Response to Reviews of the paper “Discrete k-nearest neighbor resampling for simulating multisite precipitation occurrence and adaption to climate change”

(Manuscript # gmd-2018-181-RC1,)

Interactive comment on “Discrete k-nearest neighbor resampling for simulating multisite precipitation occurrence and adaption to climate change” by Taesam Lee and Vijay P. Singh

Anonymous Referee #2 Received and published: 31 December 2018

Reply: The authors appreciate this reviewer’s comments. The authors have improved the quality of the current study according to the comments of the reviewer. Hope this reviewer is satisfied with this modification.

Present study attempts to develop a novel simulation method for multi-site precipitation occurrence, combining the k-nearest neighbor sampling technique and genetic algorithm. The coupled model has been applied in precipitation occurrence simulation in single sites. The (only) novelty probably lies in the application of this coupled technique in generating the multi-site precipitation occurrence. Authors may clarify these and may specify whether the novelty lies in the method deployed or in the application (See line 35 in the abstract and further such claims in the manuscript body).

Reply: The authors appreciate this reviewer’s insightful comment. The novelty of the current study is to propose the discrete version of KNNR-GA model in simulating multisite occurrence. The KNNR-GA model has been developed for multisite simulation of streamflow for continuous variables. The novelty of the current study is how to handle the multisite discrete binary process which is the main difference between the continuous version and the discrete version of the current study. The authors have improved the abstract and manuscript to emphasize this point. Hope this modification is satisfactory.

While, stochastic weather models (like the one deployed in this study) are commonly deployed in various applications, it would be preferable to give some physical justification to the application and comprehend the results obtained. This would bring more confidence into the purely statistical methods which otherwise may not have captured any physical relationships/behavior of the system been dealt. This is particularly significant in the present study, since multi-site occurrences might be directed by many climatic feedbacks and also controlled by many local factors also. Absence of any such physical explanation may leave the methods sound robotic and put doubt s in its generic applicability.

Reply: The authors have tried to provide the physical connection to the current results. For example, the following statement for the GA mixing process has been connected with the physical process of the proposed model.

"This can be problematic for the simulation purpose in that one of the major simulation purposes is to simulate sequences that might possibly happen in future. The wet (1) or dry (0)

for multisite precipitation occurrence is decided by the spatial distribution of a precipitation weather system. A humid air mass can be distributed randomly relying on wind velocity and direction as well as surrounding air pressure. In general, any combinations of wet and dry stations can be possible, especially when the simulation continues infinitely. Therefore, the patterns of simulated data must be allowed to have any possible combinations, here 4096 even if it has not been observed from the historical records. Also, its probability to have this new pattern must not be high since it has not been observed in the historical records and this can be taken into account by low probability of the crossover and mutation. "

"Daily precipitation occurrence, in general, shows the strongest serial correlation at lag-1 and its correlation decays as the lag gets longer. This is because a precipitation weather system moves according to the surrounding pressure and wind direction that dynamically change within a day or week. Therefore, we analyzed the lag-1 cross-correlation in the current study as the representative lagged correlation structure."

"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 measurement like a precipitation weather system dynamically changes but often it impacts the system of the following day."

In addition, the present method is compared with a method (MONR) which is developed almost two decades back. Is MONR a frequently used method for multi-site precipitation occurrence simulation? It would be convincing to compare the present technique with more recent methods deployed for multi-site precipitation occurrence simulation. More specific comments are provided below for the kind consideration of the authors.

Reply: The authors appreciate the reviewer's insightful comment. Even if MNOR model is rather old-fashioned, this model has been popularly employed in this field and its performance is more comparable to the Markov Chain model especially in multisite occurrence cases of precipitation dataset.

1. Line 68 – 74: Wilks (1998) model assumes standard normal variate and underestimates the lagged cross correlation. As mentioned before, is it really worth to compare the present method to this model, which works on an entirely different hypothesis? As mentioned by the authors in the next paragraph (lines 75-81), KNNR and KNNR-GA are proved to be efficient. Won't it be better to compare the present model (DKNNR) to compare with the above model, to highlight its applicability in multi-site precipitation occurrence, given that the novelty of the study is claimed to be in this application.

Reply: The authors appreciate the reviewer's insightful comment. The MONR model is the model of Wilks (1998) and it has been popularly employed in the literature. The present study compared the discrete version of KNNR-GA with the model of Wilks (1998), named as MONR here. See the first line of the section 2.2 as the following:

"Wilks (1998) suggested a multisite occurrence model using a standard normal random number (here, denoted as MONR) that is spatially dependent but serially independent."

2. Line 78-81: It is mentioned that KNNR model cannot produce different patterns and coupling with GA solves this drawback. Please provide more details on how GA could possibly solve this. And how the application of GA could ensure generation of similar populations. It would be interesting if some physical sense can also be provided here – how possibly GA could simulate those system behavior?

Reply: The authors appreciate the reviewer's detailed comment. Further explanation is added in the manuscript to improve the clarity in the result section.

"We further tested and discuss why the GA mixing is necessary in the proposed DKNNR model as follows. For example, assume that three weather stations are considered and observed data only has the occurrence cases of 000, 001,011,010, 011,100,111 among $2^3=8$ possible cases. In other words, no patterns for 110 and 101 is found in the observed data. Note that 0 is dry day and 1 is rainy (or wet) day. The KNNR is a resampling process in that the simulation data is resampled from the observation. Therefore, no new patterns such as 110 and 101 can be found in the simulated data.
This can be problematic for the simulation purpose in that one of the major simulation purposes is to simulate sequences that might possibly happen in future. The wet (1) or dry (0) for multisite precipitation occurrence is decided by the spatial distribution of a precipitation weather system. A humid air mass can be distributed randomly relying on wind velocity and direction as well as surrounding air pressure. In general, any combinations of wet and dry stations can be possible, especially when the simulation continues infinitely. Therefore, the patterns of simulated data must be allowed to have any possible combinations, here 4096 even if it has not been observed from the historical records. Also, its probability to have this new pattern must not be high since it has not been observed in the historical records and this can be taken into account by low probability of the crossover and mutation.
This drawback of the KNNR model frequently happens in multisite occurrence as the number of stations increases. Note that the number of patterns increases as 2^n where n is the number of stations. If n=12, then 4096 cases must be observed. However, among 4096 cases, observed cases are limited, since the number of data is limited. The GA process can mix two candidate patterns to produce new patterns. For example, in the three station case, a new pattern 101 can be produced from two observed occurrence candidates of 001 and 100 by the crossover of the first value of 001 to the first value of 100 (i.e. 001 →101), which is not in the observed data.
Note that the data employed in the case study are 40 years and 122 days (summer months) in each year. The total number of the observed data is 4880 and the number of possible cases is 4096. We checked how many of possible cases are not found in the observed data. The result shows that 3379 cases are not observed at all for the entire cases as shown in Figure 4.
We further investigated how many new patterns are generated with the probabilities $P_{cr}=0.02$, $P_m=0.001$ by the proposed GA mixing. The generated data for 100 sequences from DKNNR with the GA mixing shows that the number 3379 was reduced to 1200, which is not in the dataset among the 4096 possible patterns. Therefore, more than 2000 new patterns were simulated with the GA mixing process. The KNNR model without the GA mixing does not produce any new patterns in the 100 sequences with the same length of the historical data."

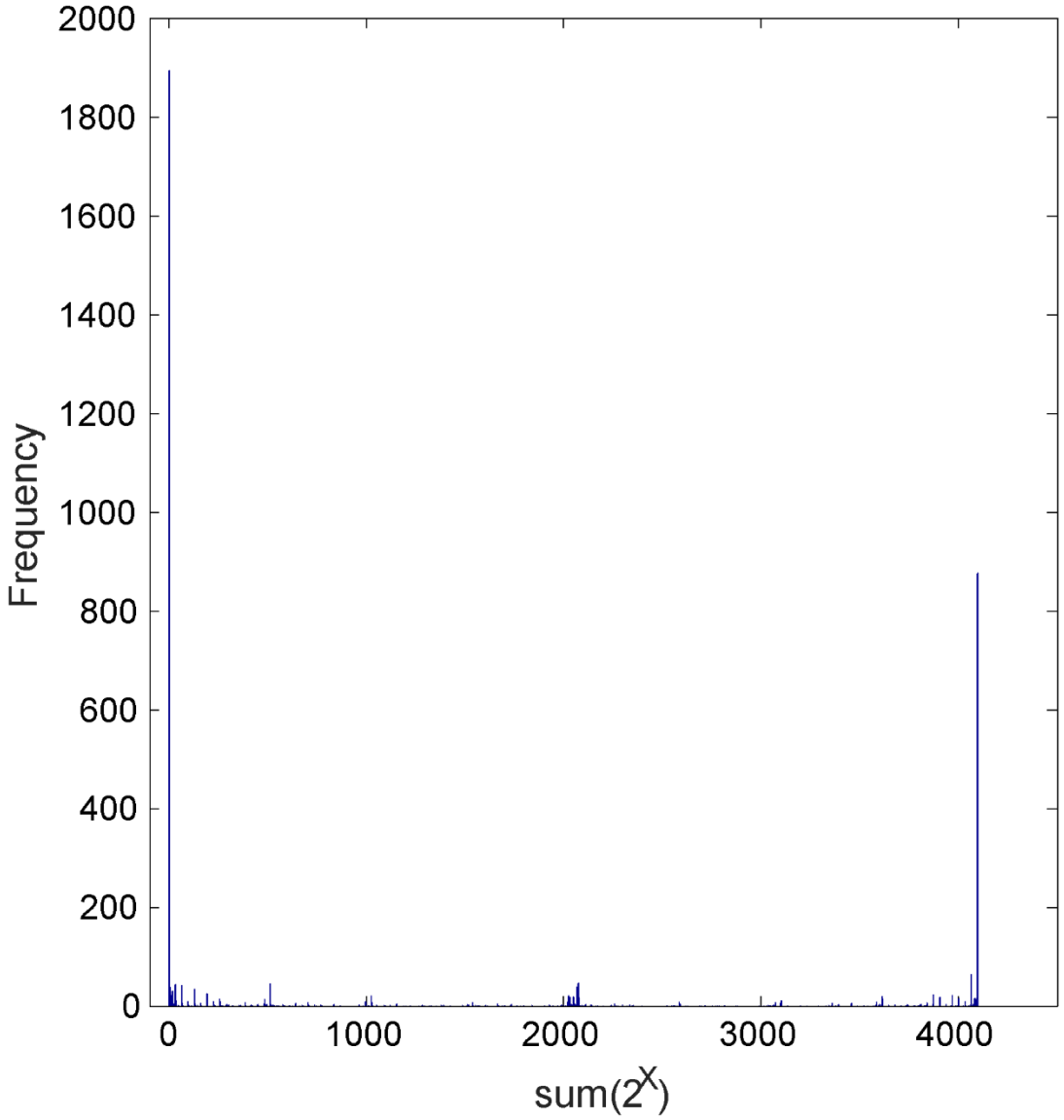

Figure S 1. Frequency of the observed patterns among all the possible cases (4096). The X coordinate indicates each pattern. All zero (0) and all one (4095) has the largest and second largest number of frequency (i.e. 1894 and 877, respectively) as expected meaning all dry and all wet stations. Note that the bars are very sporadic indicating a number of occurrence patterns are not observed.

3. Line 142: “multisite occurrence X and the observed multisite occurrence x”. Aren’t both these variables multi-dimensional and of same size? It would be ideal to denote both in capitals then.

Reply: The authors appreciate the reviewer’s detailed comment. We denote the observed occurrence with a lower case and the simulate variable with an upper case. For representing

a multisite variable, we use the bold character. This separation is inevitable to express the simulation procedure from the observed dataset (especially in KNNR model). In Eq.11, X_c^s and x_i^s represent only the simulation variable and observed data of the s^{th} station. Hope this is reasonable to this reviewer. To avoid confusion, we modify the sentence as follows:

"Estimate the distance between the current (i.e. time index: c) multisite occurrence X_c^s and the observed multisite occurrence x_i^s for the s^{th} station s=1,…,S. Here, the distance is measured for i=1,…, n-1 as

$$D_i = \sum_{s=1}^{S} \left| X_c^s - x_i^s \right| \qquad (1)$$

"

4. Line 158: When the algorithm will select the GA mixing? What is the criterion for GA mixing in the procedure?

Reply: The authors appreciate the reviewer's insightful comment. It is subjective. If one wants to simulate the dataset as the same observed pattern, this procedure can be skipped. Otherwise, the GA procedure gives the benefit of generating new patterns that we already discussed under comment 2. The sentence is modified accordingly.
"Execute the following steps for GA mixing if GA mixing is subjectively selected. Otherwise, skip this step."

5. Line 178-179: It is mentioned later in the manuscript that the changes in the mutation and cross-over probabilities may be carried out to adapt to the changes in the transition and marginal probability distributions (See lines 187-188). Considering that, would it be ideal to fix these as 0.01, following Lee et al (2010b). Shouldn't this be case specific? If not then, the later statement (lines 187-188) are questionable.

Reply: From the comment of the Reviewer 1, the estimation of parameter set was reinvestigated thoroughly. We concluded that the parameter set of P_{cr} and P_m as 0.02 and 0.003 showed the best from the result of RMSE estimated with the transition and limiting probabilities of the tested stations. The detailed results are as follows. Hope this investigation is satisfactory.

"The roles of crossover probability P_{cr} (Eq. (13)) and mutation probability P_m (Eq.(14)) were studied by Lee et al. (2010b). In the current study, we further tested by selecting an appropriate parameter set of these two parameters with the simulated data from the DKNNR model and the record length of 100,000. RMSE (Eq. (18)) of the three transition and limiting probabilities (P_{11}, P_{01}, and P_1) between the simulated data and the observed was used, since those probabilities are key statistics that the simulated data must match with the observed data and no parameterization of these probabilities was made for the current DKNNR model. Results are shown in Figure 2 and Figure 3 for P_{cr} and P_m, respectively. For P_{cr} in Figure 2, the probability of 0.02 shows the smallest RMSE in all transition and limiting probabilities. The RMSE of P_m in Figure 3 shows a slight fluctuation along with P_m. However, all three probabilities (P_{11}, P_{01}, and P_1) have relatively small RMSEs in P_m =0.003. Therefore, the parameter set 0.02 and 0.003 was chosen for P_{cr} and P_m, respectively, and employed in the current study."

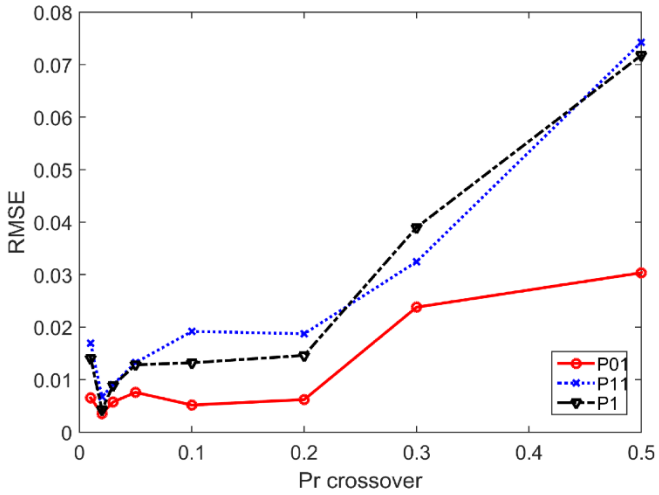

Figure 2. Testing for different probabilities of crossover Pcr. RMSE is estimated for all the tested 12 stations for each transition probability.

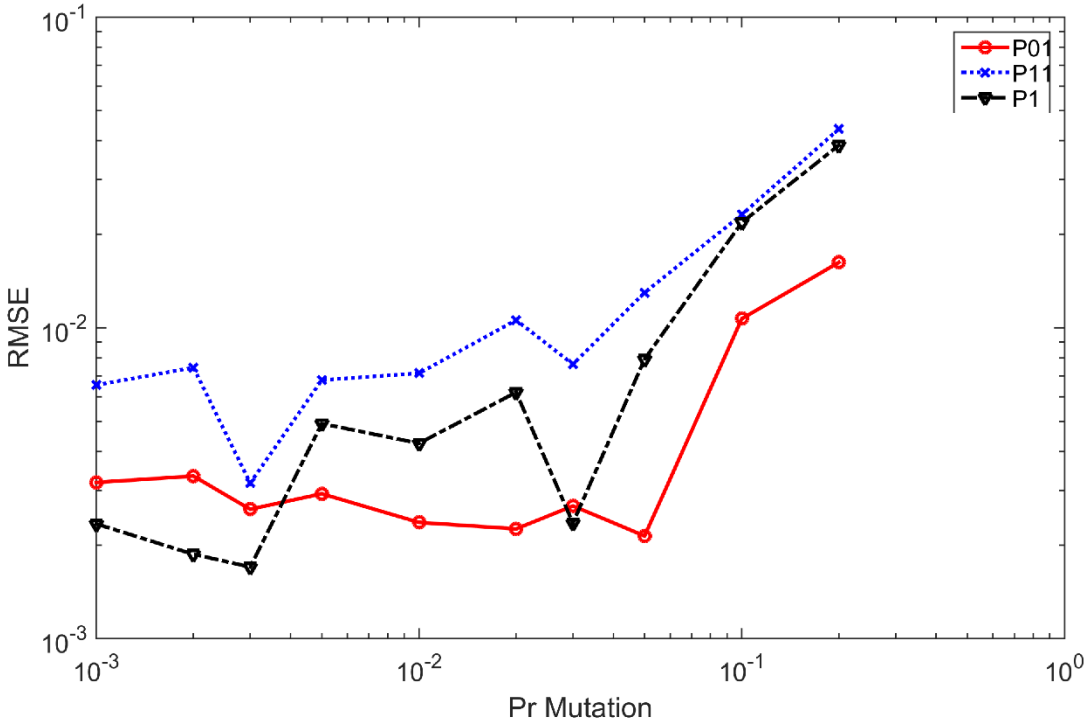

Figure 3. Testing for different probabilities of mutation Pm. RMSE is estimated for all the tested 12 stations for each transition probability.

6. Section 3.2: Authors must be pointing towards “Dealing with Non-stationarity” than “Adaptation to climate change”. It is clear that only changes in marginal and transition probabilities are been considered, by tuning the crossover and mutation probabilities? “Climate change” may refer to a larger phenomenon, which might not be addressed directly in the present study. Please explain.

Reply: The authors totally agree with the concern of the reviewer. Tuning the crossover and mutation probabilities only affected the marginal and transition probabilities. This limitation must be addressed as this reviewer commented. We added the following to address the

concern from this reviewer at the end of section 6. The authors hope that this statement is satisfactory.
"Climate change, however, may refer to a larger phenomenon, which cannot be addressed directly through modifying only the marginal and transition probabilities as in the current study. Further modeling development on systematically varying temporal and spatial cross-correlations is required to properly address the climate change of the regional precipitation system."

7. How tuning of crossover and mutation probabilities could handle the non-stationarity in the time series of multiple stations? Can the model change these parameters in between the time frame of the simulation, so as to incorporate the parameter change(s) in the probability distributions?

Reply: The authors totally agree with the concern of the reviewer as with the previous comment that tuning the crossover and mutation probabilities only effected the marginal and transition probabilities. The authors consider that it is possible that the model can change the parameter to adapt to the climate change between the time frame of the simulation to incorporate the parameter change automatically. But this capability has not been fully investigated. In addition, the focus of the current study is to propose a novel approach that simulates multisite occurrence process through the nonparametric approaches. Further development for adopting to climate change and its application is presented as a possible improvement of the proposed model in the near future and will be presented as a separate work as explained in the conclusion section as the following.

"We tested further the enhancement of the proposed model for adapting to climate change by modifying the mutation and crossover probabilities P_m and P_{cr}. The results showed that the proposed DKNNR model has the capability to adapt to the climate change scenarios, but further elaborate work is required to find the best probability estimation for climate change. Also, only the marginal and transition probabilities cannot address the climate change of regional precipitation. The variation of temporal and spatial cross-correlation structure must be considered to properly address the climate change of the regional precipitation system. Further study on improving the model adaptability to climate change will be followed in the near future.Also, the simulated multisite occurrence can be coupled with a multisite amount model to produce precipitation events, including zero values. Further development can be made for multisite amount models with a nonparametric technique, such as KNNR and bootstrapping."

8. Section 4: Please provide more details about the precipitation data used, its seasonality, rainy day characteristics etc. Are the stations selected meteorologically homogenous?

Reply: The authors appreciate the reviewer's detailed comment. The following is added to address this comment. Hope this statement is satisfactory.

"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)."

"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."

9. Section 5: This may go into the results section, if it sounds fine.

Reply: The authors appreciate the reviewer's comment. The authors separate this section to explain how the developed model is applied to the datasets and what measurements were used to show its performance. The authors consider that the separation of this application part is reasonable because there are no specific results in this section. The results of the GA mixing and its probability section in the result section are also added for the comments of the reviewer.

10. Line 222: " . . ..., since a synoptic scale weather system could result in lagged cross-correlation" – Can this statement be generalized for all locations?

Reply: The authors appreciate the reviewer's specific comment and understand his concern. The statement might not be always true. Therefore, the sentence was modified accordingly as follows:

"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)."

11. Figure 2-4: Ensemble means from MONR are close to the observed mean, than those of DKNNR model. Is MONR better in that sense? Please clarify.

Reply: The authors agree with the reviewer's comment and it is already mentioned in the manuscript as the following (see the L250-251). We also modified the sentence to include the same implication to P01 and P1 as well as P11.
"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 5 and Figure 6."*

12. Line 254-255: "Even though the transition probabilities were not employed in simulating rainfall occurrence, the DKNNR model preserved this statistic fairly well" – Is it merely by chance? Please provide justification to build confidence. Do you expect the results to vary, when deployed in different regions?

Reply: The authors appreciate the reviewer's crucial comment. The KNN resampling with the distance in Eq. (11) between the current simulation multisite occurrence (X_c^s) and the historical multisite occurrence states (x_i^s) allows to preserve the transition probabilities. The following statement is added accordingly.

"In the DKNNR modeling procedure, the simple distance measurement in Eq. (1) 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 measurement like a precipitation weather system dynamically changes but often it impacts the system of the following day."

13. Line 273-274: "Precipitation is not significantly correlated with more than one day" – Please provide reference. The statement may not hold well globally, as Box-Jenkins models of higher order are often applied for simulating precipitation events.

Reply: The authors totally agree with the reviewer's comment. The sentence was modified accordingly. Hope this modification is satisfactory.

"Daily precipitation occurrence, in general, shows the strongest serial correlation at lag-1 and its correlation decays as the lag gets longer. This is because a precipitation weather system moves according to the surrounding pressure and wind direction that dynamically change within a day or week. Therefore, we analyzed the lag-1 cross-correlation in the current study as the representative lagged correlation structure."

14. It would be better to number the stations considering its proximity. It will help in analyzing the results.

Reply: The authors appreciate the reviewer's comment. The author tried to change the numbers but consider that this may not be meaningful much since the order from west to east or north to south can be different with its numbering. Readers might be confused from this numbering. For example, the current 8,7,6, 10,2,9,1 stations can be changed to 1,2,3,4,5,6,7. The stations 3 and 4 seem close to each other due to renumbering, which is not correct. We also tested with 1,2,3,7,6,5,4. However, 1 and 7 must be far away from each other according to its numbering but they are very close to each other. We tried different numbering to consider the proximity but did not find any logical ordering. Therefore, we prefer staying as it is. Hope this can be understandable to the reviewer.

15. It would be interesting to see the results generated by the simple KNNR model in this application. Also, it would be helpful, if you may please explain how the incorporation of GA possibly helped in modeling the physical laws of the precipitation system.

Reply: The authors appreciate the reviewer's insightful comment. We produced the results without the GA process as presented in the following (See Figure S2-Figure S6). The presented results show that no significant difference from the one with the GA mixing can be found. The following is discussed in the manuscript right before the results of the probability selection (section 6.1).

"We also tested the simulation without the GA mixing procedure (results not shown). The results showed that no better result could be found from the simulation without GA mixing. The necessity of the GA mixing is further discussed in the following."

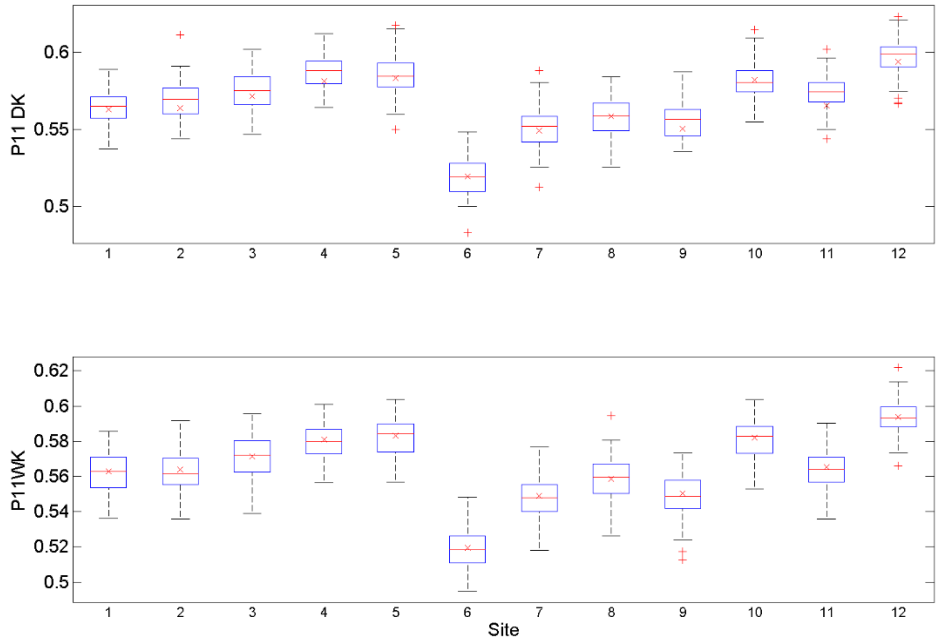

Figure S 2. Boxplots of the P11 probability for the data simulated from the DKNNR model without the GA mixing (top panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12 selected weather stations from the Yeongnam province.

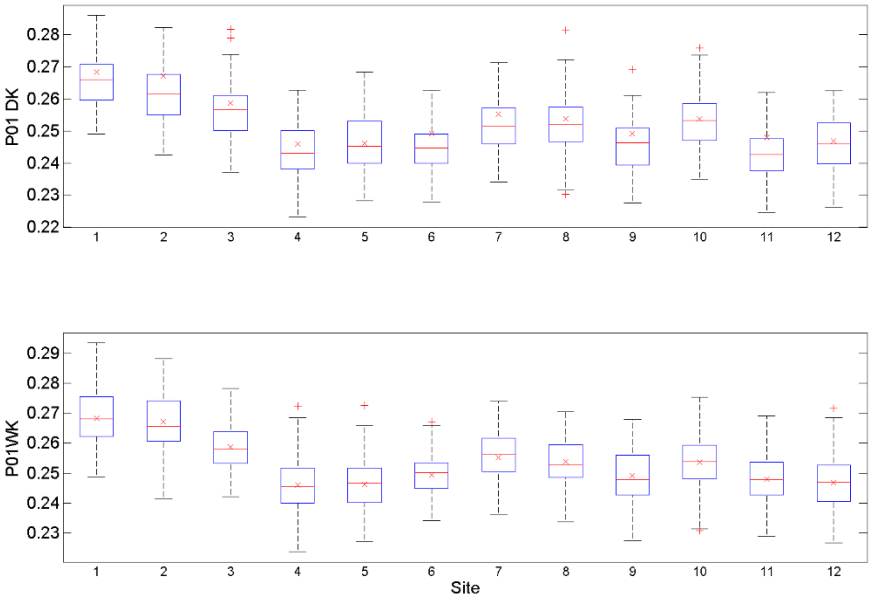

Figure S 3. Boxplots of the P01 probability for the data simulated from the DKNNR model without the GA mixing (top panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12 selected weather stations from the Yeongnam province.

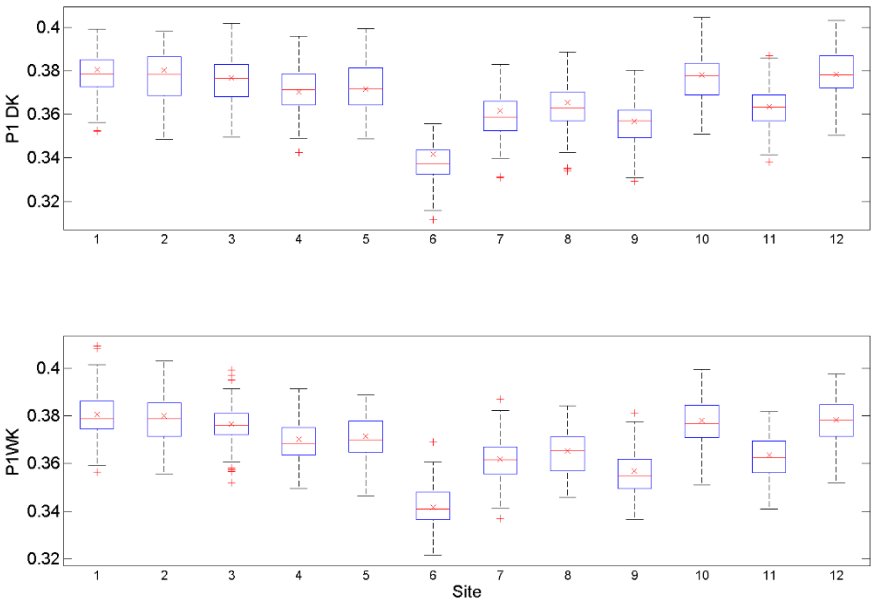

Figure S 4. Boxplots of the P1 probability for the data simulated from the DKNNR model without the GA mixing (top panel) and the MONR model (bottom panel) as well as the observed (x marker) for the 12 selected weather stations from the Yeongnam province.

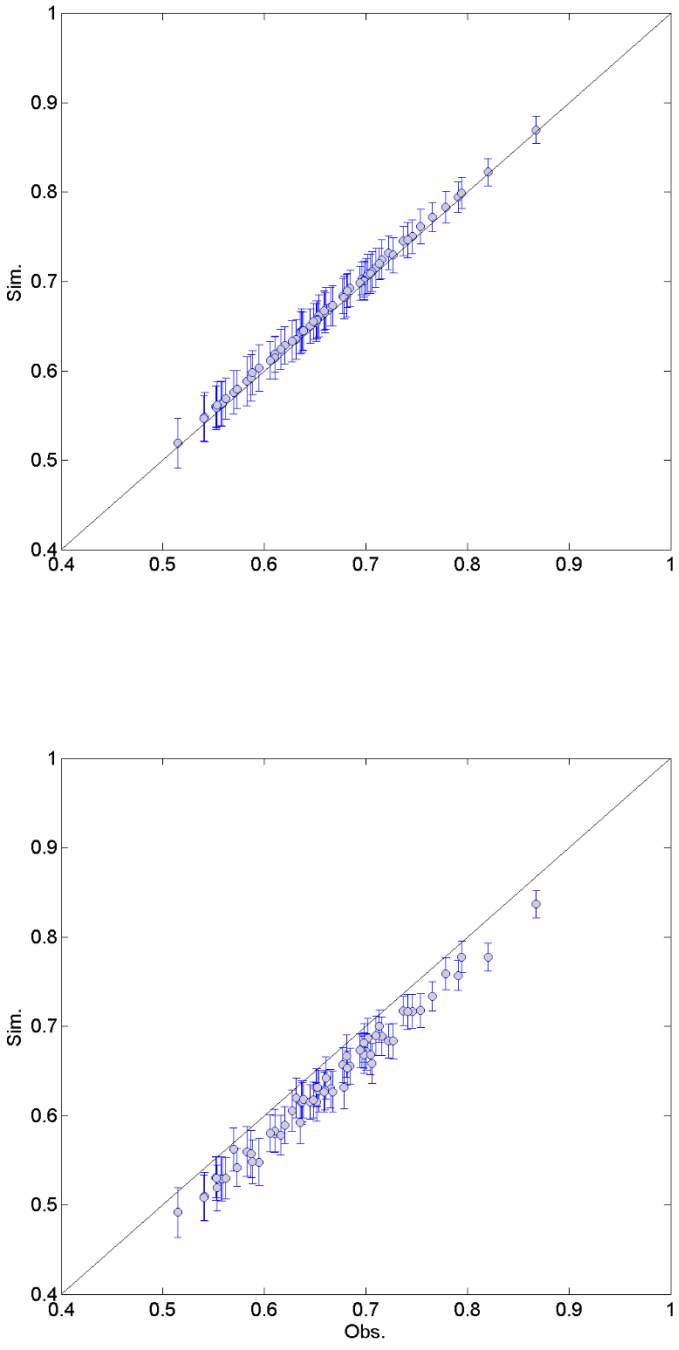

Figure S 5. Scatterplot of cross-correlations between 12 weather stations for the observed data (X coordinate) and the generated data (Y coordinate) generated from the DKNNR model without the GA mixing (top panel) and the MONR model (bottom panel). The cross-correlations from 100 generated series are averaged for the filled circle and the errorbars upper and lower extended lines indicate the range of 1.95×standard deviation.

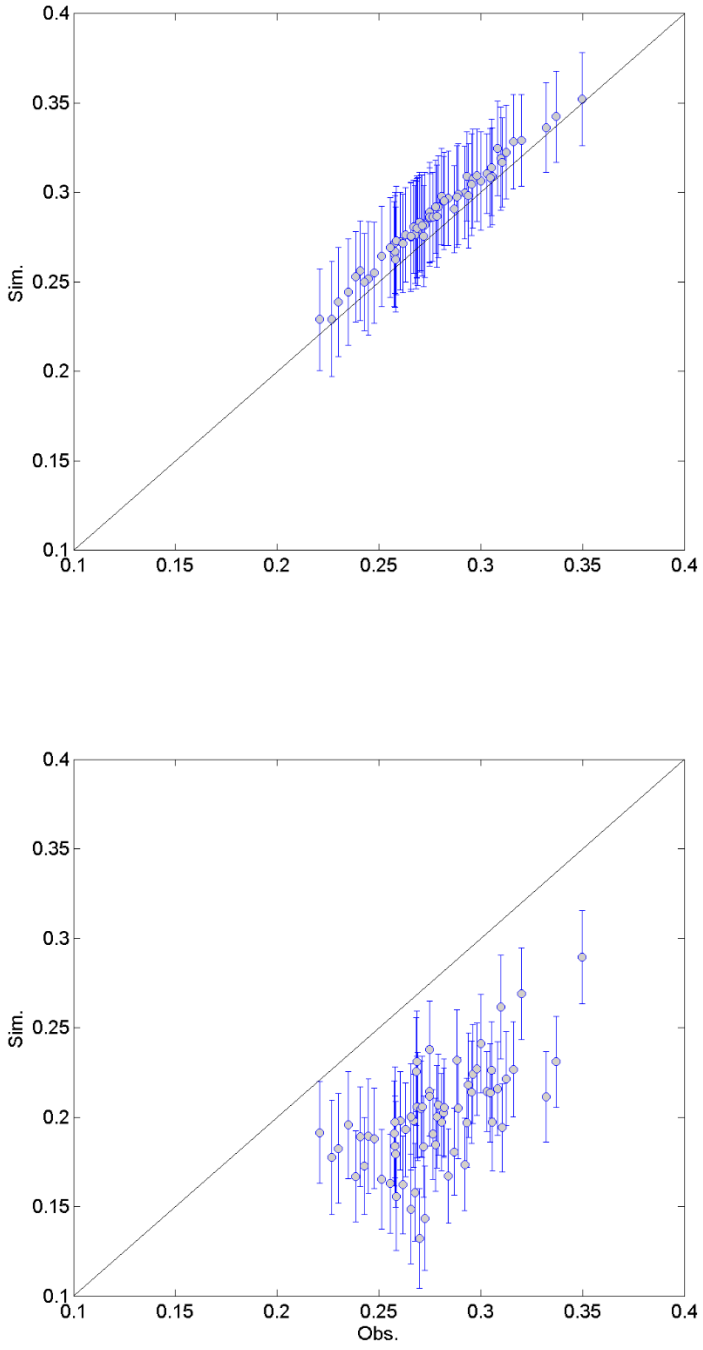

Figure S 6. Scatterplot of lag-1 cross-correlations between 12 weather stations for the observed data (X coordinate) and the generated data (Y coordinate) generated from the DKNNR model without the GA mixing (top panel) and the MONR model (bottom panel). The cross-correlations from 100 generated series are averaged for the filled circle and the errorbars upper and lower extended lines indicate the range of 1.95×standard deviation

16. Disadvantage of the simple KNNR model is the inability to simulate different patterns from the observed series. Do the stations selected exhibit significant nonstationarity? If not, will the KNNR model also serve the purpose?

Reply: The authors appreciate the reviewer's comment. The GA mixing was not applied for nonstationarity. The GA mixing is applied to overcome the disadvantage of the KNNR model that only observed pattern is repeated in the simulated data. This case is not sound for the simulation study purpose. As mentioned under comment 2, more than half of the possible patterns are not observed in the historical data. This has been covered multiple times already under previous comments. Hope this explanation can be acceptable to the reviewer.

17. Section 6.3: I am a little confused here. How can the parameters be changed in the future, for the model to adapt to the future changes, given that we may not clear information about these changes?

Reply: The authors appreciate the reviewer's comment. The authors did not fully investigate the specific changes required to be made for specific climate change assessment at this stage. As mentioned under comment 7, the focus of the current study is to propose a novel approach that simulates multisite occurrence process through nonparametric approaches. Further development for adopting to climate change and its application are partially presented as a possible improvement of the proposed model in the near future and will be presented as a separate work as explained in the conclusion. This limitation and possible development are discussed in the last section.