of this study indicate that the proposed DKNNR model reproduces the occurrence

473 and transition probabilities satisfactorily and preserves the cross-correlations better than does the

474 existing MONR model. Furthermore, not much effort is required to estimate the parameters in the

475 DKNNR model, while the MONR model requires a long stochastic simulation just to estimate

476 each parameter. Thus, the proposed DKNNR model can be a good alternative for simulating

477 multisite precipitation occurrence.

478 We tested further the enhancement of the proposed model for adapting to climate change by

479 modifying the mutation and crossover probabilities P_m and P_{cr}. Results show that the proposed

480 DKNNR model has the capability to adapt to the climate change scenarios, but further elaborate

481 work is required to find the best probability estimation for climate change. Also, only the marginal

482 and transition probabilities cannot address the climate change of regional precipitation. The

483 variation of temporal and spatial cross-correlation structure must be considered to properly address

484 the climate change of the regional precipitation system. Further study on improving the model

485 adaptability to climate change will be followed in the near future. Also, the simulated multisite

486 occurrence can be coupled with a multisite amount model to produce precipitation events,

487 including zero values. Further development can be made for multisite amount models with a

488 nonparametric technique, such as KNNR and bootstrapping.

489 **Code and Data Availability**

490 DKNNR code is written in Matlab and is available as a supplement.

491 The precipitation data employed in the current study is downloadable through

492 http://www.weather.go.kr/weather/main.jsp

493 **Author Contribution**

494 T. Lee and V. Singh conceived of the presented idea. T. Lee developed the theory and

495 programming. V. Singh supervised the findings of the current work and the writing manuscript.

496 **Acknowledgment**

497 This work was supported by the National Research Foundation of Korea (NRF) grant (NRF-

498 2018R1A2B6001799) funded by the Korean Government (MEST).

499 # Appendix A: Example of DKNNR

500 In this appendix, one example of DKNNR simulation is presented with observed dataset in

501 Table A 1 (i.e. $\mathbf{x}_i = [x_i^s]_{s \in \{1,S\}}$ for $i=1,\ldots,n$; here $S=12$ and $n=16$). The upper part of the table

502 presents the observed precipitation (unit: mm). Its occurrence data is presented in the bottom part

503   of this table. The current precipitation occurrence $\mathbf{X}_c = [X_c^s]_{s\in\{1,...,12\}}$ is shown in the second row of

504   Table A 2. The number of nearest neighbors $k = \sqrt{n} = \sqrt{16} = 4$ and the parameters for GA (i.e. P_c

505   and P_m) are 0.1 and 0.01, respectively. Simulation can be made as follows:

506   (1) Estimate the distance D_i between $\mathbf{x}_i$ and $\mathbf{X}_c$ for $i=1,…,n$-1 as in Eq.(11). For example,

507   for $i=1$,

508    $$D_1 = \sum_{s=1}^{S} \left| X_c^s - x_1^s \right| = | 0-1 | + | 1-1 | +...+ | 0-1 |= 6$$

509   All the estimated distances are shown in the last column of Table A 2.

510   (2) The daily index values are sorted according to the smallest distances shown in the first

511   two columns of Table A 3. The sorted day indices and their corresponding distances are

512   shown in the third and fourth columns of Table A 3. From the k number of sorted indices,

513   one is selected with the weight probability (see Eq.(12)), which is shown in the last

514   column of Table A 3.

515   (3) Simulate a uniform random number (u) between 0 and 1. Say $u=0.321$. This value must

516   be compared with the cumulative weighted probabilities in the last column of Table A 3

517   as [0 0.48 0.72 0.88 1.0]. The corresponding day index is assigned as to where the

518   simulated uniform number falls in the cumulative weighted probabilities, here [0 0.48].

519   Therefore, the selected day, p, is 14. The occurrences of the following day $p+1=15$ for 12

520   stations are selected as in the second row of Table A 4.

521   (4) For GA mixture, another set must be chosen as in step (3). Say $u=0.561$, which falls in

522   [0.48 0.72]. The second one should be selected. However, there are a number of days with

523   the same distances. Specifically, six days have the same distances with D_i=4. In this case,

524   one among all six days is selected with equal probability. Assume that p=4 is selected and

525   the following occurrences are selected, as shown in the third row of Table A 4.

526   (5) With two sets, crossover and mutation process is performed as follows:

527   (5-1) Crossover: For each station, a uniform random number (ε) is generated and

528   compared with P_c=0.1 here. Say ε =0.345, then skip since ε =0.345> P_c=0.1. For

529   s=6, assume the generated random number, ε (=0.051)< P_c(=0.1) and then switch

530   the 6th station value of Set 1 into the value of Set 2 (see Table A 4). The occurrence

531   state of X_{c+1}^s turns into 1 from 0 as shown in the fourth row of Table A 4 as well as

532   station 8.

533   (5-2) Mutation: For each station, a uniform random number (ε) is generated and compared

534   with P_m=0.01. For s=12, assume ε =0.009< P_m=0.01 and switch the 12th station

535   value of Set 1 with the one selected among all the observed 12th station values with

536   equal probability (here the last column, s=12, of the bottom part of Table A 1, [1 1

537   0 0 … 1]). The occurrence state of X_{c+1}^{12} turns into 0 from 1 as shown in the fourth

538   column of Table A 4.

539   (6) Repeat steps (1)-(5) until the target simulation length is reached.

540

References

541

542

543   Apipattanavis, S., Podesta, G., Rajagopalan, B., and Katz, R. W.: A semiparametric

544   multivariate and multisite weather generator, Water Resources Research, 43, Artn W11401, 2007.

545   Beersma, J. J. and Buishand, A. T.: Multi-site simulation of daily precipitation and

546   temperature conditional on the atmospheric circulation, Climate Research, 25, 121-133, 2003.

547   Brandsma, T. and Buishand, T. A.: Simulation of extreme precipitation in the Rhine basin

548   by nearest-neighbour resampling, Hydrology and Earth System Sciences, 2, 195-209, 1998.

549   Buishand, T. A. and Brandsma, T.: Multisite simulation of daily precipitation and

550   temperature in the Rhine basin by nearest-neighbor resampling, Water Resources Research, 37,

551   2761-2776, 2001.

552   Chau, K. W.: Use of meta-heuristic techniques in rainfall-runoffmodelling, Water

553   (Switzerland), 9, 2017.

554   Evin, G., Favre, A. C., and Hingray, B.: Stochastic generation of multi-site daily

555   precipitation focusing on extreme events, Hydrology and Earth System Sciences, 22, 655-672,

556   2018.

557   Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K. W., Ardabili, S. F., and Piran,

558   M. J.: Survey of computational intelligence as basis to big flood management: Challenges, research

559   directions and future work, Engineering Applications of Computational Fluid Mechanics, 12, 411-

560   437, 2018.

561   Frost, A. J., Charles, S. P., Timbal, B., Chiew, F. H. S., Mehrotra, R., Nguyen, K. C.,

562   Chandler, R. E., McGregor, J. L., Fu, G., Kirono, D. G. C., Fernandez, E., and Kent, D. M.: A

563   comparison of multi-site daily rainfall downscaling techniques under Australian conditions,

564   Journal of Hydrology, 408, 1-18, 2011.

565   Hughes, J. P., Guttorp, P., and Charles, S. P.: A non-homogeneous hidden Markov model

566   for precipitation occurrence, Journal of the Royal Statistical Society. Series C: Applied Statistics,

567   48, 15-30, 1999.

568   Jeong, D. I., St-Hilaire, A., Ouarda, T. B. M. J., and Gachon, P.: A multi-site statistical

569   downscaling model for daily precipitation using global scale GCM precipitation outputs,

570   International Journal of Climatology, 33, 2431-2447, 2013.

571   Jeong, D. I., St-Hilaire, A., Ouarda, T. B. M. J., and Gachon, P.: Multisite statistical

572   downscaling model for daily precipitation combined by multivariate multiple linear regression and

573   stochastic weather generator, Climatic Change, 114, 567-591, 2012.

574   Katz, R. W.: Precipitation as a Chain-Dependent Process, Journal of Applied Meteorology,

575   16, 671-676, 1977.

576   Katz, R. W. and Zheng, X.: Mixture model for overdispersion of precipitation, Journal of

577   Climate, 12, 2528-2537, 1999.

578   Lall, U. and Sharma, A.: A nearest neighbor bootstrap for resampling hydrologic time

579   series, Water Resources Research, 32, 679-693, 1996.

580   Lee, T.: Multisite stochastic simulation of daily precipitation from copula modeling with

581   a gamma marginal distribution, Theoretical and Applied Climatology, doi: 10.1007/s00704-017-

582   2147-0, 2017. 1-10, 2017.

583   Lee, T.: Stochastic simulation of precipitation data for preserving key statistics in their

584   original domain and application to climate change analysis, Theoretical and Applied Climatology,

585   124, 91-102, 2016.

586   Lee, T. and Ouarda, T. B. M. J.: Identification of model order and number of neighbors

587   for k-nearest neighbor resampling, Journal of Hydrology, 404, 136-145, 2011.

588   Lee, T. and Ouarda, T. B. M. J.: Stochastic simulation of nonstationary oscillation hydro-

589   climatic processes using empirical mode decomposition, Water Resources Research, 48, 1-15,

590   2012.

591   Lee, T., Ouarda, T. B. M. J., and Jeong, C.: Nonparametric multivariate weather generator

592   and an extreme value theory for bandwidth selection, Journal of Hydrology, 452-453, 161-171,

593   2012.

594   Lee, T., Ouarda, T. B. M. J., and Yoon, S.: KNN-based local linear regression for the

595   analysis and simulation of low flow extremes under climatic influence, Climate Dynamics, doi:

596   10.1007/s00382-017-3525-0, 2017. 1-19, 2017.

597   Lee, T. and Park, T.: Nonparametric temporal downscaling with event-based population

598   generating algorithm for RCM daily precipitation to hourly: Model development and performance

599   evaluation, Journal of Hydrology, 547, 498-516, 2017.

600   Lee, T., Salas, J. D., and Prairie, J.: An enhanced nonparametric streamflow

601   disaggregation model with genetic algorithm, Water Resources Research, 46, 2010a.

602   Lee, T., Salas, J. D., and Prairie, J.: An Enhanced Nonparametric Streamflow

603   Disaggregation Model with Genetic Algorithm, Water Resources Research, 46, W08545, 2010b.

604   Lee, Y.-S., Heo, J.-H., Nam, W., and Kim, K.-D.: Application of Regional Rainfall

605   Frequency Analysis in South Korea(II): Monte Carlo Simulation and Determination of

606   Appropriate Method, Journal of the Korean Society of Civil Engineers, 27, 101-111, 2007.

607   Mehrotra, R., Srikanthan, R., and Sharma, A.: A comparison of three stochastic multi-site

608   precipitation occurrence generators, Journal of Hydrology, 331, 280-292, 2006.

609   Prairie, J. R., Rajagopalan, B., Fulp, T. J., and Zagona, E. A.: Modified K-NN model for

610   stochastic streamflow simulation, Journal of Hydrologic Engineering, 11, 371-378, 2006.

611   Rajagopalan, B. and Lall, U.: A k-nearest-neighbor simulator for daily precipitation and

612 other weather variables, Water Resources Research, 35, 3089-3101, 1999.

613   Srikanthan, R. and Pegram, G. G. S.: A nested multisite daily rainfall stochastic

614 generation model, Journal of Hydrology, 371, 142-153, 2009.

615   St-Hilaire, A., Ouarda, T. B. M. J., Bargaoui, Z., Daigle, A., and Bilodeau, L.: Daily river

616 water temperature forecast model with a k-nearest neighbour approach, Hydrological Processes,

617 26, 1302-1310, 2012.

618   Taormina, R., Chau, K. W., and Sivakumar, B.: Neural network river forecasting through

619 baseflow separation and binary-coded swarm optimization, Journal of Hydrology, 529, 1788-1797,

620 2015.

621   Todorovic, P. and Woolhiser, D. A.: Stochastic model of n-day precipitation Journal of

622 Applied Meteorology, 14, 17-24, 1975.

623   Wang, W. C., Xu, D. M., Chau, K. W., and Chen, S.: Improved annual rainfall-runoff

624 forecasting using PSO-SVM model based on EEMD, Journal of Hydroinformatics, 15, 1377-1390,

625 2013.

626   Wilby, R. L., Tomlinson, O. J., and Dawson, C. W.: Multi-site simulation of precipitation

627 by conditional resampling, Climate Research, 23, 183-194, 2003.

628   Wilks, D. S.: Multisite downscaling of daily precipitation with a stochastic weather

629 generator, Climate Research, 11, 125-136, 1999.

630   Wilks, D. S.: Multisite generalization of a daily stochastic precipitation generation model,

631 Journal of Hydrology, 210, 178-191, 1998.

632   Wilks, D. S. and Wilby, R. L.: The weather generation game: a review of stochastic

633 weather models, Progress in Physical Geography, 23, 329-357, 1999.

634   Zheng, X. and Katz, R. W.: Simulation of spatial dependence in daily rainfall using

635   multisite generators, Water Resources Research, 44, 2008.

636
637

638

639   Table 1. Information on 12 selected stations from Yeongnam province, South Korea.

Order	Station Number[†]	Name	Longitude	Latitude
1	138	Pohang	129.3797	36.0327
2	143	Daegu	128.6189	35.8850
3	152	Ulsan	129.3200	35.5600
4	159	Busan	129.0319	35.1044
5	162	Tongyeong	128.4356	34.8453
6	277	Youngdeok	129.4092	36.5331
7	278	Uisung	128.6883	36.3558
8	279	Gumi	128.3206	36.1306
9	281	Youngcheon	128.9514	35.9772
10	285	Hapcheon	128.1697	35.5650
11	288	Milyang	128.7439	35.4914
12	294	Geojae	128.6044	34.8881

640   [†]The station number indicates the identification number operated by Korea Meteorological
641   Administration (KMA).

642

643

644 Table 2. Occurrence and transition probabilities of observed data and data simulated by DKNNR
645 and MONR for 12 stations from Yeongnam province, South Korea, during the summer season.
646 Note that 100 sets with the same record length as the observed data were simulated and the
647 statistics of 100 sets were averaged.

	Obs			DKNNR			MONR		
	P11	P01	P1	P11	P01	P1	P11	P01	P1
S1	0.56	0.27	0.38	0.56	0.27	0.38	0.56	0.26	0.37
S2	0.56	0.27	0.38	0.58	0.26	0.38	0.57	0.25	0.37
S3	0.57	0.26	0.38	0.58	0.26	0.38	0.56	0.26	0.37
S4	0.58	0.25	0.37	0.58	0.25	0.37	0.57	0.24	0.36
S5	0.58	0.25	0.37	0.59	0.24	0.37	0.58	0.24	0.36
S6	0.52	0.25	0.34	0.50	0.24	0.33	0.52	0.24	0.33
S7	0.55	0.26	0.36	0.56	0.25	0.36	0.55	0.24	0.35
S8	0.56	0.25	0.37	0.57	0.25	0.37	0.57	0.24	0.36
S9	0.55	0.25	0.36	0.55	0.24	0.35	0.55	0.24	0.35
S10	0.58	0.25	0.38	0.59	0.24	0.37	0.57	0.23	0.35
S11	0.57	0.25	0.36	0.58	0.24	0.36	0.56	0.24	0.35
S12	0.59	0.25	0.38	0.59	0.25	0.38	0.59	0.25	0.37

648
649
650

651 Table 3. Cross-correlation of observed data for 12 stations from Yeongnam province, South
652 Korea.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	1.00	0.70	0.70	0.64	0.58	0.70	0.65	0.63	0.75	0.64	0.66	0.59
S2	0.70	1.00	0.67	0.64	0.61	0.64	0.70	0.72	0.79	0.72	0.74	0.62
S3	0.70	0.67	1.00	0.75	0.68	0.61	0.57	0.57	0.68	0.67	0.74	0.70
S4	0.64	0.64	0.75	1.00	0.79	0.56	0.56	0.55	0.65	0.66	0.73	0.82
S5	0.58	0.61	0.68	0.79	1.00	0.51	0.54	0.55	0.61	0.65	0.70	0.87
S6	0.70	0.64	0.61	0.56	0.51	1.00	0.69	0.65	0.68	0.59	0.59	0.54
S7	0.65	0.70	0.57	0.56	0.54	0.69	1.00	0.78	0.71	0.65	0.63	0.55
S8	0.63	0.72	0.57	0.55	0.55	0.65	0.78	1.00	0.71	0.68	0.65	0.56
S9	0.75	0.79	0.68	0.65	0.61	0.68	0.71	0.71	1.00	0.68	0.71	0.62
S10	0.64	0.72	0.67	0.66	0.65	0.59	0.65	0.68	0.68	1.00	0.77	0.66
S11	0.66	0.74	0.74	0.73	0.70	0.59	0.63	0.65	0.71	0.77	1.00	0.70
S12	0.59	0.62	0.70	0.82	0.87	0.54	0.55	0.56	0.62	0.66	0.70	1.00

653
654
655

656 Table 4. Averaged cross-correlation of the 100 simulated series from the DKNNR model for 12
657 stations from Yeongnam province, South Korea.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	1.00	0.68	0.69	0.64	0.60	0.69	0.64	0.62	0.73	0.63	0.65	0.61
S2	0.68	1.00	0.67	0.63	0.62	0.63	0.68	0.72	0.77	0.74	0.73	0.63
S3	0.69	0.67	1.00	0.74	0.69	0.60	0.58	0.59	0.66	0.68	0.74	0.70
S4	0.64	0.63	0.74	1.00	0.79	0.55	0.55	0.56	0.62	0.65	0.71	0.81
S5	0.60	0.62	0.69	0.79	1.00	0.51	0.56	0.58	0.60	0.66	0.70	0.86
S6	0.69	0.63	0.60	0.55	0.51	1.00	0.68	0.64	0.65	0.59	0.58	0.53
S7	0.64	0.68	0.58	0.55	0.56	0.68	1.00	0.78	0.69	0.65	0.63	0.56
S8	0.62	0.72	0.59	0.56	0.58	0.64	0.78	1.00	0.70	0.69	0.67	0.58
S9	0.73	0.77	0.66	0.62	0.60	0.65	0.69	0.70	1.00	0.67	0.69	0.60
S10	0.63	0.74	0.68	0.65	0.66	0.59	0.65	0.69	0.67	1.00	0.77	0.67
S11	0.65	0.73	0.74	0.71	0.70	0.58	0.63	0.67	0.69	0.77	1.00	0.71
S12	0.61	0.63	0.70	0.81	0.86	0.53	0.56	0.58	0.60	0.67	0.71	1.00

658
659
660

661 Table 5. Averaged cross-correlation of 100 simulated series from the MONR model for 12
662 stations from Yeongnam province.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	1.00	0.63	0.67	0.58	0.54	0.66	0.62	0.60	0.68	0.55	0.62	0.53
S2	0.63	1.00	0.61	0.60	0.57	0.59	0.68	0.68	0.75	0.66	0.72	0.58
S3	0.67	0.61	1.00	0.71	0.67	0.57	0.56	0.53	0.65	0.61	0.71	0.69
S4	0.58	0.60	0.71	1.00	0.78	0.50	0.52	0.52	0.61	0.62	0.69	0.78
S5	0.54	0.57	0.67	0.78	1.00	0.48	0.51	0.53	0.57	0.62	0.67	0.81
S6	0.66	0.59	0.57	0.50	0.48	1.00	0.67	0.62	0.63	0.54	0.54	0.49
S7	0.62	0.68	0.56	0.52	0.51	0.67	1.00	0.75	0.70	0.61	0.62	0.52
S8	0.60	0.68	0.53	0.52	0.53	0.62	0.75	1.00	0.66	0.64	0.61	0.52
S9	0.68	0.75	0.65	0.61	0.57	0.63	0.70	0.66	1.00	0.63	0.69	0.57
S10	0.55	0.66	0.61	0.62	0.62	0.54	0.61	0.64	0.63	1.00	0.72	0.61
S11	0.62	0.72	0.71	0.69	0.67	0.54	0.62	0.61	0.69	0.72	1.00	0.66
S12	0.53	0.58	0.69	0.78	0.81	0.49	0.52	0.52	0.57	0.61	0.66	1.00

663
664
665
666

667 Table 6. The difference of RMSE of cross-correlation between MONR and DKNNR. Note that
668 the positive value indicates that the DKNNR model better performs in preserving the cross-
669 correlation, while a negative value (underlined) shows that the MONR model better performs.

MONR-DKNNR	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	0.000	0.014	0.004	0.013	0.012	0.012	0.008	0.005	0.024	0.031	0.011	0.035
S2	0.014	0.000	0.023	0.013	0.021	0.009	0.010	0.013	0.018	0.027	0.011	0.020
S3	0.004	0.023	0.000	0.015	0.004	0.014	0.003	0.022	0.009	0.028	0.011	0.004
S4	0.013	0.013	0.015	0.000	0.002	0.017	0.018	0.014	0.018	0.018	0.027	0.024
S5	0.012	0.021	0.004	0.002	0.000	0.014	0.021	0.014	0.015	0.013	0.015	0.012
S6	0.012	0.009	0.014	0.017	0.014	0.000	0.006	0.010	0.030	0.018	0.029	0.021
S7	0.008	0.010	0.003	0.018	0.021	0.006	0.000	0.005	0.008	0.024	0.012	0.023
S8	0.005	0.013	0.022	0.014	0.014	0.010	0.005	0.000	0.032	0.019	0.022	0.023
S9	0.024	0.018	0.009	0.018	0.015	0.030	0.008	0.032	0.000	0.019	0.005	0.027
S10	0.031	0.027	0.028	0.018	0.013	0.018	0.024	0.019	0.019	0.000	0.020	0.021
S11	0.011	0.011	0.011	0.027	0.015	0.029	0.012	0.022	0.005	0.020	0.000	0.022
S12	0.035	0.020	0.004	0.024	0.012	0.021	0.023	0.023	0.027	0.021	0.022	0.000

670 Note that no negative value can be found implying that the DKNNR model preserves the
671 crosscorrelation better than the MONR model.

672

673

674

675
676

677 Table 7. Lag-1 cross-correlation of observed data for 12 stations from Yeongnam province,
678 South Korea.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	0.29[‡]	0.26	0.30	0.27	0.24	0.29	0.26	0.24	0.27	0.26	0.28	0.26
S2	0.28	0.30	0.29	0.28	0.26	0.28	0.28	0.27	0.31	0.30	0.32	0.27
S3	0.28	0.26	0.31	0.30	0.27	0.27	0.25	0.24	0.27	0.27	0.30	0.27
S4	0.28	0.27	0.32	0.34	0.31	0.27	0.26	0.26	0.28	0.28	0.31	0.32
S5	0.29	0.28	0.32	0.35	0.34	0.27	0.27	0.26	0.29	0.29	0.33	0.35
S6	0.25	0.22	0.26	0.23	0.22	0.27	0.24	0.22	0.25	0.23	0.24	0.23
S7	0.25	0.26	0.27	0.25	0.25	0.28	0.29	0.27	0.27	0.27	0.28	0.26
S8	0.29	0.30	0.29	0.27	0.26	0.30	0.31	0.30	0.31	0.30	0.31	0.27
S9	0.29	0.29	0.30	0.29	0.27	0.29	0.27	0.27	0.30	0.30	0.31	0.28
S10	0.28	0.31	0.32	0.31	0.29	0.29	0.30	0.30	0.31	0.33	0.34	0.29
S11	0.27	0.29	0.31	0.30	0.27	0.27	0.27	0.27	0.29	0.30	0.32	0.29
S12	0.30	0.29	0.32	0.35	0.33	0.28	0.27	0.26	0.29	0.30	0.33	0.35

679 [‡]Shaded values represent lag-1 autocorrelation (i.e. the one lagged correlation for the same site).

680

681

682   Table 8. The difference of RMSE of lag-1 cross-correlation between MONR and DKNNR. Note
683   that a positive value indicates that the DKNNR model better performs in preserving lag-1 cross-
684   correlation, while a negative value (underlined) shows that the MONR model better performs.

MONR-DKNNR	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	0.000	0.048	0.075	0.049	0.041	0.095	0.059	0.036	0.047	0.055	0.063	0.052
S2	0.070	0.000	0.079	0.057	0.046	0.104	0.068	0.047	0.066	0.058	0.073	0.047
S3	0.067	0.054	0.000	0.046	0.031	0.096	0.072	0.056	0.055	0.052	0.056	0.025
S4	0.086	0.075	0.083	0.002	0.037	0.117	0.089	0.077	0.078	0.062	0.077	0.040
S5	0.111	0.096	0.098	0.074	0.002	0.124	0.103	0.085	0.105	0.070	0.108	0.049
S6	0.039	0.024	0.060	0.038	0.043	-0.002	0.028	0.017	0.045	0.034	0.055	0.037
S7	0.055	0.045	0.077	0.061	0.062	0.084	0.000	0.023	0.051	0.052	0.071	0.064
S8	0.092	0.078	0.104	0.079	0.068	0.115	0.079	0.000	0.094	0.078	0.101	0.074
S9	0.060	0.052	0.084	0.066	0.056	0.106	0.057	0.056	0.001	0.069	0.076	0.064
S10	0.091	0.094	0.105	0.081	0.062	0.123	0.107	0.085	0.100	0.001	0.092	0.063
S11	0.064	0.061	0.071	0.057	0.033	0.109	0.084	0.063	0.062	0.043	-0.002	0.043
S12	0.121	0.099	0.096	0.077	0.036	0.130	0.101	0.086	0.107	0.082	0.109	0.003

685

686

687

688 Table 9. Bias of lag-1 cross-correlation of the generated data from the DKNNR model. Note that
689 a positive value indicates the overestimation of lag-1 cross-correlation, while a negative value

690 shows underestimation. Note that $Bias = 1/N \sum_{m=1}^{N} \Gamma_m^G - \Gamma^h$ and see Eq. (18) for the details of each

691 term.

692

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	0.000	0.009	0.001	0.003	0.006	-0.002	0.010	0.011	0.006	0.010	0.010	0.006
S2	0.005	0.009	0.010	0.006	0.008	0.006	0.011	0.011	0.004	0.009	0.009	0.010
S3	0.002	0.010	0.001	-0.002	0.003	0.002	0.007	0.008	0.006	0.009	0.006	0.007
S4	0.006	0.009	0.004	0.001	0.007	0.003	0.008	0.008	0.009	0.010	0.010	0.005
S5	0.004	0.005	0.000	-0.001	-0.001	0.007	0.005	0.006	0.002	0.008	0.000	-0.001
S6	-0.002	0.006	0.000	0.002	-0.001	-0.002	0.004	0.003	0.002	0.005	0.004	0.001
S7	0.004	0.008	0.003	0.003	0.001	0.004	0.002	0.006	0.007	0.007	0.007	0.002
S8	0.000	0.005	0.004	0.001	0.004	-0.003	-0.003	0.000	0.001	0.004	0.006	0.003
S9	0.005	0.007	0.006	0.003	0.006	0.004	0.010	0.007	0.004	0.007	0.006	0.007
S10	0.003	0.005	0.001	-0.001	-0.001	0.001	0.001	0.001	0.003	0.000	0.002	0.001
S11	0.010	0.010	0.008	0.004	0.008	0.009	0.009	0.009	0.010	0.010	0.011	0.008
S12	0.003	0.006	0.001	-0.001	0.004	0.003	0.008	0.008	0.005	0.005	0.002	0.001

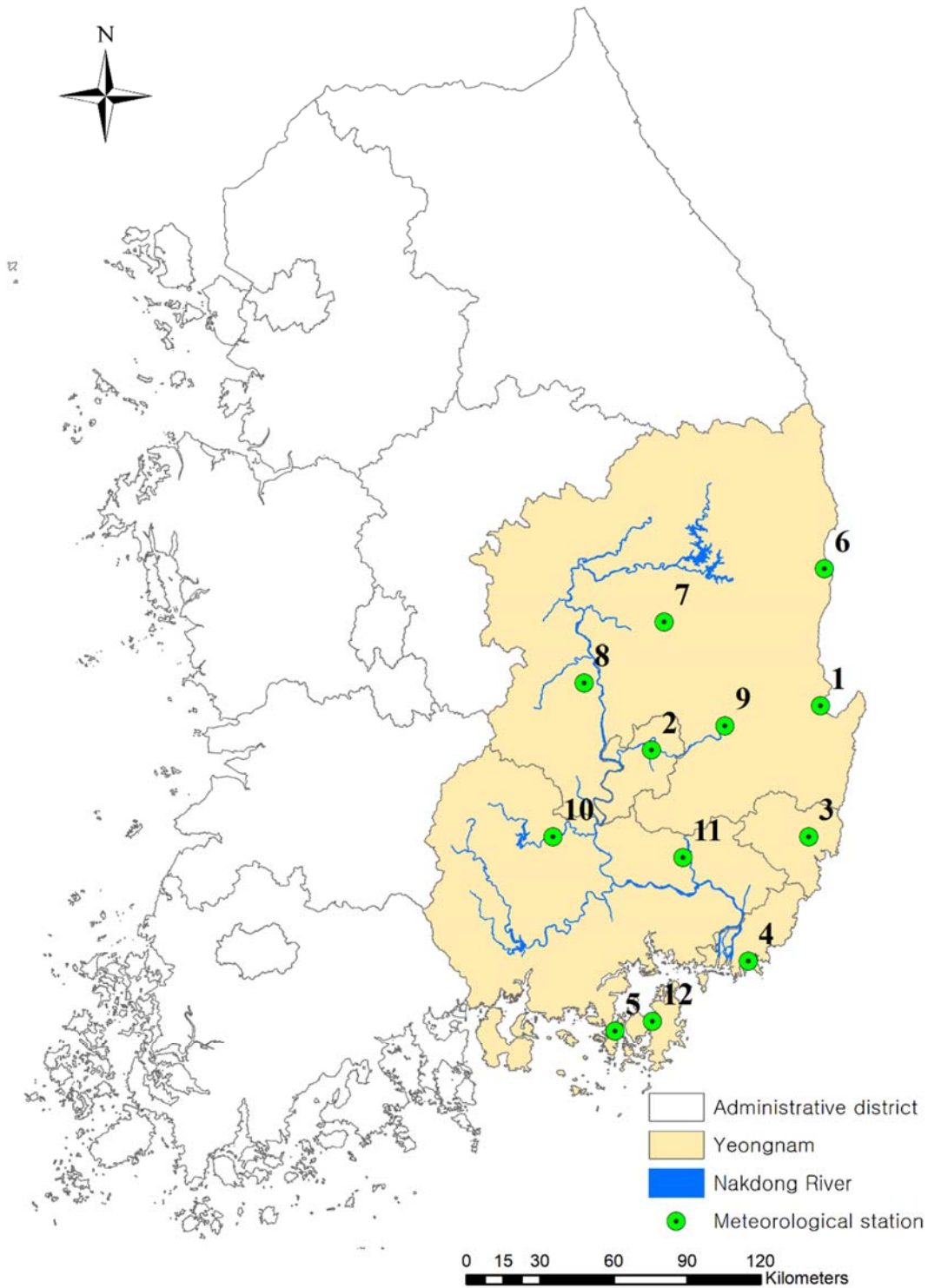
693

694

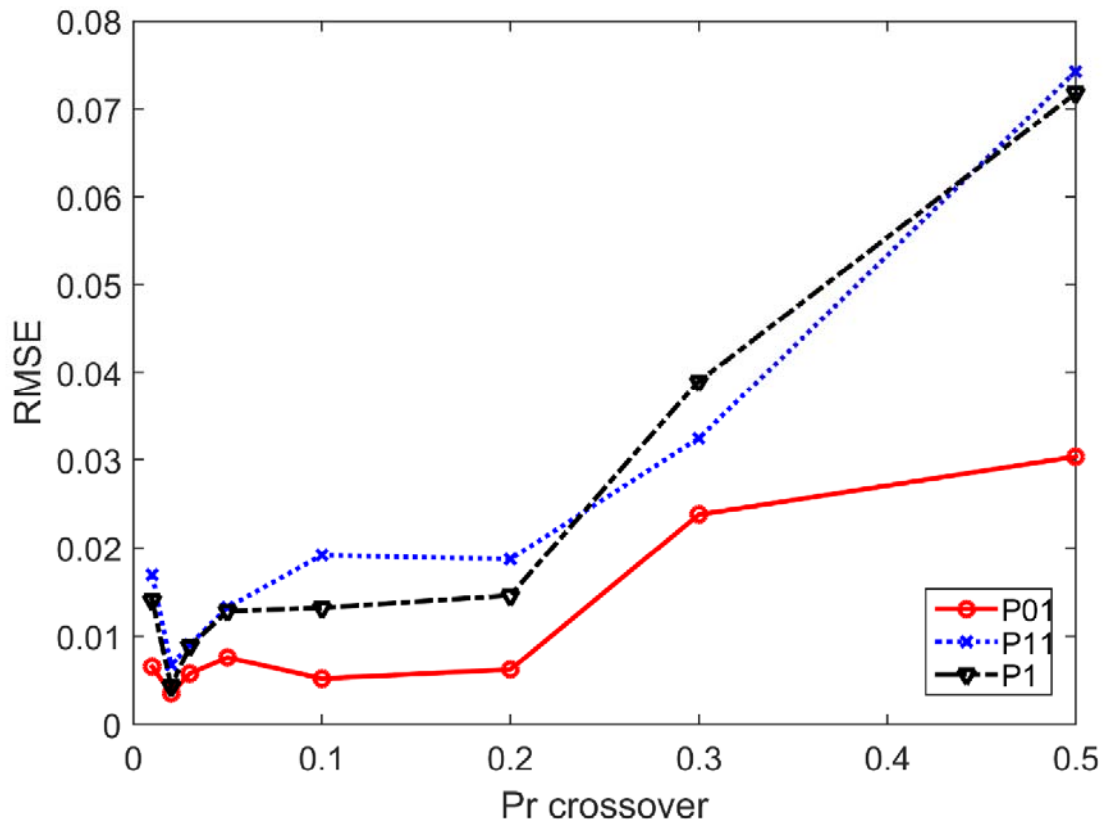
695 Table 10. Bias of lag-1 cross-correlation of the generated data from the Wilks model. Note that a
696 positive value indicates the overestimation of lag-1 cross-correlation, while a negative value
697 shows underestimation.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	-0.001	-0.062	-0.089	-0.063	-0.055	-0.106	-0.074	-0.052	-0.060	-0.070	-0.080	-0.067
S2	-0.084	0.000	-0.096	-0.072	-0.061	-0.117	-0.083	-0.063	-0.079	-0.072	-0.089	-0.063
S3	-0.080	-0.070	0.001	-0.059	-0.043	-0.110	-0.086	-0.072	-0.069	-0.066	-0.071	-0.037
S4	-0.100	-0.090	-0.097	-0.001	-0.048	-0.129	-0.103	-0.093	-0.093	-0.077	-0.092	-0.051
S5	-0.125	-0.110	-0.111	-0.087	-0.001	-0.138	-0.117	-0.100	-0.118	-0.084	-0.121	-0.060
S6	-0.053	-0.037	-0.074	-0.051	-0.057	-0.001	-0.039	-0.030	-0.060	-0.047	-0.070	-0.049
S7	-0.068	-0.058	-0.091	-0.077	-0.077	-0.098	-0.002	-0.038	-0.065	-0.065	-0.086	-0.079
S8	-0.106	-0.091	-0.119	-0.094	-0.084	-0.128	-0.093	0.001	-0.108	-0.091	-0.116	-0.088
S9	-0.074	-0.064	-0.098	-0.080	-0.070	-0.119	-0.072	-0.070	-0.001	-0.082	-0.091	-0.078
S10	-0.105	-0.107	-0.120	-0.096	-0.075	-0.136	-0.119	-0.097	-0.113	-0.001	-0.106	-0.076
S11	-0.078	-0.074	-0.085	-0.070	-0.047	-0.123	-0.097	-0.077	-0.076	-0.056	-0.001	-0.057
S12	-0.134	-0.112	-0.108	-0.088	-0.046	-0.142	-0.116	-0.101	-0.121	-0.095	-0.122	0.000

698

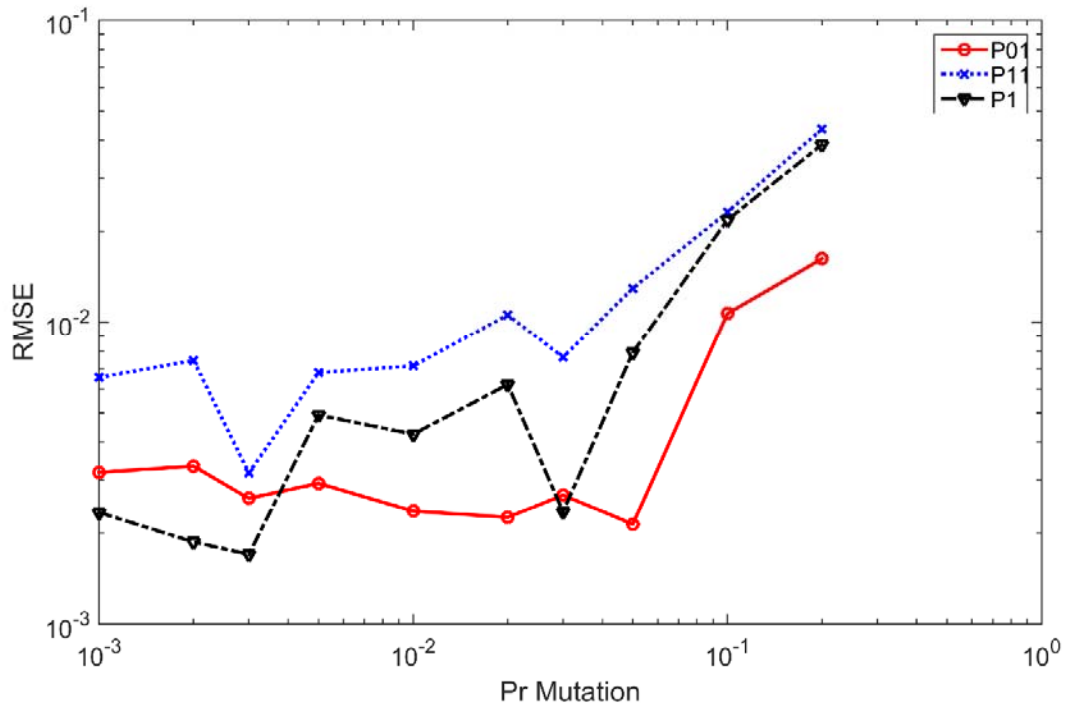
699
700

701

702   Figure 1. Locations of 12 selected weather stations at the Yeongnam province. See Table 1 for
703   further information about the stations.

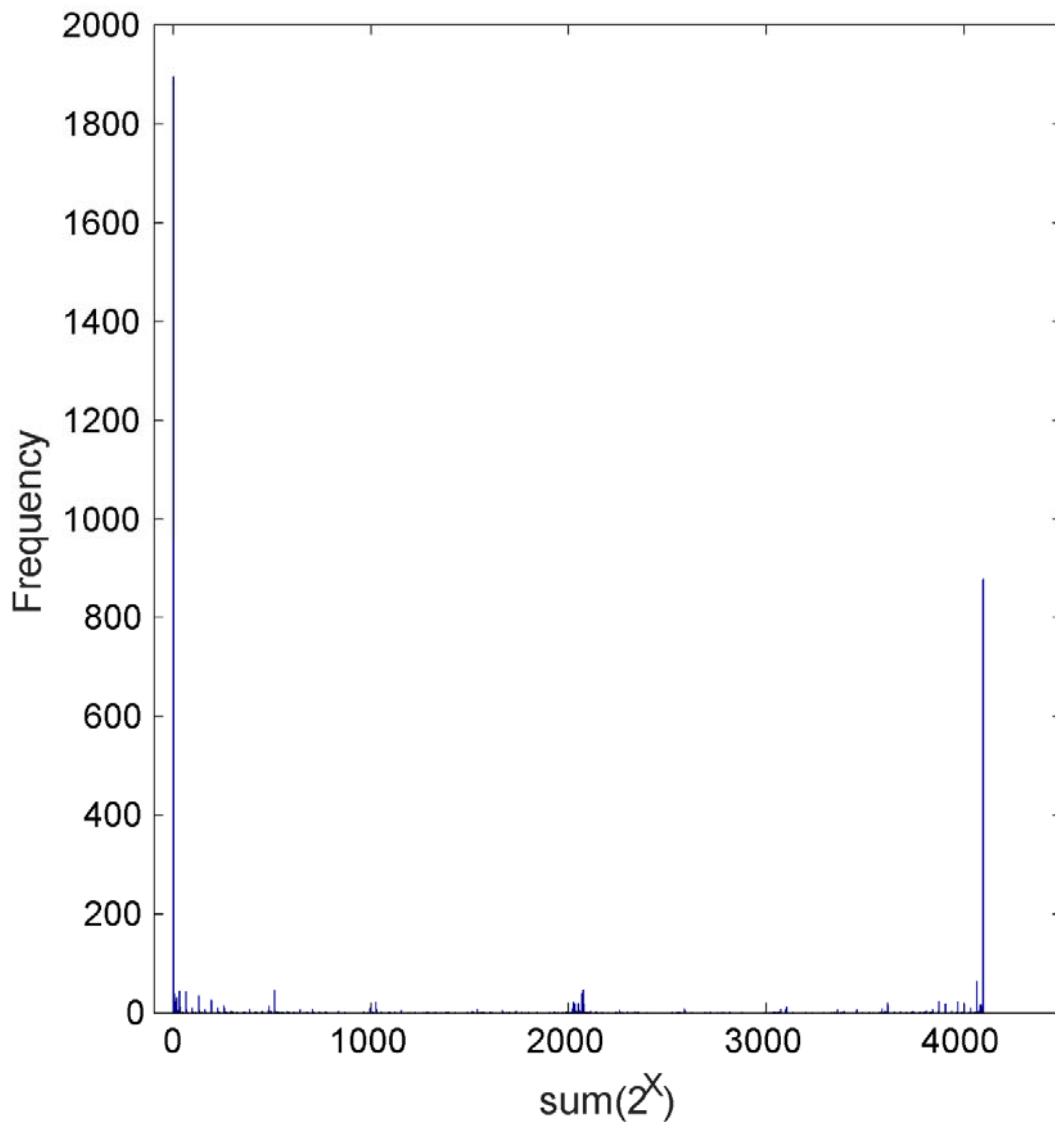
704

705
706 Figure 2. Testing for different probabilities of crossover Pcr. RMSE is estimated for all the tested
707 12 stations for each transition and limiting probability of the simulated data with the record
708 length of 100,000.

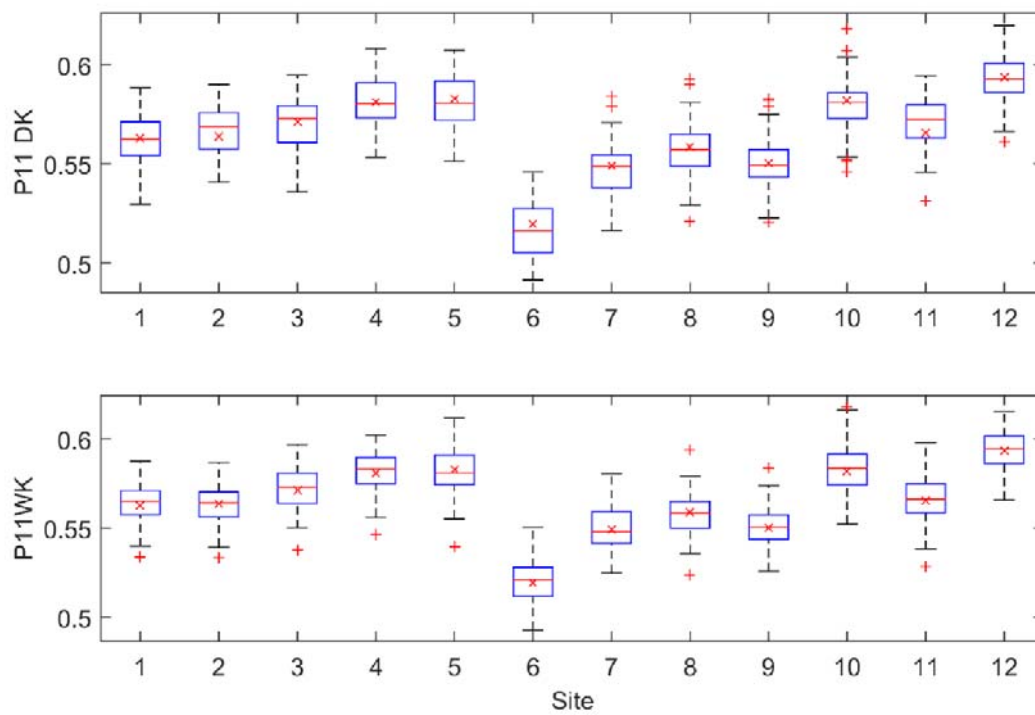
709

710

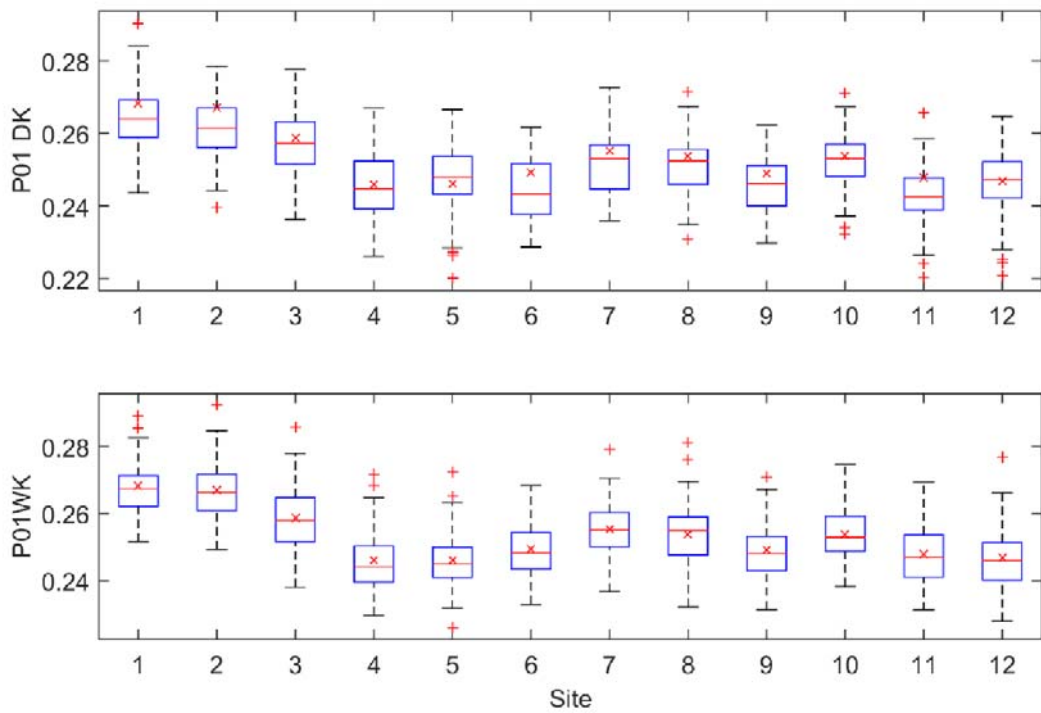
711
712

713

714 Figure 3. Testing for different probabilities of mutation P_m. RMSE is estimated for all the tested
715 12 stations for each transition and limiting probability of the simulated data with the record
716 length of 100,000.

717
718

719
720 Figure 4. Frequency of the observed patterns among all the possible cases (2^{12}=4096). The X
721 coordinate indicates each pattern with the numbering of the binary number system. All zero (0)
722 and all one (4095) has the largest and second largest numbers of frequency as 1894 and 877,
723 respectively as expected meaning all dry and all wet stations. Note that the bars are very sporadic
724 indicating a number of occurrence patterns are not observed.

725
726

727

728   Figure 5. Boxplots of the P11 probability for the simulated data from the DKNNR model (top
729   panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12
730   selected weather stations from the Yeongnam province.

731

732

733   Figure 6. Boxplots of the P01 probability for the data simulated from the DKNNR model (top
734   panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12
735   selected weather stations from the Yeongnam province.

736

737

738

739

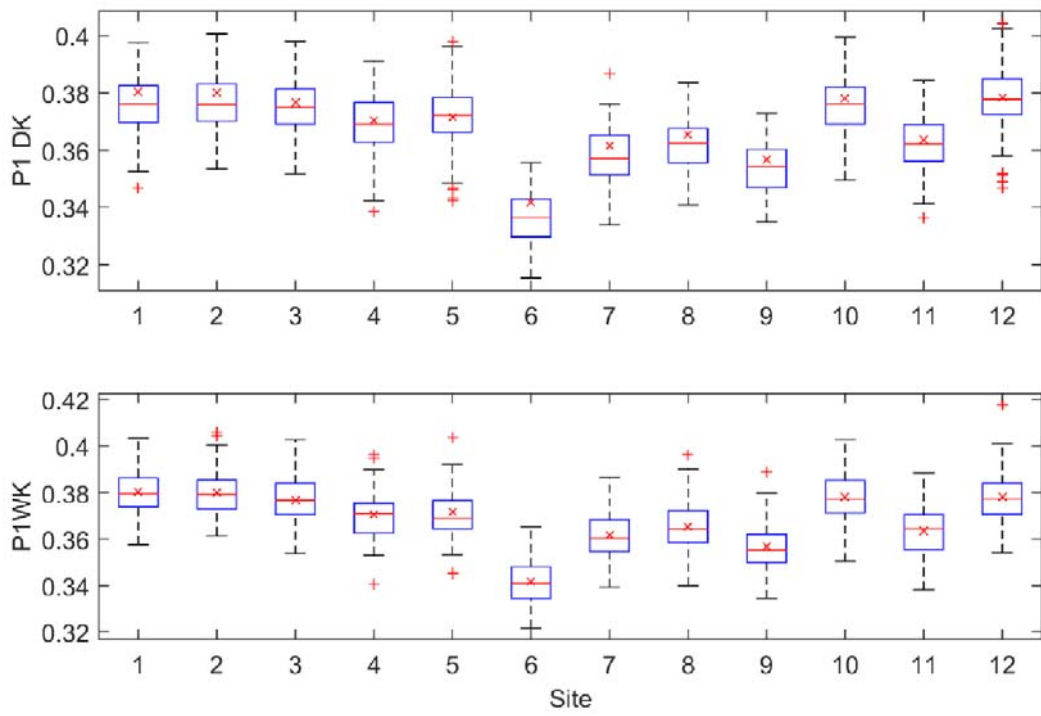
740

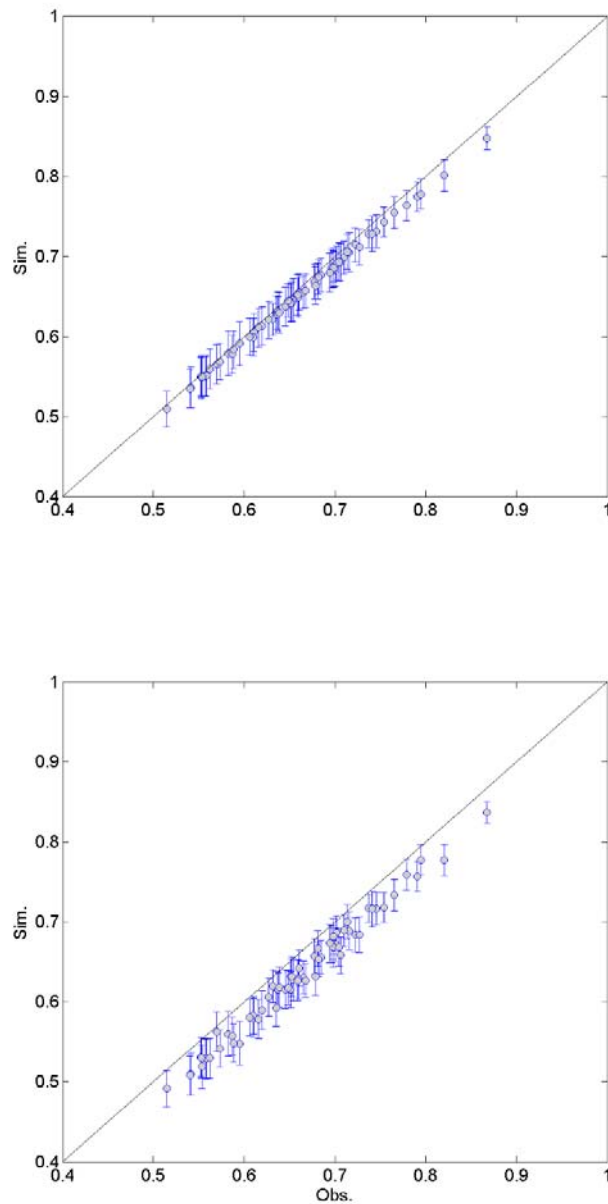
741

742

743

744

745

746

747 Figure 7. Boxplots of the P1 probability for the data simulated from the DKNNR model (top
748 panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12
749 selected weather stations from the Yeongnam province.

750

751 Figure 8. Scatterplot of cross-correlations between 12 weather stations for the observed data (X
752 coordinate) and the generated data (Y coordinate) generated from the DKNNR model (top panel)
753 and the MONR model (bottom panel). The cross-correlations from 100 generated series are
754 averaged for the filled circle and the errorbars upper and lower extended lines indicate the range
755 of 1.95×standard deviation.

756

757

758

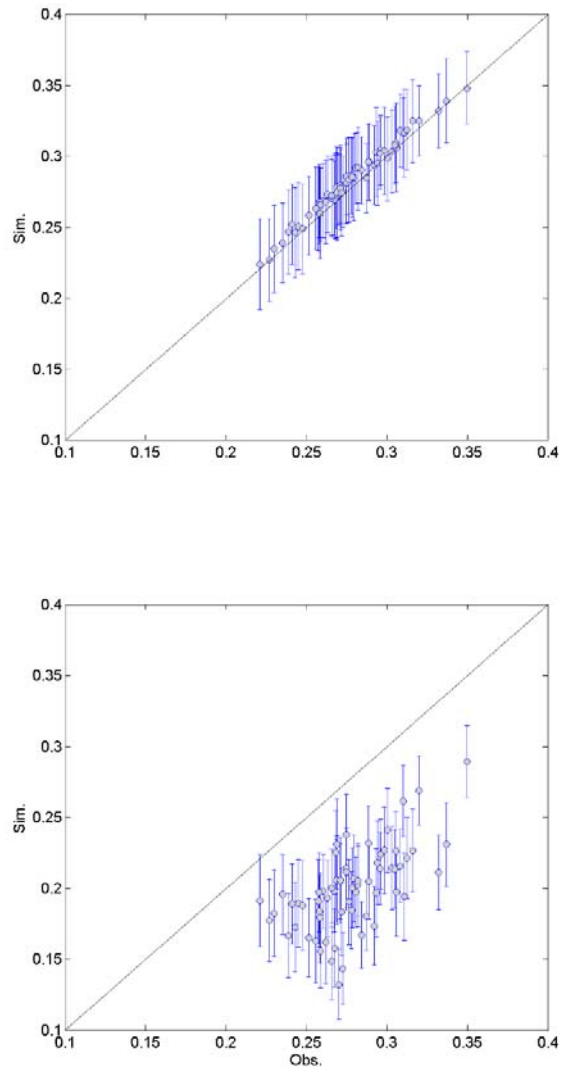



Figure 9. Scatterplot of lag-1 cross-correlations between 12 weather stations for the observed
data (X coordinate) and the generated data (Y coordinate) generated from the DKNNR model
(top panel) and the MONR model (bottom panel). The cross-correlations from 100 generated
series are averaged for the filled circle and the errorbars upper and lower extended lines indicate
the range of 1.95×standard deviation.

764

765

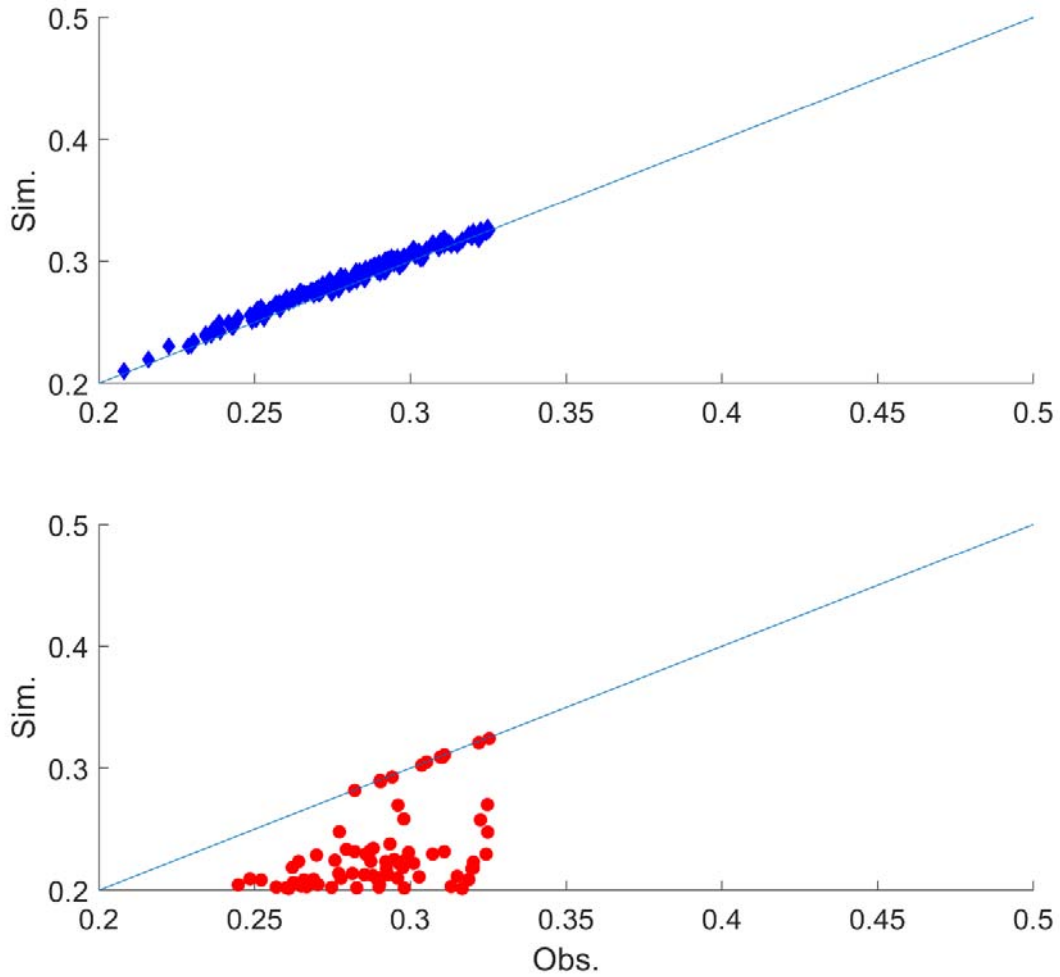



Figure 10. Scatterplot of lag-1 cross-correlations between 12 weather stations for the observed
data (X coordinate) and the generated data (Y coordinate) generated from the DKNNR model
(top panel) and the MONR model (bottom panel) with the whole year data not with the summer
season. The cross-correlations from 100 generated series are averaged.

770

771

772

773

774

775

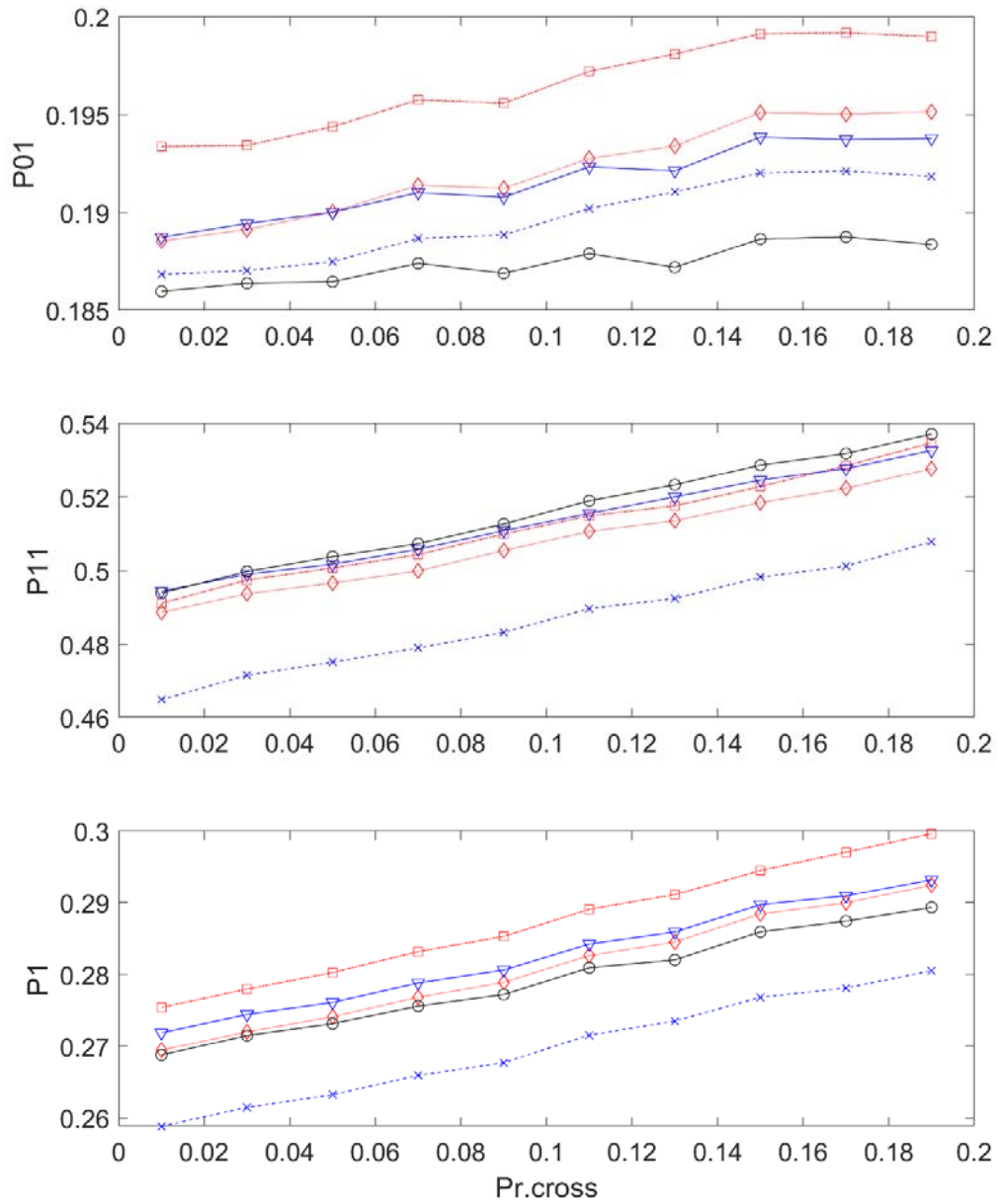
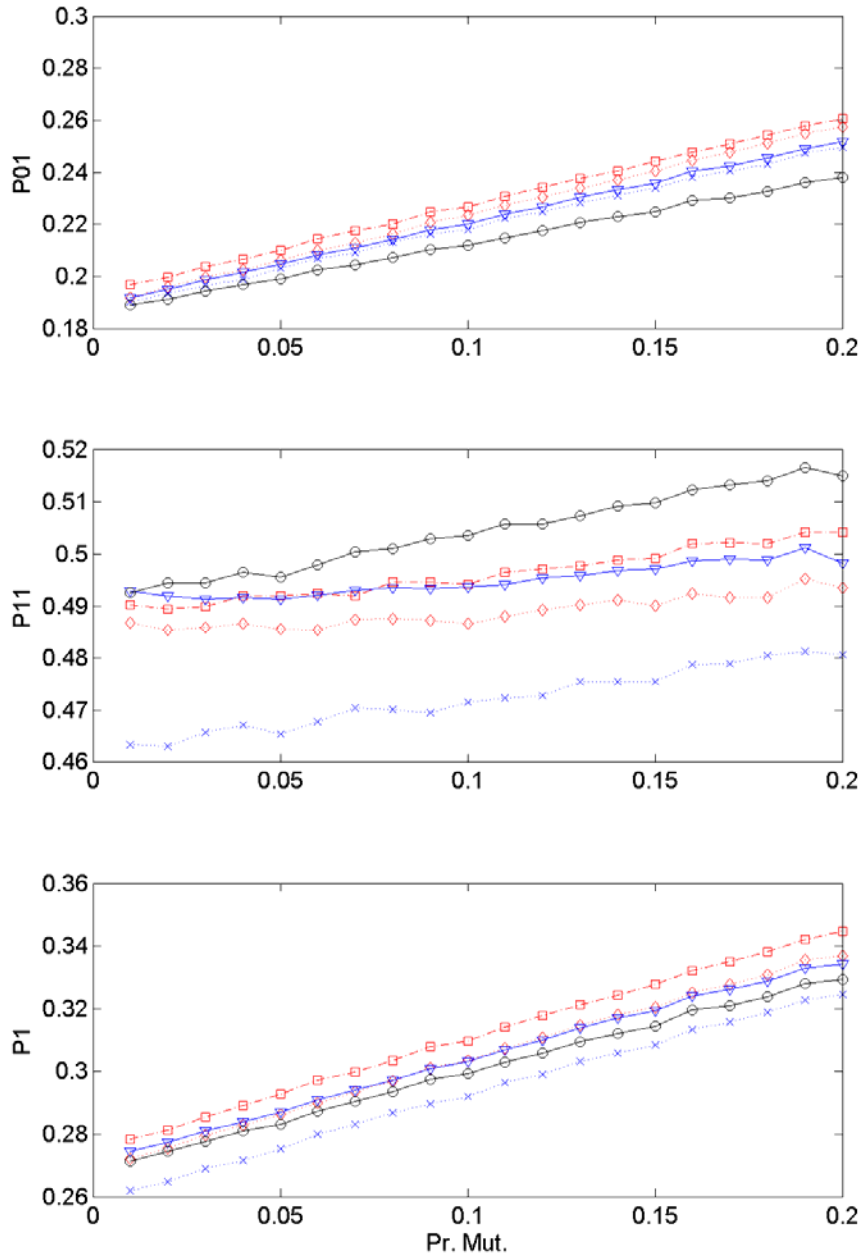



Figure 11. Transition probabilities and marginal distribution for the selected five stations along
with changing the cross-over probability P_{cr} with the condition that the candidate value is one
and the previous value is also one. See Eq.(15) for the detail.

779

780

781 Figure 12. Transition probabilities and marginal distribution along with changing the cross-over
782 probability with the condition that the mutation is processed only if the candidate value is one.
783 See Eq.(16) for the detail.

784
785

786

Table A 1. Example dataset of daily rainfall with 12 weather stations and 16 days for measured rainfall (mm) in the upper part of this table and its corresponding occurrences in the bottom part of this table.

Day	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1	2.0	2.9	1.2	0.0	0.0	1.8	4.0	8.9	2.0	4.6	1.3	0.6
2	52.6	39.8	47.2	17.4	11.8	31.0	30.0	33.7	52.0	57.8	37.0	17.5
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.2	1.0	1.4	1.9	12.3	0.0	0.0	0.0	0.7	3.1	3.5	8.1
6	14.8	0.2	0.8	0.2	5.0	0.0	0.0	18.0	0.0	0.0	0.6	3.1
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	1.0	0.0	0.4	0.0	3.8	0.0	0.1	0.0	0.0	0.0	0.0
11	7.1	6.4	12.8	12.8	13.6	2.3	2.0	5.4	6.0	7.3	16.4	20.3
12	0.0	0.0	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.0	4.3
13	10.0	1.6	11.6	14.3	1.5	5.4	0.0	0.0	2.5	0.0	2.7	16.1
14	2.3	0.0	0.7	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0
15	31.5	4.3	30.6	12.7	14.4	25.8	3.5	0.8	5.0	2.7	6.5	20.3
16	37.0	7.8	30.1	11.2	9.6	36.8	2.5	4.7	13.5	1.7	10.1	14.1
Day	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1	1	1	1	0	0	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	0	0	0	1	1	1	1
6	1	1	1	1	1	0	0	1	0	0	1	1
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	1	0	1	0	1	0	1	0	0	0	0
11	1	1	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	1	0	0	0	0	0	0	1
13	1	1	1	1	1	1	0	0	1	0	1	1
14	1	0	1	0	0	1	0	0	0	0	0	0
15	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1

790

791 Table A 2. Example dataset for estimating distances. The second row presents the current daily
792 precipitation occurrences for 12 stations and the rows below show the absolute difference
793 between the current occurrences (**Xc**) and the observed data in Table A 1. The last column
794 presents the distances in Eq. (11).

day	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Dist
Xc	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	
1	1	0	0	0	0	0	0	1	1	1	1	1	**6**
2	1	0	0	1	1	0	0	1	1	1	1	1	**8**
3	0	1	1	0	0	1	1	0	0	0	0	0	**4**
4	0	1	1	0	0	1	1	0	0	0	0	0	**4**
5	1	0	0	1	1	1	1	0	1	1	1	1	**9**
6	1	0	0	1	1	1	1	1	0	0	1	1	**8**
7	0	1	1	0	0	1	1	0	0	0	0	0	**4**
8	0	1	1	0	0	1	1	0	0	0	0	0	**4**
9	0	1	1	0	0	1	1	0	0	0	0	0	**4**
10	0	0	1	1	0	0	1	1	0	0	0	0	**4**
11	1	0	0	1	1	0	0	1	1	1	1	1	**8**
12	0	1	1	0	1	1	1	0	0	0	0	1	**6**
13	1	0	0	1	1	0	1	0	1	0	1	1	**7**
14	1	1	0	0	0	0	1	0	0	0	0	0	**3**
15	1	0	0	1	1	0	0	1	1	1	1	1	**8**
16	1	0	0	1	1	0	0	1	1	1	1	1	**8**

795

796

797

798 Table A 3. Example for selecting one sequence for $\mathbf{X_{c+1}}$. The second row presents the distances
799 in Table A 2. The third and fourth columns show the sorted days and distances for the smallest
800 distances to the largest in the second column. The fourth row presents the probabilities estimated
801 with Eq. (12). Note that there are six days whose distances are the same with each other. In this
802 case all the days are included and among six days, one is selected with equal probabilities.

Day	Dist.	Sorted Day	Sorted Dist	Prob
1	6	14	3	0.48
2	8	3	**4**	0.24
3	4	4	**4**	0.16
4	4	7	**4**	0.12
5	9	8	**4**	
6	8	9	**4**	
7	4	10	**4**	
8	4	1	6	
9	4	12	6	
10	4	13	7	
11	8	2	8	
12	6	6	8	
13	7	11	8	
14	3	15	8	
15	8	16	8	
16	8	5	9	

803

804

805 Table A 4. Example for GA mixture for X_{c+1}. The second and third rows present two selected
806 sets, while the third row shows the final set for X_{c+1} with the crossover at S6 and S8 and the
807 mutation for S12.

	Assigned day, p	Selected day, $p+1$	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Set1	14	**15**	1	0	0	1	1	0	0	1	1	1	1	1
Set2	4	**5**	1	0	0	1	1	1	1	0	1	1	1	1
Final			1	0	0	1	1	1	0	0	1	1	1	**0**

808
809

810