Other approaches which have captured researcher’s attention in the past decades are the artificial neural networks (ANNs) applied in various fields of hydrology engineering including classification, forecasting and modeling problem. ANN application in hydrology due to its high nonlinear functional characteristic has provided rapidly many advantages in river flows extrapolation Cigizoglu (2003), rainfall-runoff modeling Firat (2008), sediment forecasting Wang et al. (2008), and ET-ref modeling Kumar et al. (2002). Researchers have obtained outstanding results by using different algorithms of ANN to model the reference evapotranspiration as a function of climatic data Trajkovic et al. (2003), Keskin and Terzi (2006), Parasuraman et al. (2007), Doğan (2009), Sathishkumar and Cho (2020), Sudheer et al. (2003), and Zanetti et al. (2007) in their reference evapotranspiration estimation, simplified the ANN inputs data to air temperature, extra-terrestrial solar radiation and daily light hours. Recently, Khoob (2008) and Landeras et al. (2008) used similar data set without the daily light to estimate successfully the reference evapotranspiration. By observing the study sites of the above-mentioned studies, it is found that there is no study conducted under the climatic condition of the Sudano-Sahelian zone.

With the rise of machine learning technology and artificial intelligence, scholars began to explore how to combine intelligent algorithms with traditional estimation methods to estimate ET₀ more accurately and effectively. Kisi used temperature, solar radiation, relative humidity and wind speed as inputs in 2015 to compare the accuracy of LSSVM, MARS and MSTree in estimating monthly ET₀. The results show that LSSVM has the smallest relative error in the test period Kisi (2015). In 2017, ADNAN M first used principal component analysis (PCA) to find 5 meteorological data most relevant to ET₀ from seven meteorological data, namely, the highest, lowest, average temperature, precipitation and wind speed, and reduced their dimensions, then combined with ANN algorithm to predict ET₀. This method not only saves the time and cost of calculation, but also keeps the accuracy of prediction. Based on five meteorological data, eight input combinations were set up by Mattar (2018) who established eight ET₀ estimation models using GEP intelligent algorithm. The estimation results are very close to FAO-56 PM estimation. The results also show that GEP is more accurate than Hargreaves and Samani, Irmak, Turc in estimating ET₀. However, we found that the time complexity of these machine learning algorithms is high, and still needs more meteorological data as input.

The objective of this study is to investigate the potential of k-Nearest Neighbor algorithm (KNN) for estimating ET₀ using limited climatic data in a semi-arid environment in China.

The objective of this study is to demonstrate the adequacy of different approach for forecasting daily ET₀ using data mining and limited weather information. The data mining algorithm used is the k-Nearest Neighbor algorithm (KNN). The reference evapotranspiration is estimated by the Penman-Monteith equation (PM-56) using the meteorological data from 1951 to 2000 of the 24 weather stations in Ningxia. And then, analyzed the correlation between ET₀ and meteorological elements, in order to obtain the meteorological elements which are most closely with ET₀. These will be converted into vector space and use KNN algorithm to predict ET₀ with the meteorological data from 2001 to 2012. The next section presents a description of the study area and the methods applied in this study and provide the information about data, methodological structure and statistical indexes. The applicability of the models on evapotranspiration estimation and the results are examined in the third section. Finally, the last section provides conclusions.

Materials and methods

Study area and climate dataset

Ningxia irrigation area is China’s four major ancient irrigation districts, one of more than 2000 years of irrigation history, known as “Frontier of Jiangnan” reputation, mainly grain, cotton and oil-producing areas in Ningxia. It is also one of the 12 commodity grain bases in China. The irrigation area is located in the Ningxia Hui Autonomous Region, which located in northwest China (Geographical coordinates: east longitude 104 ° 17' ~ 107 ° 39', latitude 35 ° 14' ~ 39 ° 23'. North and south about 465 km long, 45–250 km wide) and covering some 60,000 km². With the Yellow River passing through the region, Ningxia enjoys a convenient irrigation system. There are numerous rivers, lakes and channels in the region. As shown in Figure 1.

Ningxia few precipitations, evaporation strongly, and the air is dry. For many years, the average annual rainfall is 289 mm, from north to south increasing, change in 180 ~ 800 mm. In the water evaporation capacity 1250 mm, is 4.3 times of precipitation, the trends and precipitation instead, from north to south decline, change in 1600 ~ 800 mm. The above two contrary trend decided the difference between the north and the south is drought index, by southing north change in 1 ~ 9, most areas for 3 ~ 9, belong to an arid and semi-arid area.

The weather data were collected from the Meteorological Administration of China (www.cma.
The data is composed of mean, maximum and minimum air temperatures, relative humidity, wind speed, atmospheric pressure and sunshine hours for the period 1951–2012. The data is divided into two parts. The first part (1951–2000) was used to calculate ET0 using the Penman-Monteith equation (PM-56) and train the KNN. The second part (2001–2012) was used for validation as testing period. The first 50 years (1951–2000) data were used to calculate ET0 using the Penman-Monteith equation (PM-56) and train the KNN, and the remaining data were used for validation. It should be noted that the mean 10-days values of the weather data were used for the analysis.

Penman-Monteith method

The FAO-56 Penman-Monteith equation which is given by Allen et al. (1998) as follows:

\[
ET_0 = \frac{0.408 \Delta (R_n - G) + 900 \frac{\gamma}{T_m} U_2 (e_s - e_a)}{\Delta + \gamma (1 + 0.34U_2)}
\]  

where ET0 is the reference crop evapotranspiration (mm day/1), \(R_n\) is the net radiation [MJ/(m²·d)], G is the soil heat flux [MJ/(m²·d)], \(\gamma\) is the psychrometric constant (kPa/°C), \(e_s\) is the saturation vapor pressure (kPa), \(e_a\) is the actual vapor pressure (kPa), and \(\Delta\) is the slope of the saturation vapor pressure–temperature curve (kPa/°C), \(T_m\) mean is the daily mean air temperature (°C), and \(U_2\) is the mean daily wind speed at 2 m/m/s. The computation of all data required for calculating ET0 followed the method and procedure given in Chapter 3 of FAO-56.

Correlation analysis

This study analyzed the correlation between the meteorological elements and ET0 using Pearson correlation coefficient, which defined as follows:

\[
r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2}(y - \bar{y})^2}
\]

where \(x\) and \(y\) represent the variable to be predicted and the predicted value, respectively, \(r\) is the Correlation coefficient which is in the range [-1,1]; \(r>0\) indicates a positive correlation, \(r<0\) indicates a negative correlation, \(|r|\) represents the low degree of correlation between the variables. Specially, \(r = 1\) is called perfect positive correlation, \(r = -1\) is called perfect negative correlation, and \(r = 0\) is called irrelevant. Typically, \(|r|\) is greater than 0.8, that the two variables have a strong linear correlation.

K-nearest neighbor algorithm (KNN)

The KNN technique can be used especially for classifying data into multiple categories, but it also can be successfully applied for purposes of forecasting some objective-related attributes that have a numerical value, as a result of non-linear dependencies. This confers the possibility of performing multiple category classifications and modeling non-linear data relationship (for forecasting). The KNN algorithm does not force the samples to satisfy a specific distribution. But if the sample set is normally distributed, the prediction effect will be better. The data set used in this article follows a normal distribution.

Although it is a base algorithm within instance-based learning, KNN poses some disadvantages that need to be taken into account, depending on the nature of the problem. Firstly, the algorithm does not excel in computational speed when the dataset contains a large number of instances. Secondly, the algorithm is limited to supplying an estimation of the objective-attribute value without offering other information about the instance being evaluated or the dataset as a whole.

The KNN class will implement the algorithm based on finding \(k\) of the nearest neighbors of some
instances is the dataset. The first problem the algorithm implementation is going to be facing is that of transposing these instances in a normalized form, which can be represented in a two-dimensional space as points with their corresponding coordinates. The problem lies in resolving these coordinates in order to be able to apply the geometrical formulas that represent the steps of this algorithm.

For this purpose, the KNN class will first define two helper methods. The first method computes the arithmetic average of the numerical projection of an attribute value, projection that is identified by the Numeric Value property. The value of this property is supplied by the following mathematical expression:

$$A = \frac{\sum v}{|S|}$$  \hspace{1cm} (3)

where \(v\) will be iterating through all the possible values of attribute \(A\) and \(S\) is the dataset.

The second method computes the standard deviation of an attribute using the following formula:

$$\sigma_A = \sqrt{\frac{\sum (v - \bar{A})^2}{|S|}}$$  \hspace{1cm} (4)

The parameters in Equation 4 have the same definitions as the parameters in Equation (3). These values are needed in order to be able to scale the dataset in such manner that the global arithmetic average of the entire dataset will be zero and the standard deviation will be 1. This scaling is performed by replacing the numeric value of each attribute with a new value obtained using the following calculation:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$  \hspace{1cm} (5)

where \(v\) represents the value of the attribute \(A\), which will be replaced with its new value \(v'\).

The implementation will have to create a method that will generate a matrix representation of the initial dataset. This matrix structure will be represented through the Dictionary classes, thus offering the possibility to access a value based on its attribute and the instance it can be found in. The KNN algorithm relies on evaluating distances between two points in the dataset plan; hence the need of a method that can compute this distance. Given that there is already a matrix containing the scaled values of the dataset and these can be easily accessed based on their initial source, the Calculate Distance method will be defined with two parameters of type Instance, which will be responsible for performing all the required projections for the instances; an invocation of this method could be interpreted as evaluate the distance between instances \(x\) and \(y\). The method uses the Euclidian procedure for computing the distances; the distance between two instances in the dataset identified by the (scaled) values of their attributes is defined by the Euclidian geometry through the following formula:

$$d(a, b) = \sqrt{\sum \left(\frac{a'_A - b'_A}{\sigma_A}\right)^2}$$  \hspace{1cm} (6)

where a and b are instances within the dataset; \(A\) will iterate though each attribute values in the dataset; \(x_A\) represents the value of attribute \(A\) for instance, \(x\); \(x'\) is the scaled value of that attribute.

### Parameters of error analysis

The performances of the models were evaluated using the following statistical parameters: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R²). The R², RMSE, MAE are commonly used indicators to evaluate the prediction performance of the algorithm. Generally, with a good R², the RMSE and MAE will be as small as possible. Readers interested in the formulations of the above-mentioned statistical parameters are referred to Hosseinzadeh Talaee et al. (2011).

### KNN-based ET₀ forecast model

The basic principle of this model is: receiving a new data item to be predicted, then concentrated with a set of sample data items are compared to find out the data items to be predicted with the K closest data items, and its demand means to obtain a final prediction result, as shown in Figure 2. Throughout the forecast model execution process is, first, to be predicted from meteorological data to extract features using correlation analysis, these features will be converted into vector space which like Table 1 shows, and then use KNN algorithm to predict the resulting prediction ET₀.

![Figure 2](https://via.placeholder.com/150)
Results and discussions

Correlation between ET₀ and meteorological elements

Pearson correlation analysis has been carried out between meteorological elements and daily ET₀ which calculated by FAO-56 Penman-Monteith equation. And the results are shown in Table 2.

Various meteorological elements and average daily ET₀ relevance in descending order are maximum temperature, relative humidity, average temperature, average wind speed, minimum temperatures, and sunshine hours. ET₀ negatively correlated with the relative humidity, the rest of meteorological elements are positively correlated. In addition to elements of sunshine hours was significantly correlated at the 0.05 level, other factors were significantly correlated at the 0.01 level. These data suggest that in the Ningxia region temperature and relative humidity are the main factors affecting ET₀, ET₀ increases with rising temperature, the relative humidity drops decrease. Therefore, in this study, we will use the maximum temperature, minimum temperature and relative humidity which to predict ET₀.

Results of KNN algorithm forecasting

First, calculate an average day ET₀ with PM-56 Formula, from 1951 to 2000 years. Then, based on limited meteorological data (max temperature, min temperature and relative humidity) of 2001–2012 years, using KNN algorithm (K = 1) forecast annual average daily ET₀. The predicted results and PM-56 calculated results were analyzed, the results shown in Figure 3.

Clearly, the predicted results are not satisfactory, RMSE = 1.37, MAE = −0.58, R² = 0.6453. The predicted value than the PM-56 formula calculation value much lower. This shows that there is underestimation. We found that the composition predicted characteristic vector space the dimensions are not the same, and their range is different. Such as temperature ranges from −20°C to 35°C, relative humidity is between 10% and 50%. This is clearly not conducive to precise prediction, we try to normalized features.

Results of KNN algorithm with normalization

After normalization processing features, the new forecast results shown in Figure 4. The result was better than the previous few, but there are still underestimated. RMSE = 1.014, MAE = −0.365, R² = 0.7836. So we need to continue to improve the prediction algorithm.

Results of KNN algorithm with different k value

In the KNN algorithm, K represents the selected k nearest neighbors in the forecasting process. The, what would be different k values influence on the predicted results? We tried different situations, such as K = 1, K = 3 and K = 5. The results shown in Figures 5-7 and Table 3.

We found that different values of k on the predicted effect is very large. Too small values of k or too large values of K have caused the decline in the prediction
accuracy, the error increases, $R^2$ becomes smaller. When the value of $k$ is small or large, the error of the estimation result will increase. Therefore, the algorithm must first choose an appropriate value of $k$, which is a hyperparameter. A more appropriate value of $K$, $K = 3$. At this time the prediction accuracy is acceptable.

Results of weighted KNN algorithm

We observe predictions KNN algorithm ($K = 3$), and the result compared with PM-56 is not very good. So, we try given different weights to different neighbors. Prior to this, the weight of each neighbor is the same, because we take the average of each neighbor. Gaussian function is chosen as the right of our valued function.

The reason is this function will give high weight to the near neighbors, and the relatively give a low weight to the distant neighbors, but the weight will not be zero. Weighted KNN algorithm produces prediction results

Table 3. Errors of different $k$ statistical analysis.

| K value | RMSE (mm/d) | MAE (mm/d) | $R^2$  |
|---------|-------------|------------|--------|
| $K = 1$ | 1.263       | -0.844     | 0.7972 |
| $K = 3$ | 0.687       | -0.007     | 0.8854 |
| $K = 5$ | 1.564       | -1.186     | 0.7491 |
shown in Figure 8. We can find, the curve of prediction results is consistent with the curve of PM-56 results. RMSE = 0.5778, MAE = 0.091 and R² = 0.9209.

Because the Penman-Monteith method estimates ET₀, it is necessary to input more meteorological observation data, which makes the method difficult to apply in some developing countries or regions where the observation equipment is not available. With the rise of machine learning technology and artificial intelligence, scholars have begun to explore how to combine intelligent algorithms with traditional estimation methods to estimate ET₀ more accurately and effectively.

Compared with other methods for estimating potential evapotranspiration, the method proposed in this paper uses less meteorological data, which provides a way to estimate evapotranspiration in areas where meteorological data is incomplete. In addition, the method in this study is simple, and the accuracy can meet the application of hydrology and water resources related fields.

Conclusions

In this study an attempt was made to determine the best method for estimating ET₀ in the absence of the full weather data for PM-56 method application in Ningxia which is a semi-arid environment in China. We have designed the ET₀ prediction model, which based on KNN algorithm. During the testing process, we found that the original KNN algorithm prediction accuracy is not good, so we made some improvements processing, such as normalization, different K values, and different weights. Finally, The KNN-ET₀ Forecast Model provided good agreement with the ET₀ obtained by the PM-56 method. Meanwhile, the Gaussian function was the best membership function for the KNN models. Furthermore, more weighting function will be to try to improve the prediction more accurate for the KNN Based ET₀ Forecast Model.

The k-Nearest Neighbor algorithm (KNN) is a non-parametric lazy supervised classification algorithm in machine learning technology. The principle is simple and easy to implement. The algorithm time complexity is only O(n). It has many advantages: (1) the algorithm is simple, easy to understand, and easy to implement, without the need to estimate parameters. (2) The training time is zero. It does not show training, unlike other supervised algorithms that use the training set to train a model (that is, fit a function), and then use the model to classify the validation set or test set. KNN just saves the sample and processes it when it receives the test data, so KNN training time is zero. (3) Compared with algorithms such as Naive Bayes, there is no assumption on the data, the accuracy is high, and it is not sensitive to outliers. Its disadvantages include: (1) It is impossible to give rules like decision trees. (2) It is a lazy learning method, resulting in slower prediction speed than algorithms such as logistic regression. (3) When the samples are unbalanced, the prediction accuracy rate for rare categories is low. Considering the advantages and disadvantages of this algorithm comprehensively, it needs to be treated with caution in practical applications, and take some appropriate measures to ensure stable and reliable operation of the algorithm.

The findings of this case study provide basic guidance to irrigation engineers and agriculturists as to which models will give a better estimate of ET₀, in light of data availability, for irrigation scheduling and water resources management. The KNN Base ET₀ forecast model developed here can be embedded as a module for estimating ET₀ data in hydrological modeling studies in the study area and in areas with similar hydro-meteorological characteristics. Further research is needed to test the model used here in other climates for evaluation of climate type effects.

Acknowledgments

This work was supported by the Major Innovation Projects for Building First-class Universities in China’s Western Region (No. ZKZD2017002); the Scientific Research Project of Ningxia Colleges and Universities, China (No. NGY2020006); the University First-class Discipline Construction Project of Ningxia, China. (No. NXYLK2017A03); the Natural Science Foundation of Ningxia, China (No.2019AAC03049); the China Scholarship Council (No.201708645016); and the Priority Research and Development Projects for Ningxia (No.2020BEG03021).
Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This work was supported by the China Scholarship Council [201708645016]; Major Innovation Projects for Building First-Class Universities in China's Western Region [ZKZD2017002]; The Priority Research and Development Projects for Ningxia [2020BEG03021]; University First-Class Discipline Construction Project of Ningxia, China [NXYLXK2017A03]; Natural Science Foundation of Ningxia, China [2019AAC03049]; The scientific research project of Ningxia Colleges and universities, China [NGY2020066].

References
Adnan, M., Latif, M. A., & Nazir, M. (2017). Estimating evapotranspiration using machine learning techniques. International Journal of Advanced Computer Science and Applications, 8(9), 108–113. https://doi.org/10.14569/IJACSA.2017.080915
Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). FAO Irrigation and drainage paper No. 56. Rome: Food and Agriculture Organization of the United Nations, 56 (97), e156. http://www.climasouth.eu/sites/default/files/FAO%2056.pdf
Bandoc, G., & Golumbeau, M. (2010). Climate variability influence to the potential evapotranspiration regime of the sfantu gheorghe delta shore. Journal of Environmental Protection and Ecology, 11(1), 172–181. http://www.jepe-journal.info/vol-11-no-1
Blaney, H. F. (1952). Determining water requirements in irrigated areas from climatological and irrigation data. U.S. Department of Agriculture Soil Conservation Service.
Cigizoglu, H. K. (2003). Estimation, forecasting and extrapolation of river flows by artificial neural networks. Hydrological Sciences Journal, 48(3), 349–361. https://doi.org/10.1623/hysj.48.3.349.45288
Doğan, E. (2009). Reference evapotranspiration estimation using adaptive neuro-fuzzy inference systems. Irrigation and Drainage. The Journal of the International Commission on Irrigation and Drainage, 58(5), 617–628. https://doi.org/10.1002/ird.445
Firat, M. (2008). Comparison of artificial intelligence techniques for river flow forecasting. Hydrology and Earth System Sciences Discussions, 12(1), 123–139. https://doi.org/10.5194/hess-12-123-2008
Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. Applied Engineering in Agriculture, 1(2), 96–99. https://doi.org/10.13031/2013.26773
Heydari, M., & Talaeec, P. H. (2011). Prediction of flow through rockfill dams using a neuro-fuzzy computing technique. The Journal of Mathematics and Computer Science, 2(3), 515–528. https://doi.org/10.22436/jmcs.02.03.15
Jensen, M. E., & Haise, H. R. (1963). Estimating evapotranspiration from solar radiation. Proceedings of the American society of civil engineers. journal of the irrigation and drainage division, 89, 15–41.
Keskin, M. E., & Terzi, Ö. (2006). Artificial neural network models of daily pan evaporation. Journal of Hydrologic Engineering, 11(1), 65–70. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:1(65)
Khoob, A. R. (2008). Artificial neural network estimation of reference evapotranspiration from pan evaporation in a semi-arid environment. Irrigation Science, 27(1), 35–39. https://doi.org/10.1007/s00271-008-0119-y
Kisi, O. (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology, 528, 312–320. https://doi.org/10.1016/j.jhydrol.2015.06.052
Kumar, M., Raghuwanshi, N. S., Singh, R., Wallender, W. W., & Pruitt, W. O. (2002). Estimating evapotranspiration using artificial neural network. Journal of Irrigation and Drainage Engineering, 128(4), 224–233. https://doi.org/10.1061/(ASCE)0733-9437(2002)128:4(224)
Landeras, G., Ortiz-Barredo, A., & Lopez, J. J. (2008). Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). Agricultural Water Management, 95(5), 553–565. https://doi.org/10.1016/j.agwat.2007.12.011
Makkink, G. F. (1957). Testing the Penman formula by means of lysimeters. Journal of the Institution of Water Engineers, 11, 277–288.
Matar, M. A. (2018). Using gene expression programming in monthly reference evapotranspiration modeling: A case study in Egypt. Agricultural Water Management, 189, 28–38. https://doi.org/10.1016/j.agwat.2017.12.017
Panaitescu, L., Ilie, C., Lungu, M. L., Popescu, M., Lungu, D., & Nita, S. (2014). Modern approach to the phenomenon of drought and aridity in Central and South Dobrudja. Journal of Environmental Protection and Ecology, 15(1), 110–122. http://www.jepe-journal.info/vol15-no-1-2014
Parasuraman, K., Elshorbagy, A., & Carey, S. K. (2007). Modelling the dynamics of the evapotranspiration process using genetic programming. Hydrological Sciences Journal, 52(3), 563–578. https://doi.org/10.1623/hysj.52.3.563
Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review, 100(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2
Shih, S. F. (1984). Data requirement for evapotranspiration estimation. Journal of Irrigation and Drainage Engineering, 110(3), 263–274. https://doi.org/10.1061/(ASCE)0733-9437(1984)110:3(263)
Sudheer, K. P., Gosain, A. K., & Ramasastri, K. S. (2003). Estimating actual evapotranspiration from limited climatic data using neural computing technique. Journal of Irrigation and Drainage Engineering, 129(3), 214–218. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:3(214)
Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. Geographical Review, 38(1), 55–94. https://doi.org/10.2307/210739
Trajkovic, S., Todorovic, B., & Stankovic, M. (2003). Forecasting of reference evapotranspiration by artificial neural networks. *Journal of Irrigation and Drainage Engineering, 129*(6), 454–457. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:6(454)

Turc, L. (1961). Evaluation des besoins en eau d’irrigation, évapotranspiration potentielle. *Annales Agronomiques, 12* (1), 13–49.

Wang, Y. M., Traore, S., & Kerh, T. (2008). Monitoring event-based suspended sediment concentration by artificial neural network models. *WSEAS Transactions on Computers, 5*(7), 359–368. https://doi.org/10.1109/TC.2007.70821

Zanetti, S. S., Sousa, E. F., Oliveira, V. P., Almeida, F. T., & Bernardo, S. (2007). Estimating evapotranspiration using artificial neural network and minimum climatological data. *Journal of Irrigation and Drainage Engineering, 133*(2), 83–89. https://doi.org/10.1061/(ASCE)0733-9437(2007)133:2(83)