Discrete k-nearest neighbor resampling for simulating multisite precipitation occurrence and 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

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

28 Stochastic weather simulation models are commonly employed in water resources management

29 agricultural applications, forest management, transportation management, and recreational

30 activities. Stochastic simulation of multisite precipitation occurrence is a challenge because of its

31 intermittent characteristics as well as spatial and temporal cross-correlation. Although the

32 multisite occurrence model with standard normal variate (MONR) has been used for preserving

33 key precipitation statistics and contemporaneous correlation, it does not reproduce lagged

34 crosscorrelation between stations so long stochastic simulation is required. Employing a

35 nonparametric technique, k-nearest neighbor resampling (KNNR) and coupling it with Genetic

36 Algorithm (GA), this study proposes a novel simulation method for multisite precipitation

37 occurrence, overcoming the shortcomings of the MONR model. The novel discrete version of

38 KNNR (DKNNR) model was developed and its modification for the study of climatic change

39 adaptation was tested. The datasets simulated from both the DKNNR model and the MONR

40 model were evaluated using a number of statistics, such as occurrence and transition probabilities

41 as well as temporal and spatial cross-correlations. Results showed that the proposed DKNNR

42 model simulated multisite precipitation occurrence, preserving the lagged crosscorrelation

43 between sites. When climate change was considered, the model performed satisfactorily, but

44 further improvement is required to more accurately simulate specific variations of the occurrence

45 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

69   and temporal dependence in the occurrence and amount of precipitation introduces further

70   complexity in multisite simulation.

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

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

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

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

75   1977). Compared to the MC model that requires a significant number of parameters for

76   generating multisite occurrence, the multisite occurrence model proposed by Wilks (1998)

77   transforms the standard normal variate and simulates the sequence with multivariate normal

78   distribution, and then back-transforms the multivariate normal sequence to the original domain.

79   The model is able to reproduce the contemporaneous multisite dependence structure and lagged

80   dependence only for the same site but it requires a complex simulation process to estimate

81   parameters for each site and is unable to preserve lagged dependence between sites.

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

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

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

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

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

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

88   reproduction of similar populations by mixing the simulated datasets. While KNNR is employed

89   to find historical analogues of multisite occurrence similar to the current status of a simulation

90   series, GA is applied to use its skill to generate a new descendant from the historical parent

91   chosen with the KNNR. In this procedure, the multisite occurrence of precipitation can be

92   simulated while preserving spatial and temporal correlations. Meta-heuristic techniques, such as

93   GA, have been popularly employed in a number of hydrometeorological applications (Chau,

94   2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013). Although a number of

95   variants of KNNR-GA have been applied (Lee et al., 2012; Lee and Park, 2017), none of them

96   can simulate multisite occurrence of precipitation whose characteristics are binary and

97   temporally and spatially related.

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

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

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

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

102   between stations as well as the basic occurrence probabilities. The proposed nonparametric

103   model is compared with the popular model proposed by Wilks (1998). Even though the multisite

104   occurrence data generated from the Wilks model preserves various statistical characteristics of

105   the observed data well, significant underestimation of lagged cross-correlation still exists.

106   Furthermore, the relation between standard normal variable and occurrence variable relies on

107   long stochastic simulation.

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

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

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

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

112   presented in section 7.

2. Background

2.1. Single site occurrence modeling

Let X_t^s represent the occurrence of daily precipitation for a location s ($s=1,\ldots, S$) on day t ($t=1,\ldots, n$; n is the number observed days) and let X_t^s be either zero for dry day or one for wet day. The first order Markov chain model for X_t^s is defined with the assumption that the occurrence probability of a wet day is fully defined by the previous day as

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

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

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

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

The marginal distributions of TPM (i.e. p_0 and p_1) can be expressed with TPM and its condition of $p_0 + p_1 = 1$ as:

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

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

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

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

130   probabilities as:

131
$$\rho_1(s,s) = p_{11}^s - p_{01}^s \tag{6}$$

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

133
$$X_t^s = \begin{cases} 1 & \text{if } u_t^s \le p_{i1}^s \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

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

135   1). Wilks (1998) suggested a different method using a standard normal random number w_t^s

136    $\sim N[0,1]$ as

137
$$X_t^s = \begin{cases} 1 & \text{if } w_t^s \le \Phi^{-1}(p_{i1}^s) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

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

139   **2.2. Multisite occurrence modeling**

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

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

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

143
$$\tau(q,s) = corr[w_t^q, w_t^s] \tag{9}$$

144   Also, the correlation of the original occurrence variate is

145
$$\rho(q,s) = corr[X_t^q, X_t^s] \tag{10}$$

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

147 precipitation occurrence is straightforward. Multivariate standard normal distribution is used

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

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

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

151 $\tau(q,s)$ from $\rho(q,s)$. A long sequence of the occurrences is simulated with different values of

152 $\tau(q,s)$ and its corresponding correlation of the original domain $\rho(q,s)$ is estimated with the

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

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

155 employed for parameter estimation for a real application.

3. DKNNR

156

3.1. DKNNR modeling procedure

157

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

159 precipitation variable with k-nearest neighbor resampling (KNNR) technique (Lall and

160 Sharma, 1996; Lee and Ouarda, 2011; Lee et al., 2017) for a discrete case (denoted as

161 Discrete KNNR; DKNNR) is proposed by combining a mixture mechanism with Genetic

162 Algorithm (GA). Provided the number of nearest neighbors, k, is known, the discrete k-

163 nearest neighbor resampling with genetic algorithm is done as follows:

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

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

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

167 $$D_i = \sum_{s=1}^{S} \left| X_c^s - x_i^s \right| \tag{11}$$

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

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

170   distances.

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

172   given by

173 $$w_m = \frac{1/m}{\sum_{j=1}^{k} 1/j}, \qquad m = 1,\ldots, k \tag{12}$$

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

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

176   number between 0 and S. For example, if S=2 and X_c^1=0 and X_c^2=1, the two

177   sequences have 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

178   case, a random selection procedure is required to take into account the cases with

179   the same quantity. One particular time index is randomly selected with equal

180   probabilities among the time indices of the same distances. Note that instead of the

181   random selection, one can always use the first one. In such a case, only one

182   historical combination of multisite occurrences will be selected.

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

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

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

186   Otherwise, skip this step.

187   (6-1) Reproduction: Select one additional time index using steps (1) through (4)

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

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

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

191   process by finding another individual that has similar characteristics to the

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

193   be mated and will produce a new descendant.

194   (6-2) Crossover: Replace each element x_{p+1}^s with $x_{p^*+1}^s$ at probability P_{cr}, i.e.,

195
$$X_{c+1}^s = \begin{cases} x_{p^*+1}^s & \text{if } \varepsilon < P_{cr} \\ x_{p+1}^s & \text{otherwise} \end{cases} \qquad (13)$$

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

197   new occurrence vector whose elements are similar to the historical ones is

198   generated.

199   (6-3) Mutation: Replace each element (i.e., each station, $s=1,...,S$) with one

200   selected from all the observations of this element for $i=1,...,n$ with

201   probability P_m, i.e.,

202
$$X_{c+1}^s = \begin{cases} x_{\xi+1}^s & \text{if } \varepsilon < P_m \\ x_{p+1}^s & \text{otherwise} \end{cases} \qquad (14)$$

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

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

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

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

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

208   for a simulation purpose.

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

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

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

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

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

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

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

216 as a prediction problem. However, the current multisite occurrence simulation does not

217 necessarily require an accurate value prediction and not much difference in simulation using the

218 simple heuristic approach has been reported. Also, this heuristic approach of k selection has been

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

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

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

222   ## 3.2. Adaptation to climate change

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

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

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

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

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

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

229   the condition when applied.

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

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

232
$$X_{c+1}^s = \begin{cases} x_{p*+1}^s & \text{if } \varepsilon < P_{cr} \ \& \ x_{p*+1}^s = 1 \ \& \ X_c^s = 1 \\ x_{p+1}^s & \text{otherwise} \end{cases} \qquad (15)$$

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

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

235
$$X_{c+1}^s = \begin{cases} x_{\xi+1}^s & \text{if } \varepsilon < P_m \text{ and } x_{\xi+1}^s = 1 \\ x_{p+1}^s & \text{otherwise} \end{cases} \qquad (16)$$

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

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

4. Study area and data description

For testing the occurrence model, 12 weather stations were selected from Yeongnam province which is located in the southeastern part of South Korea, as shown in Figure 1. Information on longitude and latitude (fourth and fifth columns) as well as order index and the identification number (first and second columns) of these stations operated by Korea Meteorological Administration with the area name (third column) is shown in Table 1. The employed precipitation dataset presents strong seasonality, since this area is dry from late fall to early autumn and humid and rainy during the remaining seasons, especially in summer. The employed stations are not far from each other, at most 100 km apart, and not much high mountains are located in the current study area. Therefore, this region can be considered as a homogeneous region (Lee et al., 2007).

Figure 1 illustrates the locations of the selected weather stations. All the stations are inside Yeongnam province which consists of two different regions as north and south Gyeongsang as well as the self-governing cities of Busan, Daegu, and Ulsan. Most of the Yeongnam region is drained to Nakdong River. To validate the proposed model appropriately, test sites must be highly correlated with each other as well as have significant temporal relation. The stations inside the Yeongnam area cover one of the most important watersheds, the Nakdong River basin, where the Nakdong River passes through the entire basin and its hydrological assessments for agriculture and climate change have a particular value in flood control and water resources management such as floods and droughts.

It is important to analyze the impact of weather conditions for planning agricultural operations and water resources management, especially during the summer season, because

260 around 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

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

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

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

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

5. Application

267   To analyze the performance of the proposed DKNNR model, the occurrence of

268 precipitation was simulated. The DKNNR simulation was compared with that of the MONR

269 model. For each model, 100 series of daily occurrence with the same record length were

270 simulated. The key statistics of observed data and each generated series, such as transition

271 probabilities (P_{11}, P_{01}, and P_1) and cross-correlation (see Eq.(10)), were determined. The MONR

272 model underestimated the lag-1 cross-correlation, as indicated by Wilks (1998). In the current

273 study, this statistic was analyzed, since a synoptic scale weather system often results in lagged

274 cross-correlation for daily precipitation data (Wilks, 1998). It was formulated as

275 $$\rho_1(q,s) = corr[X_{t-1}^q, X_t^s] \tag{17}$$

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

277 expressed as:

278 $$RMSE = \left(\frac{1}{N} \sum_{m=1}^{N} (\Gamma_m^G - \Gamma^h)^2 \right)^{1/2} \tag{18}$$

279   where N is the number of series (here 100), Γ_m^G is the statistic estimated from the mth generated

280   series, while Γ^h is the statistic for the observed data. Note that lower RMSE indicates better

281   performance representing the summarized error of a given statistic of generated series from the

282   statistic of the observed data.

283   The 100 simulated statistic values were illustrated with boxplots to show their variability as

284   shown in Figure 5 - Figure 7. The box of boxplot represents the interquartile range (IQR) ranging

285   25 percentile to 75 percentile. The whiskers extend to up and down 1.5×IQR. Data beyond the

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

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

288   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

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

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

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

296   P_{01}, and P_1) between the simulated data and the observed was used, since those probabilities are

297   key statistics that the simulated data must match with the observed data and no parameterization

298   of these probabilities was made for the current DKNNR model. Results are shown in Figure 2

299   and Figure 3 for P_{cr} and P_m, respectively. For P_{cr} in Figure 2, the probability of 0.02 shows the

300   smallest RMSE in all transition and limiting probabilities. The RMSE of P_m in Figure 3 shows a

301   slight fluctuation along with P_m. However, all three probabilities (P_{11}, P_{01}, and P_1) have

302   relatively small RMSEs in P_m =0.003. Therefore, the parameter set 0.02 and 0.003 was chosen

303   for P_{cr} and P_m, respectively, and employed in the current study. We also tested the simulation

304   without the GA mixing procedure (results not shown). The results showed that no better result

305   could be found from the simulation without GA mixing. The necessity of the GA mixing is

306   further discussed in the following.

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

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

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

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

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

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

313   the simulated data.

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

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

316   multisite precipitation occurrence is decided by the spatial distribution of a precipitation weather

317   system. A humid air mass can be distributed randomly relying on wind velocity and direction as

318   well as surrounding air pressure. In general, any combinations of wet and dry stations can be

319   possible, especially when the simulation continues infinitely. Therefore, the patterns of simulated

320   data must be allowed to have any possible combinations, here 4096 even if it has not been

321   observed from the historical records. Also, its probability to have this new pattern must not be

322   high since it has not been observed in the historical records and this can be taken into account by

323   low probability of the crossover and mutation.

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

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

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

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

328   candidate patterns to produce new patterns. For example, in the three station case, a new pattern

329   101 can be produced from two observed occurrence candidates of 001 and 100 by the crossover

330   of the first value of 001 to the first value of 100 (i.e. 001 $\rightarrow$101), which is not in the observed

331   data.

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

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

334   4096. We checked how many of possible cases are not found in the observed data. The result

335   shows that 3379 cases are not observed at all for the entire cases as shown in Figure 4.

336   We further investigated how many new patterns are generated with the probabilities

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

338   DKNNR with the GA mixing shows that the number 3379 was reduced to 1200, which is not in

339   the dataset among the 4096 possible patterns. Therefore, more than 2000 new patterns were

340   simulated with the GA mixing process. The KNNR model without the GA mixing does not

341   produce any new patterns in the 100 sequences with the same length of the historical data.

342

6.2. Occurrence and transition probabilities

The data simulated from the proposed DKNNR model and the existing MONR model were analyzed. The estimated transition probabilities (P_{11} and P_{01} in Eq. (3)) as well as the occurrence probability (P_1 in Eq. (5)) are shown in Table 2 and Figure 5 - Figure 7 for the observed data and the data generated from the DKNNR and MONR models. In Table 2, the observed statistic shows 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 P_1 and site 12 shows the highest P_{11}.

As shown in Figure 5, the probability P_{11} of the observed data shows that sites 6, 7, 8, and 9 located in the northern part of the region exhibited lower consistency (i.e. consecutive rainy days) than did the other sites, while sites 5 and 12 had higher probability of P_{11} than did other sites. Both models preserved well the observed P_{11} statistic. It seems that the MONR model had a slightly better performance, since this statistic is parameterized in the model as shown in section 2.2 and that is the same for P01 and P1 as shown in Figure 6 and Figure 7. Note that the MONR model employed the transition probabilities in simulating rainfall occurrence, while DKNNR model did not. The occurrence probability P_1 can be described with the combination of transition probabilities as in Eq. (5). Even though the transition probabilities were not employed in simulating rainfall occurrence, the DKNNR model preserved this statistic fairly well.

In the DKNNR modeling procedure, the simple distance measurement in Eq. (11) allows to preserve transition probabilities in that the following multisite occurrence is resampled from the historical data whose previous states of multisite occurrence (x_i^s) are similar to the current simulation multisite occurrence (X_c^s). This summarized distance (D_i) is an essential tool in the proposed DKNNR modeling. The condition of the current weather system is memorized and the system is conditioned on simulating the following multisite occurrence with the distance

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

367   system of the following day.

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

369   sites 1, 2, and 3 located in the middle part of the Yeongnam province showed a higher

370   probability than did other sites. A slight underestimation was seen for sites 2 and 11 but it was

371   not critical, since its observed value with a cross mark was close to the upper IQR representing

372   75 percentile.

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

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

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

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

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

378   statistic was preserved fairly well.

379   **6.3. Cross-correlation**

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

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

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

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

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

385   event is highly related with lagged cross-correlation.

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

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

388   moves according to the surrounding pressure and wind direction that dynamically change within

389   a day or week. Therefore, we analyzed the lag-1 cross-correlation in the current study as the

390   representative lagged correlation structure.

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

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

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

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

395   correlation. This observed cross-correlation was well preserved in the data generated from both

396   DKNNR and MONR models, as shown in Figure 8 as well as Table 4 and Table 5. However,

397   consistently slight but significant underestimation of cross-correlation was seen for the data

398   generated by the MONR model (see the bottom panel of Figure 8). Note that the errobars are

399   extended to upper and lower lines of the circles to 1.95×standard deviation. The difference of

400   RMSE in Table 6 showed this characteristic, as most of the values were positive, indicating that

401   the proposed DKNNR model performed better for cross-correlation.

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

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

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

405   magnitudes even for autocorrelation. This implies that the lag-1 cross-correlation among the

406   selected sites was as strong as the autocorrelation and as much as the transition probabilities P_{01}

407   and P_{11}, thereof.

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

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

410   significant underestimation, as seen in the bottom panel of Figure 9. The difference of RMSE

411   shown in Table 8 reflects this behavior. In the bottom panel of Figure 9, some of the lag-1 cross-

412   correlations were well preserved, that was aligned with the base line. From Table 8, the MONR

413   model reproduced the autocorrelations well with the shaded values. It is because the lag-1

414   autocorrelation was indirectly parameterized with the transition probabilities of P_{11} and P_{01} as in

415   Eq. (6). Other than this autocorrelation, the lag-1 cross-correlation was not reproduced well with

416   the MONR model. This shortcoming was mentioned by Wilks (1998). Meanwhile, the proposed

417   DKNNR model preserved this statistic without any parameterization.

418   We further tested the performance measurements of MAE and Bias. The estimates showed

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

434   not within the current scope and focus.

435   **6.4. Adaptation to climate change**

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

437   critical factor, since one of the major applications of the models is to assess the impact of climate

438   change. Therefore, we tested the capability of the proposed model in the current study by

439   adjusting the probabilities of cross-over and mutation as in Eqs.(15) and (16). A number of

440   variations can be made with different conditions.

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

442   increasing the crossover probability P_{cr} from 0.01 to 0.2 with the condition that that the

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

444   among the 12 stations (from station 1 to station 5, see Table 1 for details). The stations were

445   limited in this analysis due to computational time. In each case 100 series were simulated. The

446   average value of the simulated statistics is presented in the figure. It is obvious that the transition

447   probability P_{11} increased as intended along with the increase of P_{cr}. As expected from Eq. (5), P_1

448   presents that the change of P_1 is highly related to P_{11}. However, the probability P_{01} fluctuated

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

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

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

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

453   The change of P11 was not related with other probabilities. The combination of the adjustment of

454   P_{cr} and P_m with a certain condition to the previous state will allow the specific adaptation for

455   simulating future climatic scenarios.

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

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

458   Further modeling development on systematically varying temporal and spatial cross-correlations

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

460   # 7. Conclusions

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

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

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

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

465   probabilities and cross-correlation as well as lag-1 cross-correlation are estimated for both

466   models. Better preservation of cross-correlation and lag-1 cross-correlation with the DKNNR

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

468   seasons than the summer season), the estimated cross-correlation matrix is not a positive semi-

469   definite matrix so the multivariate normal simulation is not applicable for the MONR model,

470   because the tested sites are close to each other with high cross-correlation.

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

472   and transition probabilities fairly and preserves the cross-correlations better than the existing

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

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

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

476   multisite precipitation occurrence.

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

478   by modifying the mutation and crossover probabilities P_m and P_{cr}. The results showed that the

479   proposed DKNNR model has the capability to adapt to the climate change scenarios, but further

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

481   the marginal and transition probabilities cannot address the climate change of regional

482   precipitation. The variation of temporal and spatial cross-correlation structure must be

483   considered to properly address the climate change of the regional precipitation system. Further

484   study on improving the model adaptability to climate change will be followed in the near

485   future.Also, the simulated multisite occurrence can be coupled with a multisite amount model to

486   produce precipitation events, including zero values. Further development can be made for

487   multisite amount models with a nonparametric technique, such as KNNR and bootstrapping.

488   **Code and Data Availability**

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

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

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

492   **Acknowledgment**

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

494   2018R1A2B6001799) funded by the Korean Government (MEST).

Appendix A: Example of DKNNR

495

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

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

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

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

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

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

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

503   for $i=1$,

504
$$D_1 = \sum_{s=1}^{S} \left| X_c^s - x_1^s \right| = |0-1| + |1-1| + \dots + |0-1| = 6$$

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

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

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

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

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

510   last column of Table A 3.

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

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

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

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

| 515 | Therefore, the selected day, p, is 14. The occurrences of the following day $p+1=15$ for |
| 516 | 12 stations are selected as in the second row of Table A 4. |

(4) For GA mixture, another set must be chosen as in step (3). Say $u=0.561$, which falls in [0.48 0.72]. The second one should be selected. However, there are a number of days with the same distances. Specifically, six days have the same distances with $D_i=4$. In this case, one among all six days is selected with equal probability. Assume that $p=4$ is selected and the following occurrences are selected, as shown in the third row of Table A 4.

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

(5-1) Crossover: For each station, a uniform random number (ε) is generated and compared with $P_c=0.1$ here. Say $\varepsilon =0.345$, then skip since $\varepsilon =0.345> P_c=0.1$. For $s=6$, assume the generated random number, ε (=0.051)$< P_c(=0.1)$ and then switch the 6th station value of Set 1 into the value of Set 2 (see Table A 4). The occurrence state of X_{c+1}^s turns into 1 from 0 as shown in the fourth row of Table A 4 as well as station 8.

(5-2) Mutation: For each station, a uniform random number (ε) is generated and compared with $P_m=0.01$. For $s=12$, assume $\varepsilon =0.009< P_m=0.01$ and switch the 12th station value of Set 1 with the one selected among all the observed 12th station values with equal probability (here the last column, $s=12$, of the bottom part of Table A 1, [1 1 0 0 … 1]). The occurrence state of X_{c+1}^{12} turns into 0 from 1 as shown in the fourth column of Table A 4.

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

537

References

539

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

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

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

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

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

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

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

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

548   2761-2776, 2001.

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

550   (Switzerland), 9, 2017.

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

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

553   research directions and future work, Engineering Applications of Computational Fluid

554   Mechanics, 12, 411-437, 2018.

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

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

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

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

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

560   model for precipitation occurrence, Journal of the Royal Statistical Society. Series C: Applied

561   Statistics, 48, 15-30, 1999.

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

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

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

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

566   downscaling model for daily precipitation combined by multivariate multiple linear regression

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

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

569   Meteorology, 16, 671-676, 1977.

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

571   of Climate, 12, 2528-2537, 1999.

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

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

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

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

576   2147-0, 2017. 1-10, 2017.

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

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

579   124, 91-102, 2016.

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

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

582   Lee, T. and Ouarda, T. B. M. J.: Stochastic simulation of nonstationary oscillation

583   hydro-climatic processes using empirical mode decomposition, Water Resources Research, 48,

584   1-15, 2012.

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

586   generator and an extreme value theory for bandwidth selection, Journal of Hydrology, 452-453,

587   161-171, 2012.

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

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

590   10.1007/s00382-017-3525-0, 2017. 1-19, 2017.

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

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

593   performance evaluation, Journal of Hydrology, 547, 498-516, 2017.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

608   river water temperature forecast model with a k-nearest neighbour approach, Hydrological

609   Processes, 26, 1302-1310, 2012.

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

611   through baseflow separation and binary-coded swarm optimization, Journal of Hydrology, 529,

612   1788-1797, 2015.

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

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

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

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

617   1390, 2013.

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

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

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

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

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

623   model, Journal of Hydrology, 210, 178-191, 1998.

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

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

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

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

628
629

630

631   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

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

634

635

636 Table 2. Occurrence and transition probabilities of observed data and data simulated by DKNNR
637 and MONR for 12 stations from Yeongnam province, South Korea, during the summer season.
638 Note that 100 sets with the same record length as the observed data were simulated and the
639 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

640
641
642

643   Table 3. Cross-correlation of observed data for 12 stations from Yeongnam province, South
644   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

645
646
647

648   Table 4. Averaged cross-correlation of the 100 simulated series from the DKNNR model for 12
649   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

650
651
652

653   Table 5. Averaged cross-correlation of 100 simulated series from the MONR model for 12
654   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

655
656
657
658

659 Table 6. The difference of RMSE of cross-correlation between MONR and DKNNR. Note that
660 the positive value indicates that the DKNNR model better performs in preserving the cross-
661 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

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

664

665

666

667
668

669 Table 7. Lag-1 cross-correlation of observed data for 12 stations from Yeongnam province,
670 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

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

672

673

674   Table 8. The difference of RMSE of lag-1 cross-correlation between MONR and DKNNR. Note
675   that a positive value indicates that the DKNNR model better performs in preserving lag-1 cross-
676   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

677

678

679

680 Table 9. Bias of lag-1 cross-correlation of the generated data from the DKNNR model. Note that

681 a positive value indicates the overestimation of lag-1 cross-correlation, while a negative value

682 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

683 term.

684

	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

685

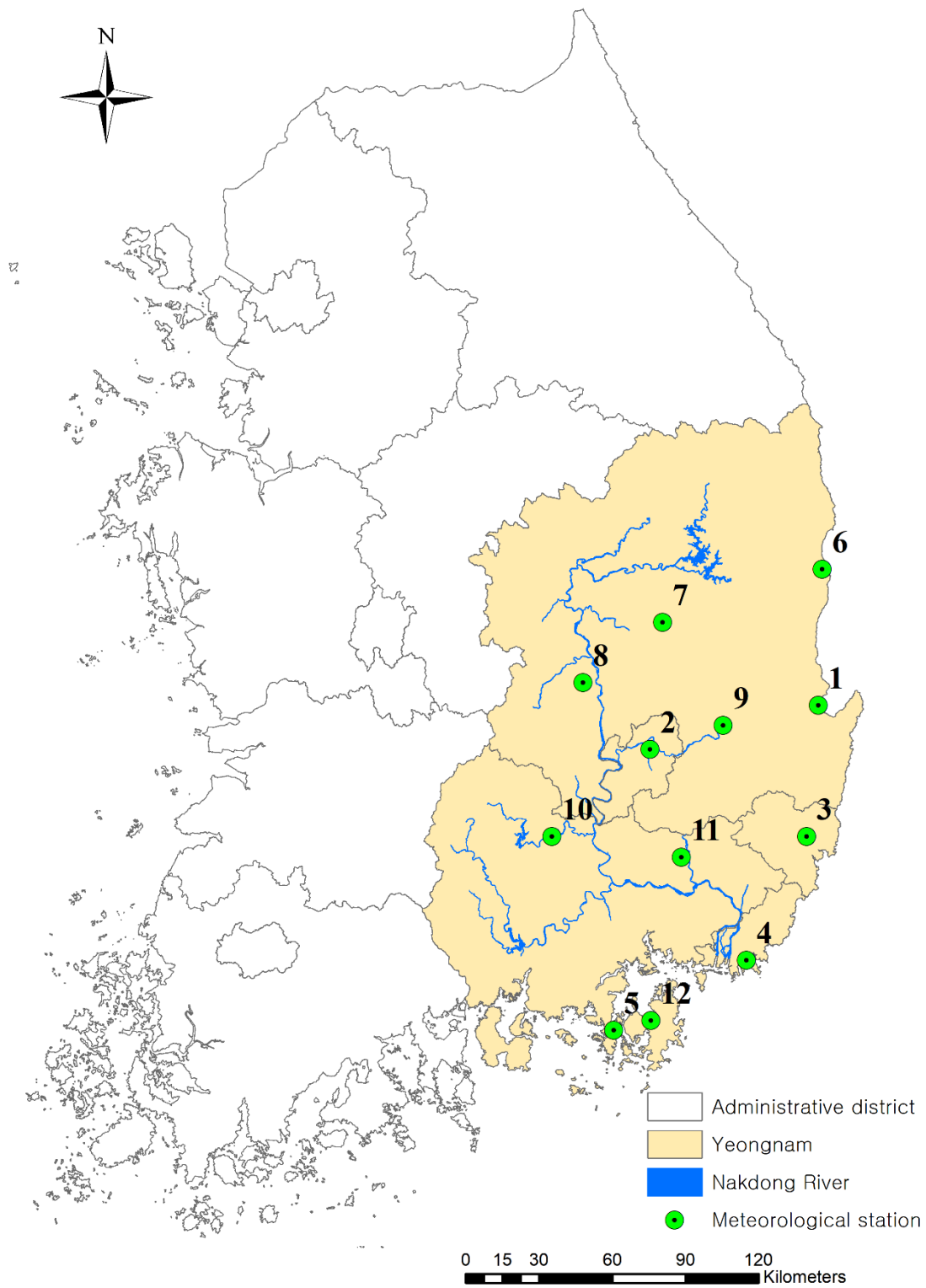
686

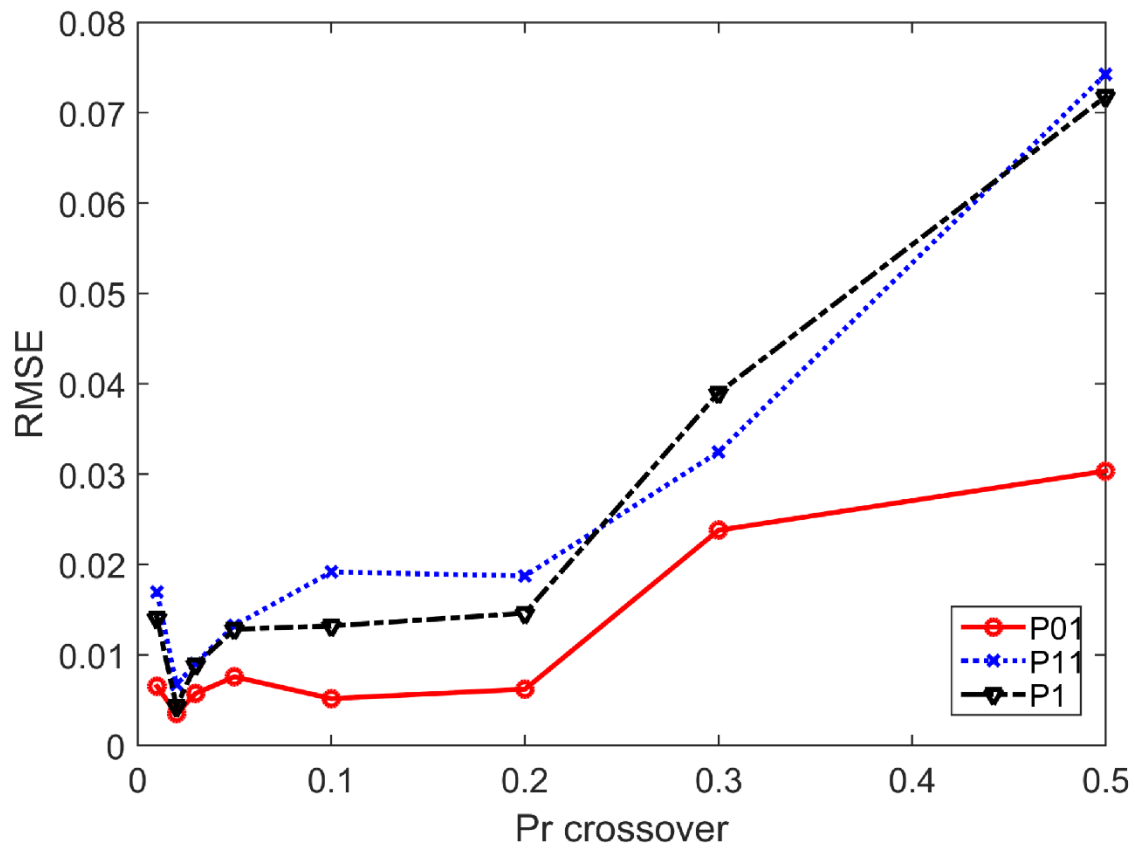
687 Table 10. Bias of lag-1 cross-correlation of the generated data from the Wilks model. Note that a
688 positive value indicates the overestimation of lag-1 cross-correlation, while a negative value
689 shows underestimation.

	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

690

691
692

693

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

696

697
698 Figure 2. Testing for different probabilities of crossover Pcr. RMSE is estimated for all the tested
699 12 stations for each transition and limiting probability of the simulated data with the record
700 length of 100,000.

701

702

705

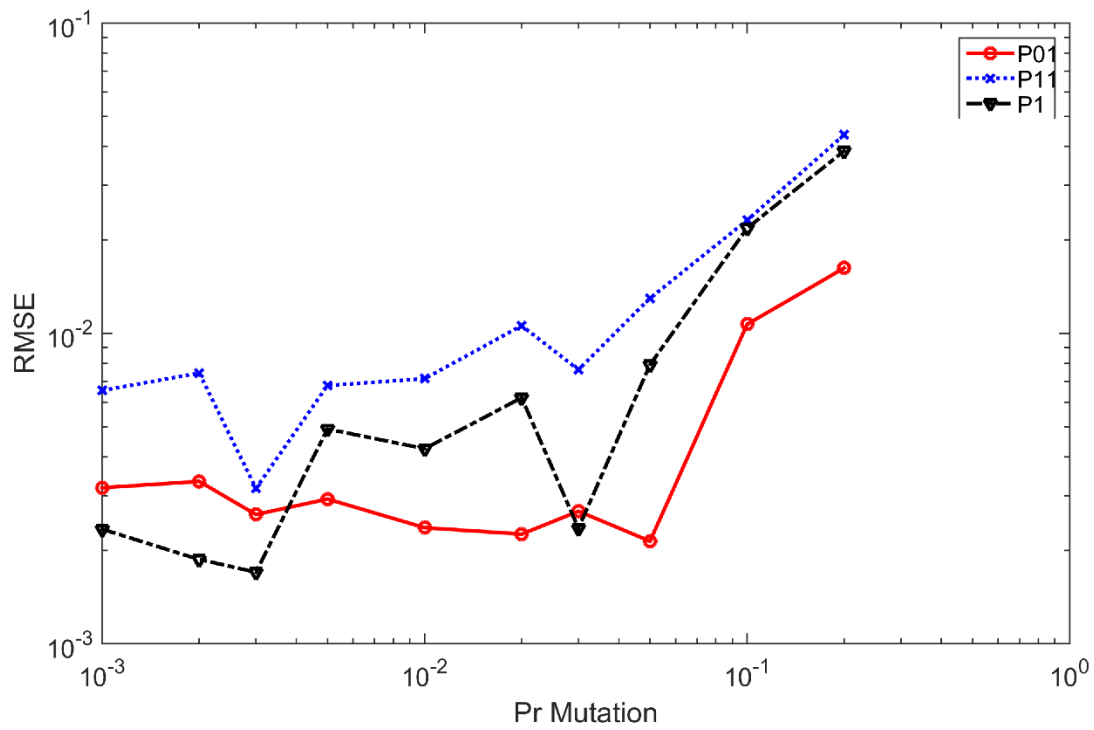
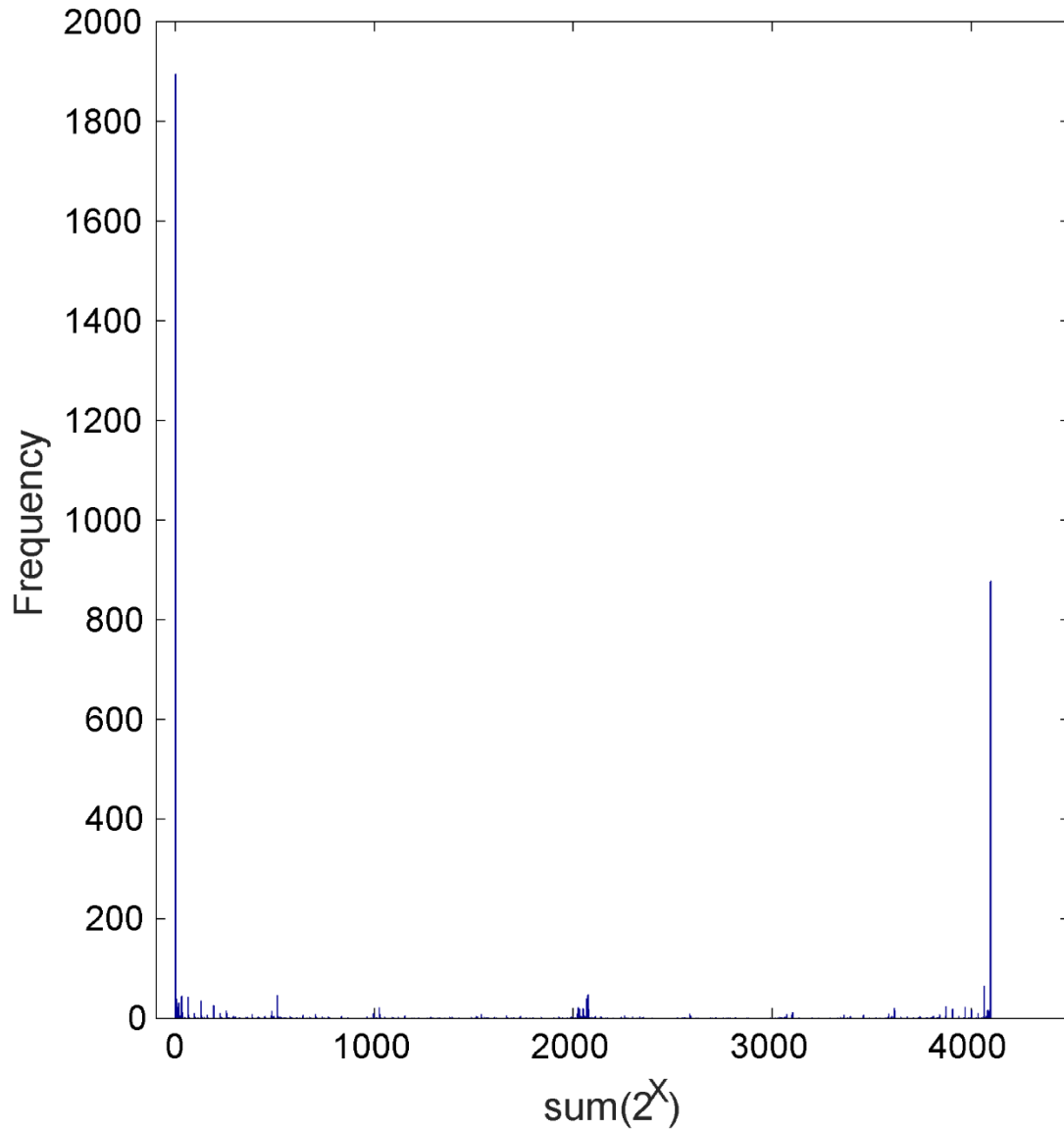
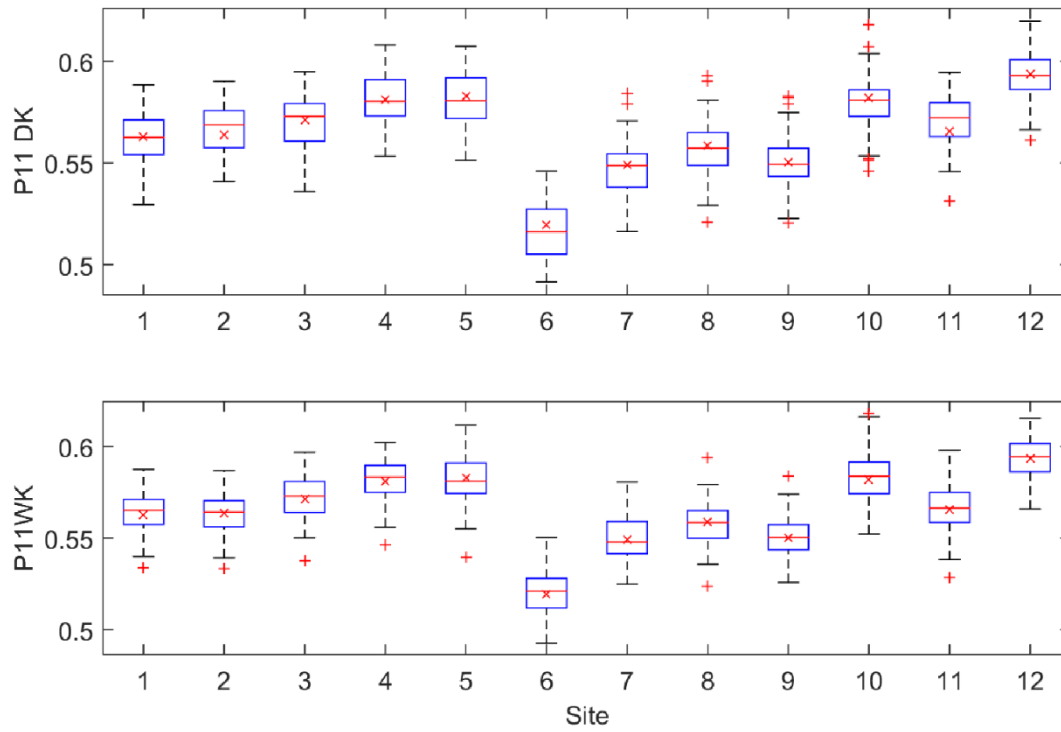
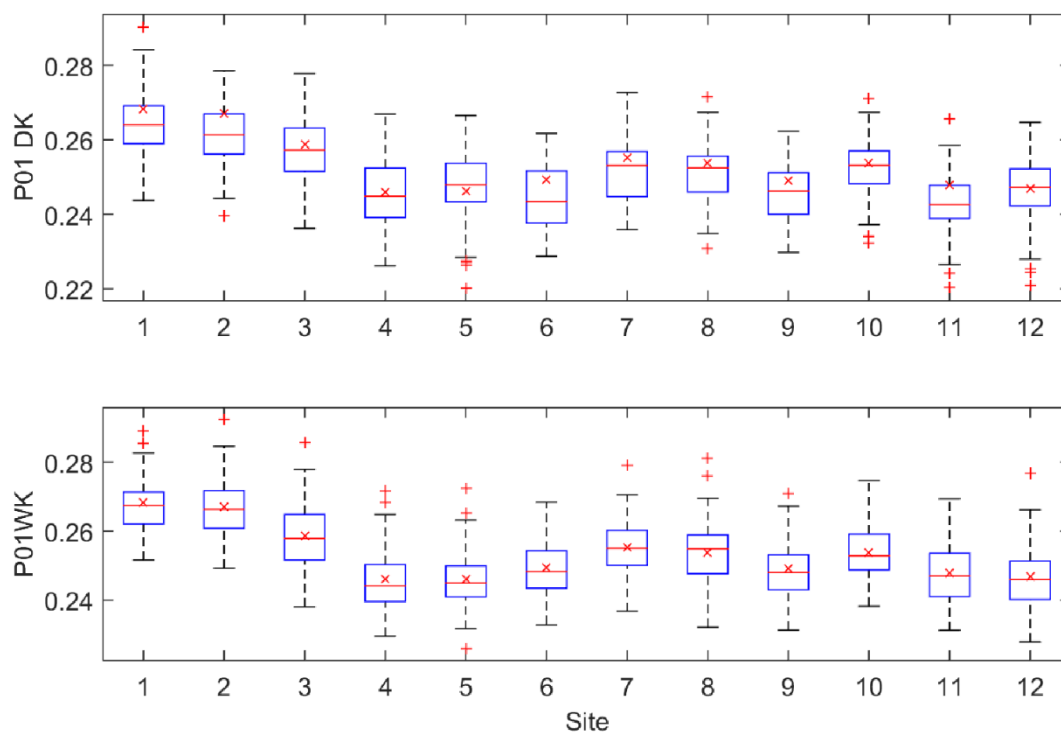



Figure 3. Testing for different probabilities of mutation P_m. RMSE is estimated for all the tested 12 stations for each transition and limiting probability of the simulated data with the record length of 100,000.

711
712 Figure 4. Frequency of the observed patterns among all the possible cases (2^{12}=4096). The X
713 coordinate indicates each pattern with the numbering of the binary number system. All zero (0)
714 and all one (4095) has the largest and second largest numbers of frequency as 1894 and 877,
715 respectively as expected meaning all dry and all wet stations. Note that the bars are very sporadic
716 indicating a number of occurrence patterns are not observed.

717
718

719

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

723

724

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

728

729

730

731

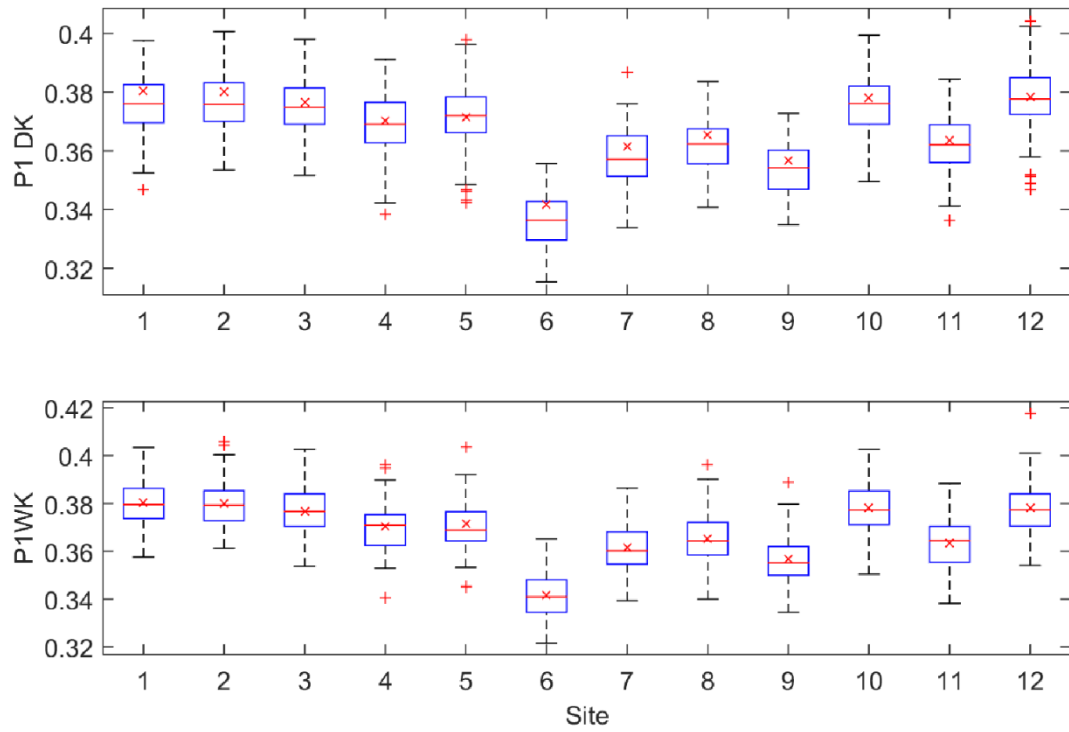
732

733

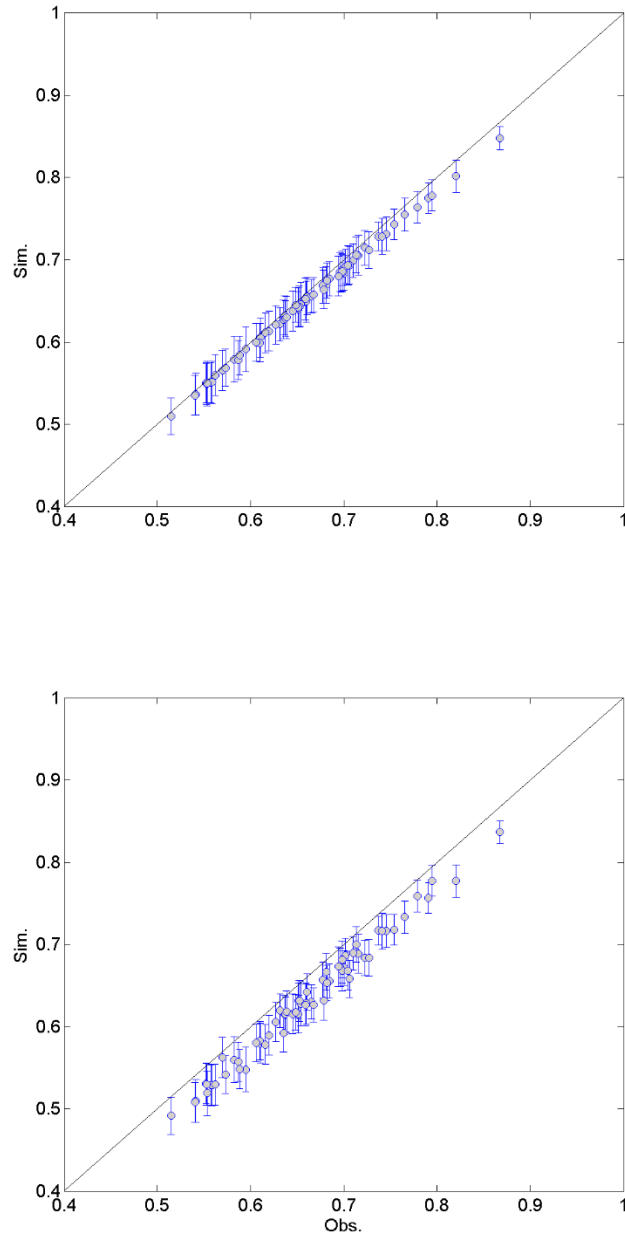
734

735

736

737

738

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

742

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

748

749

750

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.

756

757

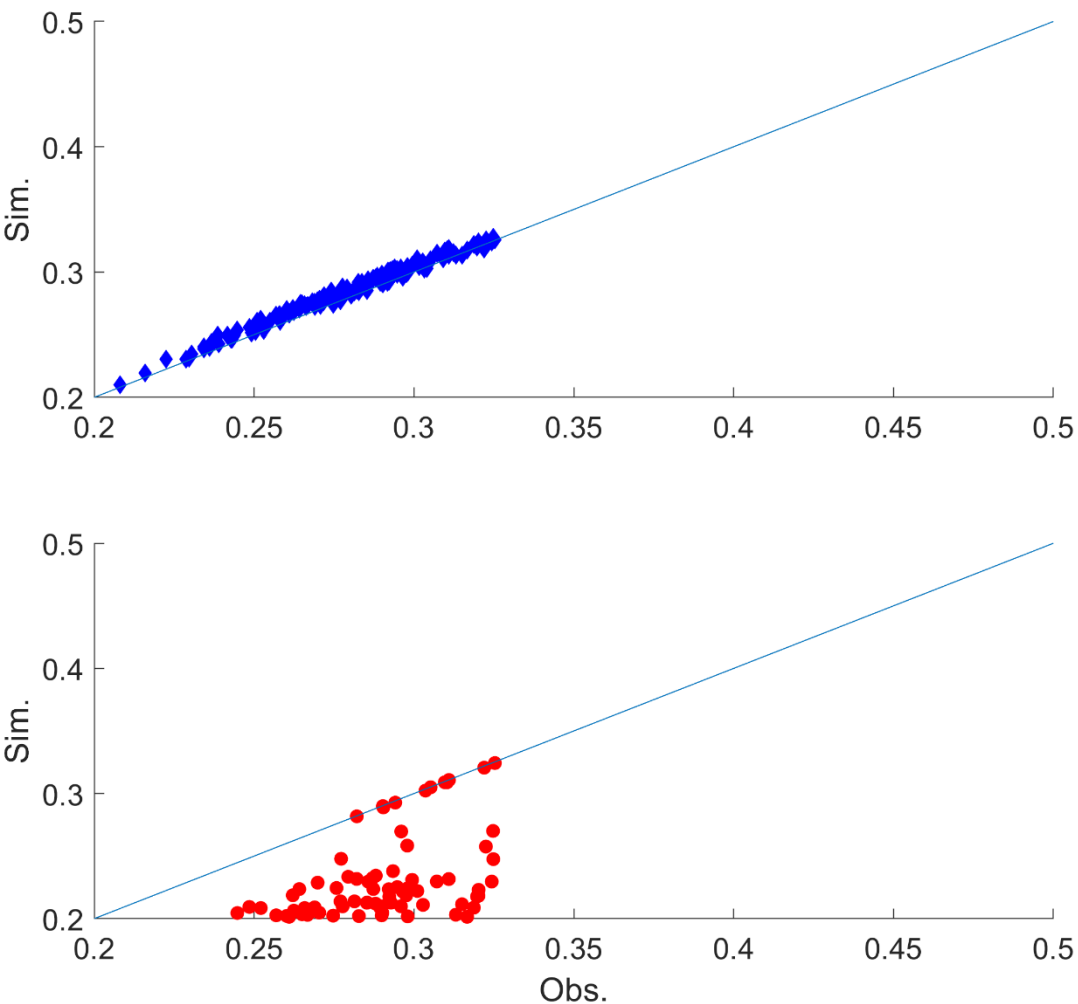



Figure 10. Scatterplot of lag-1 cross-correlations between 12 weather stations for the observed data (X coordinate) and the generated data (Y coordinate) generated from the DKNNR model (top panel) and the MONR model (bottom panel) with the whole year data not with the summer season. The cross-correlations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

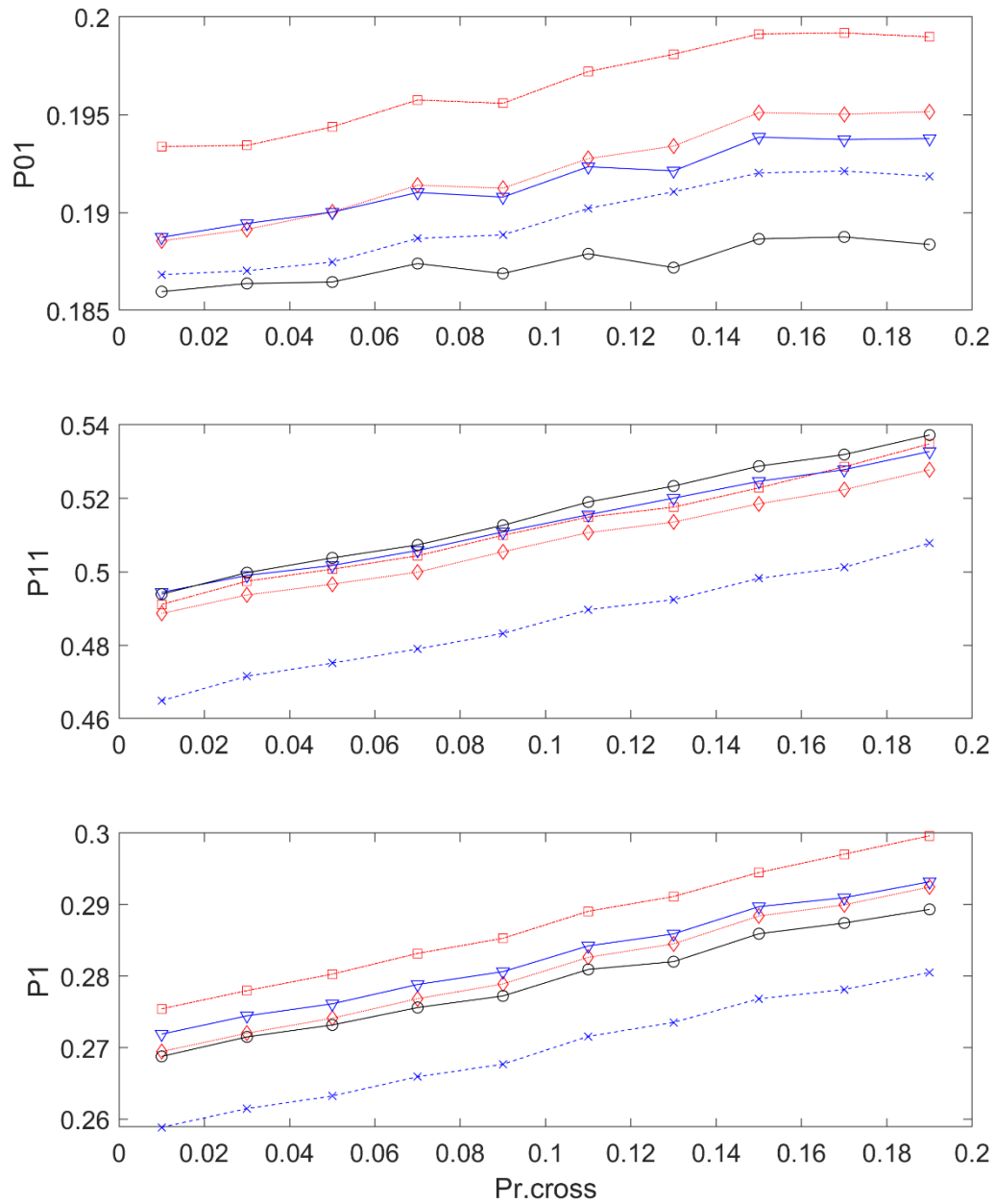
762

763

764

765

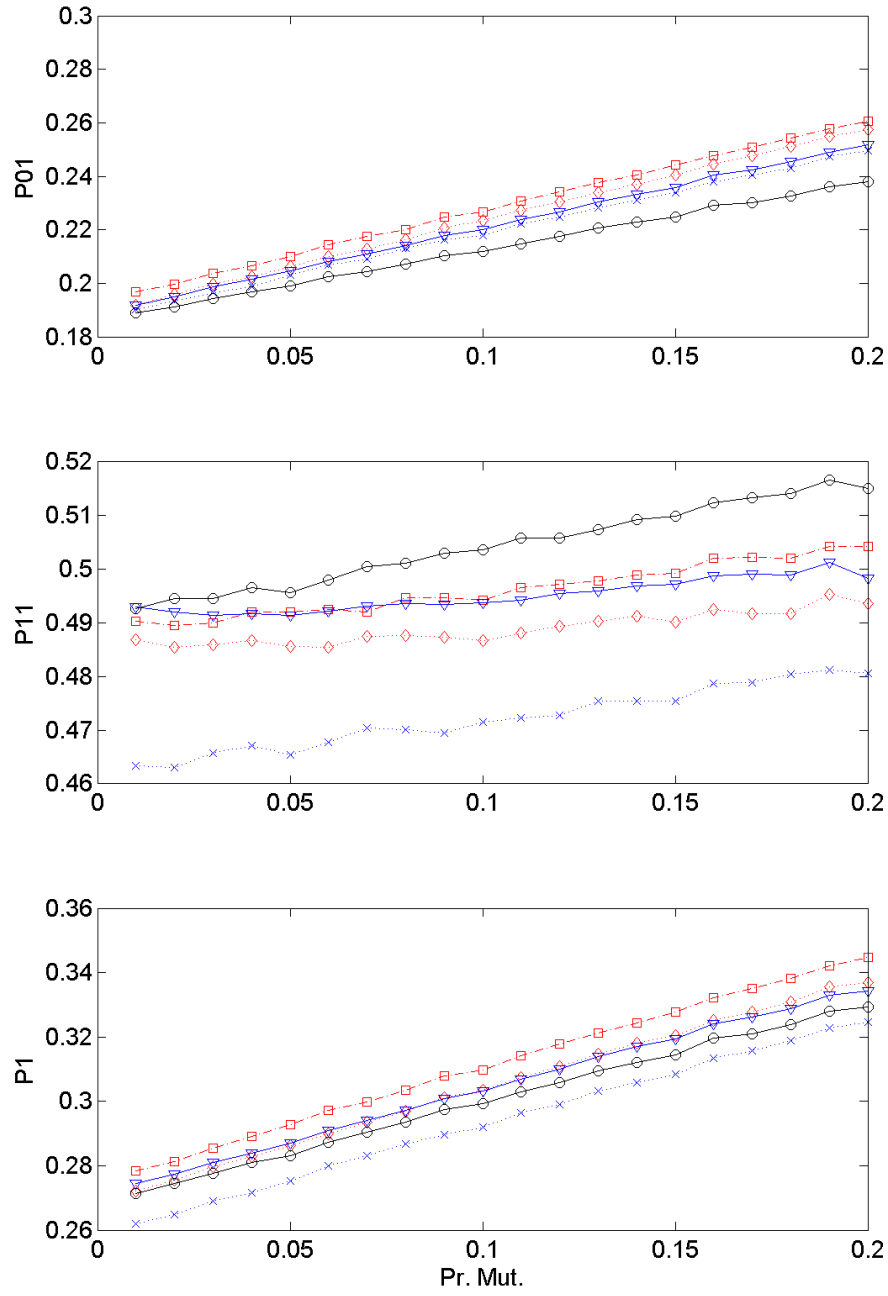
766

767

768 Figure 11. Transition probabilities and marginal distribution for the selected five stations along

769 with changing the cross-over probability P_{cr} with the condition that the candidate value is one

770 and the previous value is also one. See Eq.(15) for the detail.

771

772

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

776
777

778

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

782

783 Table A 2. Example dataset for estimating distances. The second row presents the current daily
784 precipitation occurrences for 12 stations and the rows below show the absolute difference
785 between the current occurrences (**Xc**) and the observed data in Table A 1. The last column
786 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**

787

788

789

Table A 3. Example for selecting one sequence for X_{c+1}. The second row presents the distances in Table A 2. The third and fourth columns show the sorted days and distances for the smallest distances to the largest in the second column. The fourth row presents the probabilities estimated with Eq. (12). Note that there are six days whose distances are the same with each other. In this 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	

795

796

797   Table A 4. Example for GA mixture for $\mathbf{X_{c+1}}$. The second and third rows present two selected
798   sets, while the third row shows the final set for $\mathbf{X_{c+1}}$ with the crossover at S6 and S8 and the
799   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**

800
801

802