Central Asian Journal of Environmental Science and Technology Innovation **2** (2020) 101-108

RESEARCH PAPER

A technique for simulating future climate change variable using improved K-Nearest neighbors algorithm (k-NN)

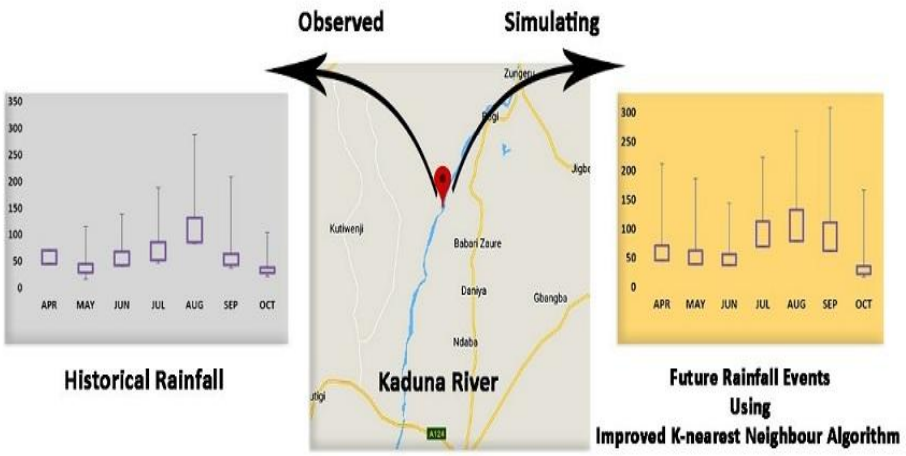
Haruna Garba *, Saminu Ahmed, Ibrahim Abdullahi

Department of Civil Engineering, Nigerian Defence Academy, Kaduna-Nigeria

Highlights

- The most critical and important need of Man was always water.
- Global environmental change is probably going to change precipitation examples and raise the recurrence of outrageous occasions.
- The K-NN algorithm model is a robust tool that can be used in predicting future rainfall values.

Graphical Abstract



Article Info

Receive Date: 09 November 2019
Revise Date: 10 January 2020
Accept Date: 15 February 2020
Available online: 20 February 2020

Keywords:
Algorithm
Generate
K-NN
Rainfall
Resampling
Simulation

Abstract

A technique for simulating future rainfall events using improved K-NN (K-nearest neighbors) algorithm is presented in this study. The K-NN calculation don't deliver new informational collections however simply reshuffles the chronicled information to create new informational indexes. A simulation day was selected in the month of August. The algorithm steps and resampling with ancient documents was applied to simulate rainfall events in Kaduna River catchment as a basis for future understanding of the characteristics of the basin. Simulated data sets for the months of April, May, June, July, August, September and October yielded nearly exact replica of the ancient documents. In performing the model, statistical characteristics such as mean, standard deviation, variance, cumulative probability, covariance, skewness, cross correlation are all preserved by the K-NN model. The technique clearly demonstrates that it can be used to generate future rainfall events that can be applied for hydrological investigation of the characteristics of a basin for future developments.

10.22034/CAJESTI.2020.02.05

E-ISSN: 2717-0519
P-ISSN: 2717-4034

* Corresponding author: h.garba1990@gmail.com (H Garba)

1. Introduction

The most critical and essential need of Man was always water. Understanding the hydrological processes includes a set of time and space scales. This includes rainstorms that happen through the span of minutes to hours and space sizes of a couple of Kilometers or less to the improvement of significant waterway bowls. As indicated, Worldwide Climate is relied upon to change altogether because of the consistent increment in levels of CO2 and other ozone-depleting substances (Sharif and Burn, 2006). Thus, numerous parts of the earth, including water assets, are foreseen to encounter genuine climatic effects that influence store activities, crop creation, disintegration forms, overflow creation, and numerous other hydrological forms. According to Intergovernmental Panel on Climate Change (IPCC) report in 2001, worldwide environmental change actuated by an increment in ozone harming substance fixations will probably change precipitation designs and most likely raised the recurrence of extraordinary occasions. Hydrological models are significant for a wide scope of uses, including water assets arranging, improvement and the executives, flood expectation and plan, and coupled frameworks displaying (Fowler et al., 2007; Pechlivanidis et al., 2011). Assets compel and constrained the scope of accessible estimation procedures to have restrictions on the accessibility of spatial-fleeting information for infrastructural and water assets improvement. Subsequently, the need to extrapolate data from a deliberate to an unmeasured circumstance in reality and evaluate the possible effect of future framework reaction to the atmosphere and land the executives changes.

The inverse distance weighting technique was used for rainfall interpolation that was considered a suitable method of predicting missing rainfall records (Garba et al., 2018). Results of analysis through optimization of steps were entirely satisfactory, using the radius of influence to estimate the optimum parameter values; the smaller the optimum value, the better the prediction, and the forecast's accuracy increases at short optimum radii (Young, 1994). A small amount and long-duration rainfall values enhance the prediction potential of the Inverse Distance Weighting (IDW). In the other study, a system was presented to look at future precipitation situations utilizing the Fuzzy Clustering Technique from Global Circulation Models (GCM) projections (Ghosh and Mujumdar, 2008; Folorunsho et al., 2012). In the methodology, the dimensionality of data sets was reduced by Principle Component Analysis (PCA). The fluffy grouping method was then applied to arrange the rule part distinguished by PCA. The fluffy enrolment esteem was then utilized in a relapse model. The relapse model is changed with a standard term. The proposed system is computationally straightforward and can display precipitation with a high decency of fit (R2) esteem.

An improved climate creating model, which permits closest neighbour resampling with the change of memorable information, was applied to produce climate information dependent on the conceivable situation to sufficiently reproduce extraordinary occasions for five positions (Sharif and Burn, 2007). Because of the reproduction results, the expanding precipitation situation was distinguished as the basic variable for evaluating hazards related to the examination of floods in a streaming bowl, while the expanding temperature variable shows up the most basic for the investigation of dry spells (Chinn, 1993). Recurrence investigation was then done to decide the effect of potential environmental change on the events of tempest occasions of some random size.

A K-NN resampling plan mimics day by day climate factors, occasional atmosphere, and spatial and worldly conditions for various stations in a district. The K-NN Algorithm utilized Mahalanobis separation has the bit of leeway that factors don't need to be normalized nor have a prerequisite to reassigns weight to factors like the measurement for neighbour determination (Yates et al., 2003). The calculation was utilized to create elective atmosphere situations dependent on recommended moulding rules. Precipitation events and sum are created autonomously, and other climatic factors are produced dependent on stochastically produced precipitation as analysed by (Nick and Harp 1980; Richardson 1981).

The paper aims to improve and apply a climate generator dependent on parametric factual procedures utilizing precipitation as the driving variable for the executives' infrastructural water assets (Rajagopolan and Lall, 1999).

2. Materials and Methods

Stochastic climate generators are routinely utilized in water; horticulture and disintegration control the board. Climate generators dependent on parametric factual method ordinarily use (precipitation) as a driving factor in various models, where precipitation events and sum are created freely, and different aspects produced dependent on the stochastically produced precipitation (Nicks and Harp 1980; Richardson 1981). Synchronous testing of the climate factors including precipitation and temperature engaged with K-NN approach The testing is done from the watched information with the swap, for example, to demonstrate factors for another day, t+1 days with comparative attributes as those watched, day t is chosen from the notable record (Buishand and Brandsma, 2001). A characterized likelihood dissemination or Kernel then chooses one of the closest neighbors. The day's watched qualities, resulting in that nearest neighbour, are embraced as model and incentive for day t+1. Models create dependent on the K-NN Mahalanobis separation method can be stretched out to multisite expectation of climate information while charging the four-dimensional association organization of the notable information unblemished. The spatial conditions are saved because that day's climate is embraced as the climate for all stations. Besides this, transient conditions are probably going to be safeguarded as the anticipated qualities for t+1 are moulded on the qualities for day t. The cross-relationship among the factors is protected as a square of factors is resampled from the watched information.

Think through that day by day noteworthy climate vector comprises of p factors, p=3, which incorporates most extreme temperature (T max), least temperature (T min), and precipitation (PPT). Accept the amount positions measured in the model is q and information are accessible for N years, and X_t^j signify the vector of climate factors for day t and station j, where *t*=1……..T and *j*=1……..*q;* T being the all-out number of days in the watched time arrangement. The element vector for day t can be communicated in an extended structure as X_t^j= [X_ (1.t,) ^j X_ (2.t) ^j… ….X_ (p.t) ^j] where X_ (1.t,)^j speak to the estimation of climate inconstant I for station j.

Assume that the recreation starts on a day t comparing to January 1. The calculation pushes through the means underneath to acquire the climate for day t+1. The methodology proceeds for every one of the 365, and the technique is rehashed to create information for whatever number a very long time as could be expected under the circumstances.

2.1. Algorithm steps

Compute provincial methods for the p factors over the q positions for every day of the memorable record.

$$\overline{X_t} = [\overline{X_{1.t,}X_{2.t}}\ldots\ldots\ldots\ldots\overline{X_{p.t}}] \quad (1)$$

Where:

$$\overline{X_{1.t,}} = \frac{1}{q}\sum_{j=1}^{q} X_{i.t}^{j} \ ;i=1\ldots..p,t=1\ldots\ldots T \quad (2)$$

a) Decide the magnitude L of information obstructs that incorporates every probable neighbour to the present component vector, which the resampling is to be finished. A transitory window is measured as a possible possibility to the current component vector. Utilized a brief window of 14 days, which suggested that if the present day is January 15, at that point, the window of days comprise of the entire days between January 10 and January 24 for all N years yet barring January 15 for the given year (Yates et al., 2003). In this manner, the information square of possible neighbors from which to resample comprise of L= (w+1)×N-1 days. Register mean vectors across q stations every day in the information square comprising possible neighbors utilizing the articulation in condition 1.

b) Calculate the covariance matrix, C_1 using the data block of size $L\times P$

c) Control the quantity of first K-closest neighbors to be utilized for resampling out of the absolute L neighbors. Choosing K by using the summed up cross approval score (GCV) (Lall and Sharma, 1996). As it is possible, in another study, the use of experimental methods for picking K was shown, according to which K = √L, the exhibition of the calculation with this worth, was seen as acceptable (Yates et al., 2003).

d) Register the Mahalanobis separation between the mean vector of the present day's climate, $(X\_t)^{-}$ and the mean vector $(X\_i)^{-}$ for the day I where I =1 L. The separation measurements can be characterized as;

$$d_i = \sqrt{(\overline{X_t} - \overline{X_\iota}) C_t^{-1} (\overline{X_t} - \overline{X_\iota})^{\mathrm{T}}} \quad (3)$$

T addresses the transverse movement, and C_t^ (- 1) is the opposite of the covariance lattice.

e) Category the Mahalanobis removes in rising request and hold the principal K-closest neighbors. A discrete likelihood conveyance that offers greater loads to the nearer neighbors was utilized for resampling from the K-closest neighbors. Loads are doled out to every one of these j neighbors as per the measurement characterized by;

$$W_j = \frac{1/j}{\sum_{i=1}^{K} 1/i} \quad (4)$$

The combined possibilities P_j are given by;

$$P_j = \sum_{i=1}^{j} W_i \quad (5)$$

The neighbour with the littlest separation is doled out the most noteworthy loads, while the neighbour with the most elevated separation gest the least weight. This capacity was created by Lall and Sharma (1996) through a nearby Poisson guess of the likelihood.

f) Decide the closest neighbour of the present day by utilizing the likelihood metric in Eq (5). Produce an arbitrary number r ⊂ (0.1) in the event that P_1< r<Pk; at that point, the j for which r is Closest is chosen to P_1 is chosen. On the off chance that r≤P_1, the day relating to d_i is selected, and on the off chance that r ≥P_K, at that point, the day was comparing to d_k is chosen. The observed qualities for the day resulting in the chosen closest neighbors are embraced to speak to the Weather for day t+1 (Eum et al., 2010; Stott, 2016). In this altered methodology, the information focuses resampled utilizing the essential K-NN approach, including an irregular segment, as portrayed in the means underneath.

g) For each station and every factor, non-parametric dissemination is fitted to the K standard deviation σ and data transmission λ. Annoyances of the estimations of climate factors acquired utilizing the fundamental K-NN approach is completed with the accompanying advances;

h) Let σ_i^j be the contingent standard deviation of variable I for station j figured from the K closest neighbour. Let Z_(t+1) be an irregular variety for day t+1 in the recreation time frame from an ordinary conveyance with zero mean and unit fluctuation. The new estimation of climate variable I for day t +1 and station j is given by;

$$y_{i.t+1}^{j} = X_{i.t+1}^{j} + \lambda \sigma_i^j Z_{t+1} \quad (6)$$

Where,

$X_{i.t+1}^{j}$ Is the estimation of the climate variable for t+1 and station j got from the essential K.NN model, y(i.t+1)^j is the relating esteem gotten after annoyance and λ transfer speed (an element of the number of tests)? Steps (f-j) are rehashed to create the same number of long periods of engineered information as required.

2.2. Kaduna river catchment

Kaduna River catchment (Fig. 1) has a total drainage area of approximately 18,244.87 km² within the catchment. There are seven meteorological data collection points located at Kaduna North, Kaduna South, Zaria, Zonkwa, Kaura, Saminaka, and Kangimi. Kaduna State, which involves a focal situation in the Northern topographical locale of Nigeria and exists in the Northern Savana Zone of Nigeria, is situated on scope 9030'N and scope 11045'N; longitude 70E and 8030'E. It covers a complete landmass of 2,896,000 km².

2.3. Historical data

Precipitation in the study area has an uneven spatial and temporal distribution. The average annual precipitation is usually below 300 mm, mostly concentrated between May and September. There are seven meteorological data collection points in the Kaduna River catchment, located at Kaduna North, Kaduna South, Zaria, Kauru, Kangimi, Zonkwa, and Saminaka. The geological area of the stations as decided from Latitudes and Longitudes appears in Fig. 1. Missing historical data records for the stations were unfilled with mean values in the study (Garba et al., 2018). Available data consists of records from the stations from 1975-2000.

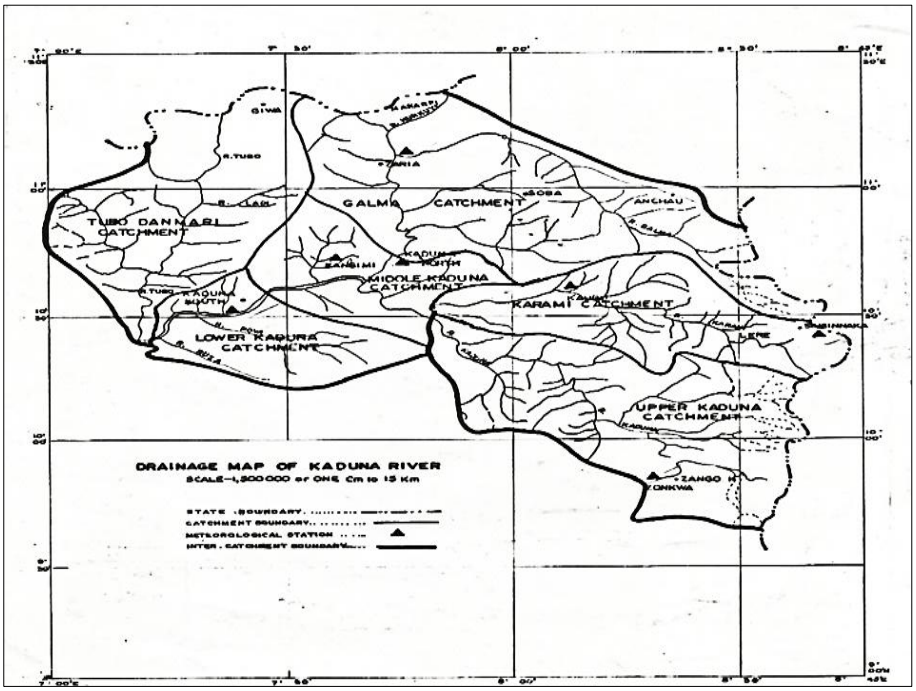


Fig. 1. Drainage Map of Kaduna River.

2.4. Model application

Evaluating the K-NN model's act be influenced by a large extent of choosing the K-nearest neighbors and the width of the temporary window w. A temporary window of 14 was adopted in the study (Yates et al., 2003). August 8 was considered the simulation data, and therefore the window of days comprises the entire days between August 1 and August 15, excluding August. The data block of potential neighbors from which resampling was done is 375. The K quantities of closest neighbors were legitimate and proposed excellent execution of the model. The Mahalanobis distance was used as the frequency of the weighing factor from which the discrete probability distribution of the variable of interest was evaluated. The produced arbitrary variable was contrasted, and the discrete likelihood dispersion and the day were chosen as the climate for locale under survey. The resampling of data was applied to perturb the historical data and generate data outside the historical record by estimating conditional standard deviation for all the K-nearest neighbors (Sharif and Burn 2006). Calculation stages 1 to 6 were applied to register the territorial methods, decided the size of information square, figure mean vectors, process the covariance, decided the quantity of the closest neighbor to be held for resampling, and the climate for the principal reproduction day of the verifiable information. Future simulation of the variable was achieved by adopting the algorithm steps 6 to 10. The climate for day t + 1, which is the first

re-enactment, was controlled by; processing the Mahalanobis separates between the mean vector and current days climate; the separations are arranged in the climbing request, A discrete likelihood that gives high weight to the nearest neighbor was chosen. A random number is selected which is to the cumulative probability based on the weight is attached. The watched qualities ensuing to the chose closest neighbor was chosen as and received to speak to the climate for day t + 1. The calculation stages 6 to 10 can be rehashed to produce the same number of long stretches of engineered information as required. Figs. 2 and 3 shows box plots of watched and reproduced month to month precipitation for the catchment zone. The measurements of watched and recreated information every day esteems were totalled.

3. Results and Discussion

Box plots are employed for groups of data scales and scores. They enable the study and understanding of the characteristics of a group of data as well as the level of scores. Data are sorted and placed in groups, which are 25% of all scores.

The box plots in Figs. 2 and 3 are the historical and simulated rainfall events for April, May, June, July, August, September, and October. From the results, 25% of the data is less than 50 mm and 75% less than 150mm. The results show a high agreement between simulated and observed rainfall for April, June, July, August, and September. The long whisker in July, August, and September create excellent variability in rainfall. It can be seen that the algorithm process yielded the exact reproduction of historical data. In determining the nearest neighbor of the simulation date, a random number generated was observed to be close to the neighbor's cumulative probabilities with the smaller distances. Due to the historical data perturbations, a maximum rainfall value of 150 mm was observed in the simulated and historical data sets. However, there is prominent variability between the simulated and historical data for May.

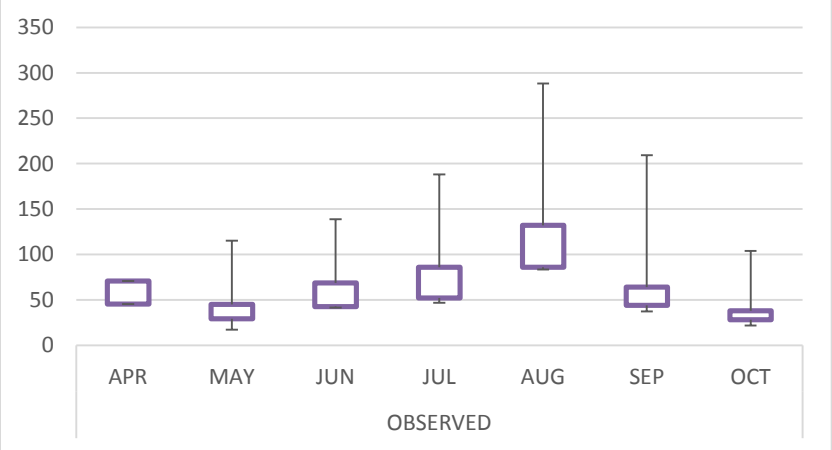


Fig. 2. Box plot of historical rainfall.

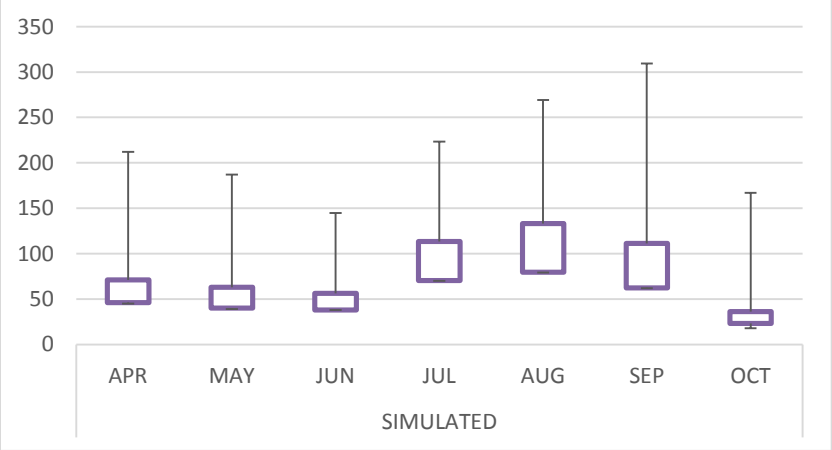


Fig. 3. Box plot of simulated rainfall.

4. Conclusions

A technique for simulating future rainfall events using an improved K-nearest neighbor algorithm is presented in this study to assess the algorithm processes in simulating future rainfall events from historical data sets. The results presented by Box plots clearly showed that a model is a robust tool that can be used in predicting future rainfall values.

Correlation of the chronicled month to month esteems with the re-enacted qualities indicated that the model could replicate the authentic qualities satisfactorily. The conspicuous distinction between boxes plots for the months under audit requires further examination.

Reference

Buishand, T.A., Brandsma, T., 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest neighbour resampling. *Water. Resour. Res.*, **37**(11), 2761-2776. https://doi.org/10.1029/2001WR000291

Chinn, T.J., 1993. Physical hydrology of the dry valley lakes. *Phys. Biogeochem. Process. Antarct. lakes*, **59**, 1-51. https://doi.org/10.1029/AR059p0001

Eum, H.I., Simonovic, S.P., Kim, Y.O., 2010. Climate change impact assessment using k-nearest neighbor weather generator: case study of the Nakdong River basin in Korea. *J. Hydrol. Eng.*, **15**(10), 772-785. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000251

Folorunsho, J.O., Iguisi, E.O., Mu'azu, M.B., Garba, S., 2012. Application of adaptive neuro fuzzy inference system (Anfis) in river Kaduna discharge forcasting. *Res. J. Appl. Sci. Eng. Technol.*, **4**(21), 4272-4283. ISSN: 20407459

Fowler, H.J., Blenkinsop, S., Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol. J. Roy. Meteorol. Soc.*, **27**(12), 1547-1578. https://doi.org/10.1002/joc.1556

Garba, H., Tilli, L.G., Ahmed, S. Ibrahim, A., 2018. Rainfall interpolation analysis on River Kaduna catchment for climate change assessment. *Niger. J. Technol.*, **37**(3), 806-812. https://doi.org/10.4314/njt.v37i3.33

Ghosh, S., Mujumdar, P.P., 2008. Statistical downscaling of GCM simulations to stream flow using relevance vector machine. *Adv. Water. Resour*., **31**(1), 132-146. https://doi.org/10.1016/j.advwatres.2007.07.005

Lall, U., Sharma, A., 1996. A nearest neighbour bootsrap for time series resampling. *Water. Resour. Res.*, **32**(3), 679-693. https://doi.org/10.1029/95WR02966

Nicks, A.D., Harp, J.F., 1980. Stochastic generation of temperature and solar radiation data. *J. Hydrol.*, **48**(1-2), 1-7. https://doi.org/10.1016/0022-1694(80)90062-1

Pechlivanidis, I.G., Jackson, B.M., Mcintyre, N.R., Wheater, H.S., 2011. Catchment scale hydrological modelling; A review of model types, calibration approaches and uncertainty analysis methods in the context of recent development in technology and applications. *Global. NEST. J.*, **13**(3), 193-214. https://doi.org/10.30955/gnj.000778

Rajagopolan, B., Lall, U., 1999. A K-nearest- neighbour simulator for daily precipitation and other variables. *Water. Resour. Res.*, **35**(10), 3089-3101. https://doi.org/10.1029/1999WR900028

Richardson, C.W., 1981. Stochastic simulation of daily precipitation temperature and solar radiation. *Water. Resour. Res.*, **17**(1) 182-190. https://doi.org/10.1029/WR017i001p00182

Sharif, M., Burn, D.H., 2006. Simulating climate change scenarios using Improved K-nearest neighbour model. *J. hydro.* **325**(1-4), 179-169. https://doi.org/10.1016/j.jhydrol.2005.10.015

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

Stott, P., 2016. How climate change affects extreme weather events. *Sci.*, **352**(6293), 1517-1518. https://doi.org/10.1126/science.aaf7271

Yates, D., Gangopadhyay, S., Rajagopolan, B., Strzepek, K., 2003. A technique for generating regional climate scenarios using a nearest- neighbour algorithm. *Water. Resour. Res.,* **39**(7) 1-15. https://doi.org/10.1029/2002WR001769

Young, K.C., 1994. multivariate chain model for simulating climate parameters from daily data. *J. Appl. Meteorol.*, **33**(6), 661-671. https://doi.org/10.1175/1520-0450(1994)033<0661:AMCMFS>2.0.CO;2

How to cite this paper:
Garba, H., Ahmed, S., Abdullahi, I., 2020. A technique for simulating future climate change variable using improved K-Nearest neighbors algorithm (k-NN). *Cent. Asian J. Environ. Sci. Technol. Innov.* **2**, 101-108.