Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-181-AC1, 2018

Geoscientific Model Development Discussions

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

Taesam Lee and Vijay P. Singh

tae3lee@gnu.ac.kr

Received and published: 12 December 2018

Author response to the 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

1. The manuscript presents discrete k-nearest neighbor resampling for simulating multisite precipitation occurrence and adaption to climate change, which is interesting. The subject addressed is within the scope of the journal. Reply: The authors appreciate

this reviewer's comment.

2. However, the manuscript, in its present form, contains several weaknesses. Appropriate revisions to the following points should be undertaken in order to justify recommendation for publication. Reply: The authors appreciate the reviewer's comments. the authors improved the quality of the current study according to the given comments. Hope this reviewer is satisfactory to this modification.

3. For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction. Reply: The authors appreciate the reviewer's comment. Accordingly, the introduction and abstract were improved as follows. Hope the modification is satisfactory.

Abstract: Stochastic weather simulation models are commonly employed in water resources management agricultural applications, forest management, transportation management, and recreational activities. The data simulated by these models, such as precipitation, temperature, and wind, are used as input for hydrological and agricultural models. Stochastic simulation of multisite precipitation occurrence is a challenge because of its intermittent characteristics as well as spatial and temporal cross-correlation. The multisite occurrence model with standard normal variate (MONR) has been used for preserving key statistics and contemporaneous correlation, but it cannot reproduce lagged crosscorrelation between stations and long stochastic simulation is therefore required to estimate its parameters. Employing a nonparametric technique, k-nearest neighbor resampling (KNNR), and coupling it with Genetic Algorithm (GA), this study proposes a novel simulation method for multisite precipitation occurrence, overcoming the shortcomings of the existing MONR model. The proposed discrete version of KNNR (DKNNR) model is compared with an existing parametric model, called multisite occurrence model with standard normal variate (MONR). The datasets simulated from both the DKNNR model and the MONR model are tested using a number of statistics, such as occurrence and transition probabilities as well as temporal and spa-

tial cross-correlations. Results show that the proposed DKNNR model can be a good alternative for simulating multisite precipitation occurrence, while preserving the lagged crosscorrelation between sites and simulating multisite occurrence from a simple and direct procedure without no parameterization. We also tested the model capability to adapt climate change. It is shown that the model is capable but further improvement is required to have specific variations of the occurrence probability due to climate change. Combining with the generated occurrence, the multisite precipitation amount can then be simulated by any multisite amount model. .

Introduction: Wilks (1998) presented a multisite simulation model for the occurrence process (i.e. X) using the standard normal variable that is spatially dependent, representing the relation between the occurrence variable and the standard normal variable with simulation data. Originally, the occurrence of precipitation had been simulated with discrete Markov Chain (MC) model (Katz, 1977). Compared to the MC model requiring a significant number of parameters to generate multisite occurrence, the multisite occurrence model proposed by Wilks (1998) transforms the standard normal variate and simulates the sequence with multivariate normal distribution, and then back-transforms the multivariate normal sequence to the original domain. The model is able to reproduce the contemporaneous multisite dependence structure and lagged dependence only for the same site while requiring a complex simulation process to estimate parameter for each site and being unable to preserve lagged dependence between sites. Meanwhile, Lee et al. (2010a) proposed a nonparametric-based stochastic simulation model for hydrometeorological variables. They overcame the shortcoming of a previous nonparametric simulation model (Lall and Sharma, 1996), called k-nearest neighbor resampling (KNNR) such that the simulated data cannot produce patterns different from those of the observed data (Brandsma and Buishand, 1998; Mehrotra et al., 2006; St-Hilaire et al., 2012). In addition to this KNNR, Lee et al. (2010a) used a meta-heuristic algorithm Genetic Algorithm (GA) that led to the reproduction of similar populations by mixing the simulated dataset. While the KNNR is employed to find similar historical analogues of multisite occurrence to the current status of a simulation

series, GA is applied to use its skill to generate a new descendant from the historical parent chosen with the KNNR. In this procedure, the multisite occurrence of the precipitation variable can be simulated while preserving spatial and temporal correlations. Note that meta-heuristic techniques to GA have been popularly employed in a number of hydrometeorological applications (Chau, 2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013). A number of variants of KNNR-GA have since been applied (Lee et al., 2012; Lee and Park, 2017). None of these models can adopt the multisite occurrence in precipitation whose characteristics are binary and temporally and spatially related. Therefore, in the current study we propose a novel stochastic simulation method for multisite occurrence of the precipitation variable with the KNNR-GA based nonparametric approach that (1) simulates multisite occurrence with a simple and direct procedure without parameterization of all the required occurrence probabilities; and (2) reproduces the complex temporal and spatial correlation between stations as well as the basic occurrence probabilities. Note that the proposed nonparametric model is compared with the most popularly employed model proposed by Wilks (1998). Even though the multisite occurrence data from this model (Wilks, 1998) preserves various statistical characteristics of the observed data well, significant underestimation of lagged cross-correlation still exists. Furthermore, the relation between standard normal variable and occurrence variable relies on long stochastic simulation. The paper is organized as follows. The next section presents a mathematical background of existing multisite occurrence modeling. The modeling procedure is discussed in section 3. The study area and data are reported in section 4. The model is applied in section 5. Results of the proposed model are discussed in section 6, and summary and conclusions are presented in section 7."

4. It is shown in the reference list that the authors have several publications in this field. This raises some concerns regarding the potential overlap with their previous works. The authors should explicitly state the novel contribution of this work, the similarities and the differences of this work with their previous publications. Reply: The authors appreciate th reviewer's thoughtful comment. We have explicitly described the detailed

difference and stated the novel contribution of this study as follows

"Meanwhile, Lee et al. (2010a) proposed a nonparametric-based stochastic simulation model for hydrometeorological variables. They overcame the shortcoming of a previous nonparametric simulation model (Lall and Sharma, 1996), called k-nearest neighbor resampling (KNNR) such that the simulated data cannot produce patterns different from those of the observed data (Brandsma and Buishand, 1998; Mehrotra et al., 2006; St-Hilaire et al., 2012). In addition to this KNNR, Lee et al. (2010a) used a meta-heuristic algorithm Genetic Algorithm (GA) that led to the reproduction of similar populations by mixing the simulated dataset. While the KNNR is employed to find similar historical analogues of multisite occurrence to the current status of a simulation series, GA is applied to use its skill to generate a new descendant from the historical parent chosen with the KNNR. In this procedure, the multisite occurrence of the precipitation variable can be simulated while preserving spatial and temporal correlations. Note that meta-heuristic techniques to GA have been popularly employed in a number of hydrometeorological applications (Chau, 2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013). A number of variants of KNNR-GA have since been applied (Lee et al., 2012; Lee and Park, 2017). None of these models can adopt the multisite occurrence in precipitation whose characteristics are binary and temporally and spatially related."

5. It is mentioned in p.2 that k-nearest neighbor resampling coupling with genetic algorithm is adopted to simulate multisite precipitation occurrence. What are other feasible alternatives? What are the advantages of adopting this particular soft computing technique over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The authors appreciate the reviewer's comment. While the KNNR is employed to find similar historical analogues of multisite occurrence to the current status of a simulation series, GA is applied to use its skill to generate a new descendant from the historical parent chosen with the KNNR. In this procedure, the multisite occurrence of the precipitation variable can be simulated with preserving

spatial and temporal correlations. We added the following in the manuscript accordingly. Note that the location of the page has been changed especially in the introduction section for replying the comment 3 of this reviewer.

"While the KNNR is employed to find similar historical analogues of multisite occurrence to the current status of a simulation series, GA is applied to use its skill to generate a new descendant from the historical parent chosen with the KNNR. In this procedure, the multisite occurrence of the precipitation variable can be simulated while preserving spatial and temporal correlations. Note that meta-heuristic techniques to GA have been popularly employed in a number of hydrometeorological applications (Chau, 2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013). A number of variants of KNNR-GA have since been applied (Lee et al., 2012; Lee and Park, 2017). None of these models can adopt the multisite occurrence in precipitation whose characteristics are binary and temporally and spatially related."

6. It is mentioned in p.2 that multisite occurrence model with standard normal variate is adopted as benchmark for comparison. What are the other feasible alternatives? What are the advantages of adopting this particular model over others in this case? How will this affect the results? More details should be furnished. Reply: The authors thanks to this reviewer's comment. Another alternative is to use a multisite version of Markov Chain (M-MC) model by estimating the transition matrix of multisite occurrence. However, this M-MC model requires a number of parameters even difficult to handle and very often no data exist to estimate some of parameters. If this model is applied for comparison, the proposed model shows much better performance than the MONR model. The following is added in the manuscript accordingly.

"Wilks (1998) presented a multisite simulation model for the occurrence process (i.e. X) using the standard normal variable that is spatially dependent, representing the relation between the occurrence variable and the standard normal variable with simulation data. Originally, the occurrence of precipitation had been simulated with discrete Markov Chain (MC) model (Katz, 1977). Compared to the MC model requiring a signif-

icant number of parameters to generate multisite occurrence, the multisite occurrence model proposed by Wilks (1998) transforms the standard normal variate and simulates the sequence with multivariate normal distribution, and then back-transforms the multivariate normal sequence to the original domain. The model is able to reproduce the contemporaneous multisite dependence structure and lagged dependence only for the same site while requiring a complex simulation process to estimate parameter for each site and being unable to preserve lagged dependence between sites"

7. It is mentioned in p.8 that a random selection procedure is adopted to take into account the cases with the same quantity. What are other feasible alternatives? What are the advantages of adopting this particular procedure over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The authors appreciate this reviewer's comment. Other than the random selection, one can use always the first one. In such a case, only one historical combination of occurrence will be selected among the combinations with the same distance. The following is added in the manuscript. Hope this modification is satisfactory to this reviewer.

"For example, if S=2 and Xc1=0 and Xc2=1, the two sequences has the same D=1 as [xi1=0 and xi2=0] and [xi1=1 and xi2=1]. In this case, a random selection procedure is required to take into account the cases with the same quantity. One particular time index is randomly selected with the equal probabilities among the time indices of the same distances. Note that instead of the random selection, one can choose always the first one. In such a case, only one historical combination of multisite occurrences will be selected."

8. It is mentioned in p.9 that the reproduction procedure in (6-1) is adopted in this study. What are other feasible alternatives? What are the advantages of adopting this particular approach over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The authors appreciate this reviewer's comment. This reproduction process is a mating process by finding another individual that has similar characteristics with the current one xp+1. With this procedure, a similar

vector to the current vector will be mated and produce a new descendant. Alternatively, this procedure can be skipped. Then all the elements of the generated vector will be the same as the historical. The following is added accordingly in the manuscript.

"This reproduction process is a mating process by finding another individual that has similar characteristics to the current one xp+1. With this procedure, a similar vector to the current vector will be mated and produce a new descendant."

9. It is mentioned in p.9 that Eq.(13) is adopted for crossover. What are other feasible alternatives? What are the advantages of adopting this particular crossover type over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The same answer as in the comment 8 can be made to this comment for the feasible alternative. The advantage of this crossover is that a new occurrence vector whose elements are similar to the historical is generated. The following is added in the manuscript accordingly.

"From this crossover, a new occurrence vector whose elements are similar to the historical is generated."

10. It is mentioned in p.9 that Eq.(14) is adopted for mutation. What are other feasible alternatives? what are the advantages of adopting this particular mutation type over others in this case? How will this affect the results? The authors should provide more details on this. Reply: Another alternative can be skipped this procedure. Then always similar multisite occurrence to historical combinations would be generated, which is not feasible for a simulation purpose. The advantage of this mutation is to allow a totally new combination of multisite occurrence to be simulated with this mutation process compared to historical records. The following is added in the manuscript accordingly.

"This mutation procedure allows to generate a multisite occurrence combination that is totally different from the historical records. Without this procedure, always similar multisite occurrences to historical combinations are generated, which is not feasible for a simulation purpose."

11. It is mentioned in p.9 that a simple selection method is adopted for the selection of the number of nearest neighbors. What are other feasible alternatives? What are the advantages of adopting this particular method over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The authors appreciate this reviewer's critical comment. Another alternative is to use generalized cross-validation (GCV) as shown in Sharma and Lall1996 and Lee and Ouarda 2011 by treating this simulation as a prediction problem. However, the current multisite occurrence simulation does not necessarily require accurate value prediction and not much difference on simulation using the simple heuristic approach is reported. Also, this heuristic approach of k selection has been popularly employed for hydrometeorological stochastic simulations (Lall and Sharma, 1996; Lee and Ouarda, 2012; Lee et al., 2010b; Prairie et al., 2006; Rajagopalan and Lall, 1999). The following is added in the manuscript accordingly.

"One can use generalized cross-validation (GCV) as shown in Sharma and Lall1996 and Lee and Ouarda 2011 by treating this simulation as a prediction problem. However, the current multisite occurrence simulation does not necessarily require accurate value prediction and not much difference on simulation using the simple heuristic approach is reported. Also, this heuristic approach of k selection has been popularly employed for hydrometeorological stochastic simulations (Lall and Sharma, 1996; Lee and Ouarda, 2012; Lee et al., 2010b; Prairie et al., 2006; Rajagopalan and Lall, 1999)."

12. It is mentioned in p.11 that 12 weather stations were selected from Yeongnam province are adopted as the case study. What are other feasible alternatives? What are the ad- vantages of adopting this particular case study over others in this case? How will this affect the results? The authors should provide more details on this. Reply: The authors appreciate this reviewer's comment. The object of the current study is to build a simulation model for multisite precipitation occurrence. To validate the proposed model appropriately, tested sites must be highly correlated with each other as well as significant temporal relation. The employed stations inside the Gyeongnam area cover

one of the most important watersheds, the Nakdong River basin, where the Nakdong river pass through the entire basin and its hydrological assessments for agriculture and climate change has particular values in water resources management such as floods and droughts. The following has been added accordingly.

"To validate the proposed model appropriately, tested sites must be highly correlated with each other as well as significant temporal relation. The employed stations inside the Yeongnam area cover one of the most important watersheds, the Nakdong River basin, where the Nakdong river pass through the entire basin and its hydrological assessments for agriculture and climate change has particular values in water resources management such as floods and droughts."

13. It is mentioned in p.11 that historical records of 1976 to 2008 are taken. Why are more recent data not included in the study? Is there any difficulty in obtaining more recent data? Are there any changes to situation in recent years? What are its effects on the result? Reply: The authors appreciate this reviewer's comment. This dataset was employed to illustrate the performance of the proposed model especially for the base period. This dataset has been well evaluated from a number of the previous studies (Lee, 2017). According to this comment, more recent data up to the year2015 whose quality has been checked was added and all the results were modified accordingly. Not much significant difference was found from the results of the previous dataset.

14. It is mentioned in p.12 that the root mean square error is adopted to evaluate statistics from 100 generated series. What are the other feasible alternatives? What are the advantages of adopting this particular evaluation metric over others in this case? How will this affect the results? More details should be furnished. Reply: Another alternative would be MAE and Bias. The estimates showed that MAE has no difference from RMSE and Bias of the lag-1 correlation presents significant negative values implying the underestimation of the lag-1 correlation. The following is added in the manuscript including Table 9 and Table 10.

"We further tested the performance measurements of MAE and Bias. The estimates showed that MAE has no difference from RMSE. In addition, Bias of the lag-1 correlation presents significant negative values implying its underestimation for the simulated data of the MONR model shown in Table 9 while Table 10 of the DKNNR model shows much smaller bias."

15. It is mentioned in p.16 that ": : :Special remedy should be applied, such as decreasing cross-correlation by force, but further remedy was not applied in the current study since: : :" More justification should be furnished on this issue. Reply: The authors appreciate this reviewer's comment. We tried to discuss about the possible improvement of the existing MONR model not the proposed model in the current study. The improvement of the existing model is not within the scope of the current study. Following study can be doable for this issue. The authors consider that no further justification was necessary in the current study since the MONR model has not been proposed in the current study. Hope this reviewer understand this. We improved the sentence as the following to avoid the confusion of the model we discuss.

"Special remedy for the existing MONR model should be applied, such as decreasing cross-correlation by force, but further remedy was not applied in the current study since it was not within the current scope and focus."

16. It is mentioned in p.17 that ": : :However, the probability P01 fluctuated along with the increase of Pcr. Elaborate work to adjust all the probabilities is however required: : :" More justification should be furnished on this issue. Reply: The authors appreciate this reviewer's insightful comment. We agree that more justification and application might be needed to show the capability of the proposed model. However, the current study is focused on proposing a novel approach that simulates multisite occurrence process. Further development for adopting climate change and its application will be presented as a separate work as explained in the conclusion in the following. Hope this reviewer understand the intention of the authors.

"We tested further enhancement of the proposed model for adapting climate change through modifying the mutation and crossover probability Pm and Pcr with the current and previous states. The results show that the current model has the capability to adapt to the climate change scenarios, but elaborate work is required however. Further study on improving the model adaptability to climate change will be followed in near future."

17. Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results? Reply: The authors appreciate this reviewer's critical comment. The authors totally agree with this comment. Accordingly, we tested the key parameters for the proposed DKNNR method found that the parameter set of Pcr and Pm as 0.02 and 0.003 shows the best from the result of RMSE estimated with the transition and limiting probabilities of the tested stations. Hope this result is satisfactory to this reviewer. The following is added in the manuscript:

"The roles of crossover probability (Eq. 13) and mutation probability (Eq.14) were studied by Lee et al. [2010a]. In the current study, we further tested to select 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 transition and limiting probabilities (P11, P01, and P1) between the simulated data and the observed was used since those probabilities are key statistics that the simulated data must be met with the observed and no parameterization on these probabilities has been made for the current DKNNR model. The results are shown in Figure 2 and Figure 3 for Pcr and Pm, respectively. For Pcr in Figure 2, the probability of 0.02 shows the smallest RMSE in all transition and limiting probabilities. The RMSE of Pm in Figure 3 shows slight fluctuation along with Pm. However, all three probabilities have relatively small RMSEs in Pm =0.003. Therefore, the parameter set 0.02 and 0.003 is chosen for Pcr and Pm, 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.

âĂČ

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

18. Some assumptions are stated in various sections. Justifications should be provided on these assumptions. Evaluation on how they will affect the results should be made. Reply: The authors appreciate this reviewer's comment. Following the comments from the above, we tried our best to show how the assumption may affect the results. Hope the modification following the previous comment meet this reviewer's expectation.

19. The discussion section in the present form is relatively weak and should be strengthened with more details and justifications. Reply: The authors appreciate this reviewer's critical comment. The discussion has been intensified at the conclusion section. Hope this modification is satisfactory to this reviewer. Note that there is no separate discussion section in the current manuscript. If this reviewer implies other specific section, please let us know.

20. Moreover, the manuscript could be substantially improved by relying and citing more on recent literatures about contemporary real-life case studies of soft computing techniques in hydrological forecasting such as the followings: ïAËŻnËĞ Fotovatikhah, F., et al., "Survey of Computational Intelligence as Basis to Big Flood Management: Challenges, research directions and Future Work," Engineering Applications of Computational Fluid Mechanics 12 (1): 411-437 2018. (Fotovatikhah et al., 2018) ïAËŻnËĞ Wu, C.L., et al., "Rainfall-Runoff Modeling Using Artificial Neural Network Coupled with Singular Spectrum Analysis", Journal of Hydrology 399 (3-4): 394-409 2011. (Wu and Chau, 2011) ïAËŻnËĞ Taormina, R., et al., "Neural network river forecasting through optimization", Journal of Hydrology 529 (3): 1788-1797 2015. (Taormina et al., 2015) ïAËŻnËĞ Wang, W.C., et al., "Improved annual rainfall-runoff forecasting using PSO-SVM model based on EEMD," Journal of Hydroinformatics 15 (4): 1377-1390 2013.

(Wang et al., 2013) ïAËŻnËĞ Cheng, C.T., et al., "Flood control management system for reservoirs," Environmental Modeling & Software 19 (12): 1141-1150 2004.(Cheng and Chau, 2004) ïAËŻnËĞ Chau, K.W.,et al., "Use of Meta-Heuristic Techniques in Rainfall-Runoff Modelling" Water 9(3): article no. 186, 6p 2017. 21. (Chau, 2017) Reply: The authors appreciate relevant works. Almost all the suggested papers that are relevant with this study were included in the current study as the following:

"In this procedure, the multisite occurrence of the precipitation variable can be simulated with preserving spatial and temporal correlations. Note that meta-heuristic techniques to GA have been popularly employed in a number of hydrometeorological applications (Chau, 2017; Fotovatikhah et al., 2018; Taormina et al., 2015; Wang et al., 2013)."

Some inconsistencies and minor errors that needed attention are: ïAËŻnËĞ Replace ": : :had a slight better: : :" with ": : :had a slightly better: : :" in line 250 of p.13 22. In the conclusion section, the limitations of this study and suggested improvements of this work should be highlighted. Reply: The authors appreciate this reviewer's detailed comment. The suggested minor error was corrected accordingly, and the conclusion was modified accordingly as the following.

"In the current study, a nonparametric simulation model, based on discrete KNNR and DKNNR, is proposed to overcome the shortcomings of the existing MONR model such as long stochastic simulation for the parameter estimation and underestimation of the lagged crosscorrelation between sites. Occurrence and transition probabilities and cross-correlation as well as lag-1 cross-correlation are estimated for both models. Better preservation of cross-correlation and lag-1 cross-correlation with the DKNNR model than the MONR model is observed. For some cases (i.e., the whole year data and other seasons than the summer season), the estimated cross-correlation matrix is not a positive semi-definite matrix so the multivariate normal simulation is not applicable for the MONR model, because the tested sites are close to each other with high cross-correlation. Results of this study indicate that the proposed DKNNR

model reproduces the occurrence and transition probabilities fairly well and preserves the cross-correlations better than the existing MONR model. Furthermore, not much effort is required to estimate the parameters in the DKNNR model, while the MONR model requires a long stochastic simulation just to estimate each parameter. Thus, the proposed DKNNR model can be a good alternative for simulating multisite precipitation occurrence."

Please also note the supplement to this comment:
https://www.geosci-model-dev-discuss.net/gmd-2018-181/gmd-2018-181-AC1-supplement.pdf

---

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-181, 2018.