Advanced Engineering Informatics 33 (2017) 89–95

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

# A novel approach for precipitation forecast via improved *K*-nearest neighbor algorithm

Mingming Huang [a,*], Runsheng Lin [a], Shuai Huang [b,*], Tengfei Xing [c]

[a] Beijing Meteorological Information Center, 44 Zizhu Road, Haidian District, Beijing 100089, PR China
[b] Institute of Crustal Dynamics, China Earthquake Administration, 1 Anningzhuang Road, Haidian District, Beijing 100085, PR China
[c] Tencent Beijing, 66 Zhongguncun East Road, Beijing 100080, PR China

ARTICLE INFO

*Article history:*
Received 12 December 2016
Received in revised form 5 April 2017
Accepted 8 May 2017

*Keywords:*
KNN algorithm
Improved KNN algorithm
Precipitation
Precipitation forecast

ABSTRACT

The prediction method plays crucial roles in accurate precipitation forecasts. Recently, machine learning has been widely used for forecasting precipitation, and the *K*-nearest neighbor (KNN) algorithm, one of machine learning techniques, showed good performance. In this paper, we propose an improved KNN algorithm, which offers robustness against different choices of the neighborhood size *k*, particularly in the case of the irregular class distribution of the precipitation dataset. Then, based our improved KNN algorithm, a new precipitation forecast approach is put forward. Extensive experimental results demonstrate that the effectiveness of our proposed precipitation forecast approach based on improved KNN algorithm.



## 1. Introduction

Precipitation plays crucial roles in climate because it not only is vital for agriculture, forestry and the energy industry but also provides stable habitats for great varieties of species [1–4]. Nevertheless, heavy rains in a short time usually result in natural disasters such as flash floods, mud-rock flows, urban waterlogging and landslides, which causes tremendous losses in lives and properties of people [5,6]. And for this reason, reliable precipitation forecast is highly important and essentially needed. Unfortunately, the precipitation forecast is a major challenge in meteorology due to formation mechanism of precipitation, not completely understood so far, involves a rather complex physics [7–9]. And currently the precipitation forecast is far from being satisfactory [10]. Accordingly, the interest in precipitation forecasts has grown significantly in recent years [11–14].

There exist two possible approaches to forecast precipitation in the literature [6,15].The first is the physically-based approach, which means that the underlying physical laws of precipitation are modeled by studying the rainfall processes. The other approach is the pattern recognition methodology based on the use of machine learning techniques to implement which means that precipitation patterns are attempted to recognize based on their features using machine learning techniques. However, it is a general notion that the physically-based approach may not be feasible, since significant variables to reason precipitation are interconnected in an extremely complicated way, and the volume of precipitation calculations requires sophisticated mathematical tools [16,17]. And the pattern recognition methodology based on the use of machine learning techniques to implement has been proven to be very effective in precipitation forecasting [18,19,17]. In recent years, much effort has been invested in developing precipitation forecasts using various machine learning techniques that yield good forecast.

Ramírez et al. [10] established a nonlinear mapping between meteorological variables and surface rainfall data to form an artificial neural network (ANN) model trained by the resilient propagation learning algorithm, which can generate specific quantitative forecasts of daily rainfall in the São Paulo State, Brazil. Hong [17] employed support vector machine (SVM) to implement rainfall forecasting in Northern Taiwan and the empirical results revealed that this approach yields well forecasting performance. Zeng et al. [20] proposed an improved *K*-nearest neighbor (KNN) algorithm to forecast precipitation, in which one optimal *k* is chosen according to the number of exiting weather event $k^+$ and the number of no weather event $k^-$, which are computed to match different weather events owing diverse probability based on the crossing verification method in searching *k*-nearest neighbor process. In addition, Chen et al. [7] developed a novel weighted *k*-nearest neighbor algorithm for forecasting precipitation in Nanjing city.

* Corresponding authors.
*E-mail addresses:* mmhuang1205@163.com (M. Huang), huangshuai3395@163.com (S. Huang).

http://dx.doi.org/10.1016/j.aei.2017.05.003
1474-0346/

It is worth noting that $K$-nearest neighbor algorithm [21], which is one of the most well-known algorithms in pattern recognition for supervised non-parametric classification, has been widely used in precipitation forecasts. However, there exists one major outstanding issue in the KNN algorithm that it is sensitive to the choice of the neighborhood size $k$ [22]. And the classification performance of many KNN-based methods will degrade dramatically due to the problem above, particularly in the small sample size cases and the data with an uneven distribution [23,24]. Unfortunately, historical rainfall data is imbalanced, which means that the number of training samples in some classes is much larger than that of the others. For example, the number of torrential rainfall events, as compared to the number of dry days, is significantly few because the torrential rainfall events rarely occurred in a year. As a consequence, although the KNN algorithm has been proven to be very effective in making forecast of rainfall, the forecast performance is often limited by the influence of the sensitiveness to the selection of the neighborhood $k$ when the precipitation data with an uneven distribution. To address the issue of sensitivity of different choices of $k$, there exists one kind of main works, namely distance-based vote weighting schemes for the KNN classifier [25]. Dudani [26] proposed a weighted voting method named the distance-weighted $k$-nearest neighbor (WKNN) rule, which is the first distance-based vote weighting schemes. In this approach, the farthest neighbor is weighted with 0, the closest with 1 and the others are scaled between by a linear mapping. Gou et al. [27] presented a dual weighted $k$-nearest neighbor (DWKNN) rule that extended the linear mapping of Dudani, in which the closest and the farthest neighbors are weighted the same way as the linear mapping, but those between them are assigned smaller values.

In this paper, we propose an improved KNN algorithm, which offers robustness against different choices of the neighborhood size $k$, particularly in the case of the irregular class distribution of the precipitation dataset. Then, based our improved KNN algorithm, a new precipitation forecasting model is put forward. Extensive experimental results show that the proposed precipitation forecasting paradigm achieves much better forecast performance.

The reminder of this paper is organized as follows. In Section 2, we give a brief overview of many algorithms, including KNN, WKNN and DWKNN. Section 3 presents details of our improved KNN algorithm. The construction of our proposed precipitation forecasting model is described in Section 4. Experiment results are discussed in Section 5. Finally, we conclude this paper in Section 6.

## 2. Related work

Before presenting in detail our improved KNN algorithm, we briefly review of KNN, WKNN and DWKNN that form the basis for our work.

### 2.1. KNN

The $k$-nearest neighbor algorithm is a powerful nonparametric classifier which assigns an unclassified pattern to the class represented by a majority of its $k$ nearest neighbors. In the general classification problem, let $T = \{\boldsymbol{x}_n \in R^d\}_{n=1}^N$ denote a training set with $M$ classes consisting of $N$ training samples in $d$-dimensional feature space, and the class label of one sample $\boldsymbol{x}_n$ is $c_n$. Given a query point $\boldsymbol{x}$, the KNN rule is carried out as follows.

- Find $k$ nearest neighbors from the set $T$ for the unknown query point $\boldsymbol{x}$, and let $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^k$ indicate the set of $k$ nearest neighbors for $\boldsymbol{x}$. The distance between $\boldsymbol{x}$ and the neighbor $x_i^{NN}$ is measured by the Euclidean distance metric

$$d(x, x_i^{NN}) = \sqrt{(x - x_i^{NN})^{\mathrm{T}}(x - x_i^{NN})} \tag{1}$$

- The class label of the query point $\boldsymbol{x}$ is predicted by the majority voting of its neighbors

$$c' = \arg\max_c \sum_{(x_i^{NN}, c_i^{NN}) \in \overline{T}} \delta(c = c_i^{NN}) \tag{2}$$

where $c$ is a class label and $c_i^{NN}$ denotes the class label for the $i$-th nearest neighbor among its $k$ nearest neighbors. $\delta(c = c_i^{NN})$, an indicator function, takes a value of one if the class $c_i^{NN}$ of the neighbor $x_i^{NN}$ is the same as the class $c$ and zero otherwise.

### 2.2. WKNN and DWKNN

To overcome the negative effect of the neighborhood $k$, the WKNN assigning weights to the neighbors according to their distance from the unclassified sample. Formally, the weighted function of WKNN is represented as

$$\omega_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \qquad d_i = \sqrt{(x - x_i^{NN})^T (x - x_i^{NN})} \tag{3}$$

where $x_i^{NN}$ denotes the $i$-th nearest neighbor among its $k$ nearest neighbors $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^k$ sorted in an increasing order according to their corresponding Euclidean distance to the query point $\boldsymbol{x}$.

Accordingly, the class of the query object predicted by the majority weighted voting defined as

$$c' = \arg\max_c \sum_{(x_i^{NN}, c_i^{NN}) \in \overline{T}} \omega_i \cdot \delta(c = c_i^{NN}) \tag{4}$$

To further overcome the negative effect of the neighborhood $k$, the dual distance-weighted function of DWKNN extends the linear mapping of WKNN. The dual distance-weighted function of DWKNN can be formulated as

$$\omega_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1} \cdot \frac{d_k + d_1}{d_k + d_i}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \tag{5}$$

## 3. Our improved KNN algorithm

Although WKNN and DWKNN algorithms perform well in comparisons with the traditional KNN approach, the sensitivity of the classification performance to the choices of the neighborhood size $k$ still exits, especially in the data imbalance situations. And we noticed that the exponential of some distance, which is chosen as the weighting scheme, exhibits better classification accuracy and lower variance [28]. Inspired by the effectiveness of the exponential of some distance for classification, we believe that it should be a better candidate which is chosen as the weighting scheme. In this paper, we design an improved KNN algorithm in order to further mainly conquer the influence of the neighborhood $k$ in the case of the precipitation dataset with an uneven distribution. Suppose a training set $T = \left\{\boldsymbol{x}_n \in R^d\right\}_{n=1}^N$ with $M$ classes, where $N$ is the sample numbers of $T$, and $d$ is the feature dimension. In our proposed improved KNN algorithm, the class label of a query point $\boldsymbol{x}$ is yielded by the following steps.

- Find $k$ nearest neighbors for the unknown query point $\boldsymbol{x}$ in the training set $T$, and let $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^{k}$ denote the set of $k$ nearest neighbors for $\boldsymbol{x}$, and the $k$ nearest neighbors $x_1^{NN}, x_2^{NN}, \ldots, x_k^{NN}$ are sorted in an increasing order according to their corresponding Euclidean distance to $\boldsymbol{x}$, described in Eq. (1).
- Allocate different weights to $k$ nearest neighbors, and the weight $\omega_j$ of the $j$-th nearest neighbor is determined as

$$\omega_j = \begin{cases} \exp\left(-\left(\left(\frac{d_k - d_j}{d_k - d_1}\right) \cdot \left(\frac{d_k + d_1}{d_k + d_j}\right)\right)\right), & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \tag{6}$$

- Classify the query point $\boldsymbol{x}$ into the class $\bar{c}$ by a majority weighted voting of its neighbors.

$$\bar{c} = \arg\max_{c} \sum_{\left(x_j^{NN}, c_j^{NN}\right) \in \overline{T}} \omega_j \cdot \delta\left(c = c_j^{NN}\right) \tag{7}$$

Note that our proposed improved KNN rule is equal to DWKNN, WKNN and KNN rule when $k = 1$ because the nearest neighbor gets weight of 1 when $k = 1$.

## 4. Improved KNN algorithm based precipitation forecast for Beijing area

### 4.1. Precipitation data and predictors

The area considered herein is Beijing, located in the northeast of China. There are four main seasons including spring (March-May), summer (June-August), autumn (September-November) and winter (December- February). About 75% of the annual precipitation over this area (lat: $39.26^\circ - 41.03^\circ$N, lon: $115.25^\circ - 117.30^\circ$E) occurs in summer which is from June to August. For the prediction of the level of rainfall in a calendar day, daily precipitation data from June to August, which are collected from twenty national automatic weather stations in Beijing for the period from 1990 to 2012, have been obtained from the Beijing Meteorological Information Center. Fig. 1 shows the Beijing area and the locations of twenty national automatic weather stations. According to the partition rule of grade of precipitation in flood protection departments, total rainfall amount for 24 h is divided into one of five grades of precipitation which are no rain, light rain, moderate rain, heavy rain and torrential rain respectively; while the five grades of precipitation are labeled as 0, 1, 2, 3 and 4 respectively, as depicted in Table 1.

All of the prognostic variables are chosen from the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis dataset with a spatial resolution of $2.5^\circ \times 2.5^\circ$ [29], which are commonly used to forecast precipitation by meteorologists [18,30]. In our work, six prognostic variables chosen from the NCEP-NCAR reanalysis dataset is taken as predictors, as illustrated in Table 2. And the area between $35^\circ$ and $42.5^\circ$N, and $112.5^\circ$ and $120^\circ$E (16 grid-points in the total) over the study area is taken as the selection area for predictors, i.e. the precipitation field. Thus the feature of each sample which is daily precipitation data from June to August in the year of 1990–2012 is obtained by concatenating six predictors from the precipitation field.

### 4.2. Normalization

Since the range of each predictor is significantly different and the test results might rely on the values of a few predictors, they are preprocessed using a normalization [14]. We calculated the upper and lower bound of each predictor in the precipitation field from June to August in the year of 1990–2012. The results are shown in Table 3. The process for the used normalization is represented as

$$\begin{aligned} \bar{y}_i &= \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \\ y_{\min} &= \min\{\mathbf{y}\} = \min_{j=1,\ldots,n}\{y_j\} \\ y_{\max} &= \max\{\mathbf{y}\} = \max_{j=1,\ldots,n}\{y_j\} \end{aligned} \tag{8}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is each predictor. Accordingly, the value of each predictor is normalized to between 0 and 1 based on the Eq. (8).

### 4.3. Improved KNN algorithm based precipitation forecast

As a classification algorithm, our improved KNN algorithm could be more robust to the change of $k$ in the case of the precipitation dataset with an uneven distribution, and tends to yield more reliable forecast performance. Therefore, we proposed to use our improved KNN algorithm to make precipitation forecast. Let $\boldsymbol{X}$ denote a set of precipitation sample, and suppose $\boldsymbol{X}$ is $X = \left\{x_n \in R^d\right\}_{n=1}^{N}$, where $x_i$ represents the feature of the $i$-th precipitation sample, $N$ is the total number of features, and $d$ is the feature dimension. Let $c_i \in \{0, 1, 2, 3, 4\}, i = 1, 2, \ldots, N$ be the grade of precipitation, which is the class attribute of each precipitation sample $x_i = (x_{i1}, x_{i2}, \ldots, x_{id}), i = 1, 2, \ldots, N$. Accordingly, the training set can be formulated as follows

$$\begin{bmatrix} x_1 & c_1 \\ x_2 & c_2 \\ \vdots & \vdots \\ x_N & c_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & c_1 \\ x_{21} & x_{22} & \cdots & x_{2d} & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} & c_N \end{bmatrix} \tag{9}$$

Given the unknown precipitation sample $\boldsymbol{x}$, we find $k$ nearest neighbors of $\boldsymbol{x}$, $\overline{T} = \{(x_i^{NN}, c_i^{NN})\}_{i=1}^{k}$, in the training set as shown in Eq. (9) based on our improved KNN algorithm described in Section 3. Then we sort the $k$ nearest neighbors in an increasing order according to their corresponding Euclidean distance with $\boldsymbol{x}$, $d_1, d_2, \ldots, d_k$. Hence the unknown precipitation sample $\boldsymbol{x}$ is classified into the class $\bar{c}$ by the precipitation forecast model which is given in Eqs. (6) and (7), as depicted in Section 3.

In our precipitation forecast model, we use daily precipitation data from June to August in the year of 1990–2008 and those in the year of 2009–2012 as training samples and testing samples respectively. Thus the training dataset contains 1748 samples and testing set is 368.

## 5. Experiments and results

In this section, we will investigate our proposed precipitation forecast approach based on improved KNN algorithm. Obviously, the only difference in our precipitation forecast experiments is the classification algorithm in our proposed precipitation forecast model. So the final forecast performance only relies on the distinctiveness of the classification algorithm. The following forecast experiments will show whether our proposed improved KNN algorithm will achieve better forecast performance. For the purpose of comparison our proposed approach, we have also built other three precipitation forecast approach including precipitation forecast model based on KNN algorithm [21], precipitation forecast model based on WKNN algorithm [26] and precipitation forecast model based on DWKNN algorithm [27].

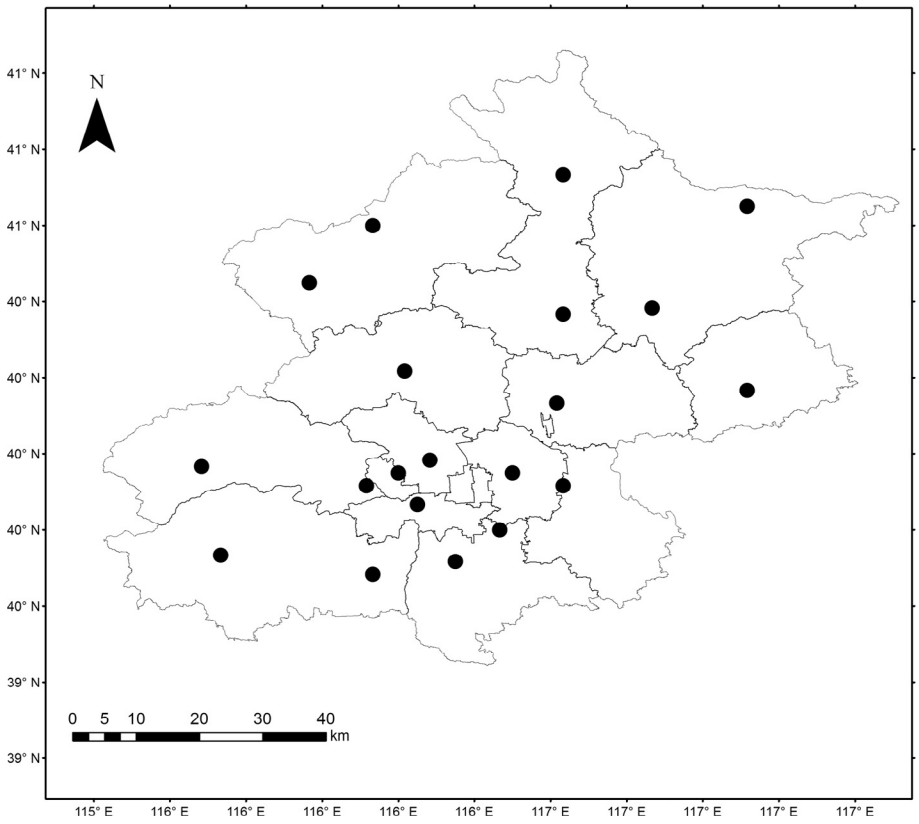


**Fig. 1.** Map of Beijing with positions of the stations (•).

**Table 1**
The grade of precipitation divided by flood protection departments ($R$ denotes total rainfall amount for 24 h in mm).

| Precipitation grade | $R$ (mm) | Label of precipitation grade |
|---|---|---|
| No rain | $R < 0.1$ | 0 |
| Light rain | $0.1 \leqslant R < 10$ | 1 |
| Moderate rain | $10 \leqslant R < 25$ | 2 |
| Heavy rain | $25 \leqslant R < 50$ | 3 |
| Torrential rain | $R \geqslant 50$ | 4 |

**Table 2**
The names and descriptions of the prognostic variables.

| Prognostic variables | Description |
|---|---|
| hgt500 | 500hpa geopotential height |
| hgt1000 | 1000hpa geopotential height |
| air850 | 850hpa air temperature |
| rhum850 | 850hpa humidity |
| vwnd700 | 700hpa meridional wind |
| pr_wtr | precipitable water |

**Table 3**
The upper and lower bound ranges of predictors.

| Predictors | Unit | Upper bound | Lower bound |
|---|---|---|---|
| 500hpa geopotential height | m | 5946.15 | 5404.40 |
| 1000hpa geopotential height | m | 185.13 | −109.33 |
| 850hpa air temperature | degK | 307.15 | 275.03 |
| 850hpa humidity | % | 102.12 | 3.000001 |
| 700hpa meridional wind | m/s | 25.3 | 0.000002 |
| precipitable water | kg/m$^2$ | 76.35 | 1.300001 |

### 5.1. Criteria for evaluating precipitation forecast performance

Five different types of standard evaluation criteria, which involves accuracy, threat score (TS), summary alarm rate (SAR), missing alarm rate (MAR) and precision-recall (PR) curve, were employed to evaluate the performance of various precipitation forecast approaches developed in this paper. The accuracy, computed based on the percentage of all test samples classified correctly, is used to evaluate the prediction performance of different types of precipitation described in Section 4.1. Accuracy tells us about the number of samples which are correctly forecasted. Furthermore, the prediction performance of two different types of weather events, rain and no-rain, is evaluated by four measures including TS, SAR, MAR and PR curve. The measure for evaluating precipitation forecast performance, accuracy, threat score (TS), summary alarm rate (SAR), missing alarm rate (MAR) and precision-recall (PR) curve, are defined as follows

$$\text{Accuracy} = \frac{\#\text{test samples forecasted correctly}}{\#\text{test samples}} \tag{10}$$

$$\text{TS} = \frac{\text{hit alarms}}{\text{hit alarms} + \text{missing alarms} + \text{false alarms}} \tag{11}$$

$$\text{SAR} = \frac{\text{hit alarms}}{\#\text{test samples}} \tag{12}$$

$$\text{MAR} = \frac{\text{missing alarms}}{\text{missing alarms} + \text{hit alarms}} \tag{13}$$

$$\text{Precision} = \frac{\text{hit alarms}}{\text{hit alarms} + \text{false alarms}} \tag{14}$$

**Table 4**
Contingency table.

| Event forecasted | Event observed | |
|---|---|---|
| | Rain(yes) | No-rain(no) |
| Rain(yes) | Hit alarms | False alarms |
| No-rain(no) | Missing alarms | Correct alarms |

$$\text{Recall} = \frac{\text{hit alarms}}{\text{hit alarms} + \text{missing alarms}} \tag{15}$$

where the #test samples denotes the total number of test samples, the #test samples forecasted correctly is the number of test samples which is forecasted correctly, and the hit alarms, missing alarms and false alarms are computed in Table 4.

### 5.2. Precipitation forecast

We evaluate our proposed precipitation forecast approach based on improved KNN algorithm in two different applications: grade forecast and rain and no-rain forecast. The first is grade forecast, which means that we make a prediction about different grades of precipitation. The other is rain and no-rain forecast, which means that we make a prediction about two different types of weather events which are rain and no-rain. The proposed approach is compared with the other three popular methods using DWKNN, WKNN and KNN algorithm respectively. In fact, the code of DWKNN, WKNN and KNN algorithm in the other three popular precipitation forecast methods are not given directly from the literature. Therefore, we implemented DWKNN, WKNN and KNN algorithm by ourselves based on the literature. Moreover, we also compare our proposed approach to the state-of-the-art precipitation forecast approach using SVM algorithm which is with different kernels, including linear kernel and radial basis function kernel, in grade forecast experiments. The forecast approaches using DWKNN, WKNN, KNN and SVM algorithm are abbreviated to DWKNN, WKNN, KNN and SVM.

#### 5.2.1. Grade forecast evaluation

In this section, we conducted extensive experiments on grade forecast to evaluate the performance of our proposed approach. In our grade forecast experiments, we use daily precipitation data from June to August in the year of 1990–2008 and those in the year of 2009–2012 as training samples and testing samples respectively. Thus the training dataset contains 1748 samples and testing set is 368. We are interested to see how the performance changes if we modify the value of the neighborhood size $k$. Inspired by Ref. [24] and taking into consideration the number of training samples in which torrential rain events, one of five types of precipitation, rarely occurred in a year, the parameter $k$ is ranged from 1 to 15 with an interval of 1.

The experimental results are shown in Table 5. It can be found that the accuracy of our proposed precipitation forecast approach based on improved KNN algorithm is somewhat better than those of DWKNN, WKNN and KNN in almost all the test cases. That is to say, our proposed improved KNN algorithm almost gives an improvement over the other three approaches with the increase of the neighborhood size $k$. Consequently, it suggests that the proposed improved KNN algorithm has the robustness to the sensitivity of different choices of the neighborhood size $k$ with the satisfactory forecast performance to some degree.

Furthermore, we also conducted experiments on grade forecast to compare our proposed precipitation forecast approach based on improved KNN algorithm with precipitation forecast approach based on SVM algorithm. In this experiment, the neighborhood size $k$ is fixed as 15. Detailed comparison results are shown in Table 6.

**Table 5**
Precipitation forecast results with different neighborhood sizes.

| $k$ | Accuracy (%) | | | |
|---|---|---|---|---|
| | Our proposed | DWKNN | WKNN | KNN |
| 1 | **42.12** | 42.12 | 42.12 | 42.12 |
| 2 | **42.39** | 42.12 | 42.12 | 39.95 |
| 3 | **41.85** | 42.12 | 42.12 | 40.22 |
| 4 | **43.21** | 42.66 | 42.93 | 42.93 |
| 5 | **45.11** | 42.39 | 42.12 | 42.93 |
| 6 | **47.01** | 41.85 | 41.85 | 46.19 |
| 7 | **45.11** | 42.66 | 42.39 | 45.11 |
| 8 | **45.65** | 42.93 | 42.93 | 43.48 |
| 9 | **46.47** | 42.66 | 43.21 | 44.57 |
| 10 | **46.74** | 43.21 | 44.29 | 44.29 |
| 11 | **47.01** | 44.02 | 44.84 | 45.38 |
| 12 | **46.74** | 44.29 | 45.38 | 45.11 |
| 13 | **48.64** | 45.11 | 45.65 | 47.28 |
| 14 | **48.36** | 45.92 | 45.92 | 45.38 |
| 15 | **49.46** | 45.65 | 45.92 | 48.37 |

**Table 6**
Grade forecast results comparison of different approaches.

| Algorithm | Accuracy (%) |
|---|---|
| Our proposed | 49.46 |
| SVM | |
| Linear | 49.18 |
| Radial basis function | 49.45 |

As can be seen in Table 6, our proposed precipitation forecast approach based on improved KNN algorithm obtains the performance of 49.46%; while the precipitation forecast approach based on SVM with linear kernel and radial basis function kernel achieve a forecast accuracy of 49.18% and 49.45%, respectively. It can be found that the performance of our proposed precipitation forecast approach is somewhat better than SVM with linear kernel and SVM with radial basis function kernel.

To demonstrate intuitively the effectiveness of our proposed method, we make a comparison between observed and predicted rainfall grade at 31 samples, which are daily precipitation data on July 2010 for testing samples, by four different forecast approaches. Meantime, we choose daily precipitation data from June to August in the year of 1990–2008 as training samples. Detailed comparison results are shown in Fig. 2. From the experimental results illustrated in Fig. 2, we can see that the output of our proposed forecast approach, simulated with testing data, shows a good agreement with the target. We can observe that our proposed precipitation forecast approach based on improved KNN algorithm has better results than the other three approaches including DWKNN, WKNN and KNN in accuracy.

#### 5.2.2. Rain and no-rain forecast evaluation

To further evaluate the performance of our proposed precipitation forecast approach based on improved KNN algorithm, we also conduct experiments to see the prediction performance for two different types of weather events including rain and no-rain. The threat score, summary alarm rate, missing alarm rate and precision-recall curve are used as the evaluation criterion. And the experimental comparisons in terms of threat score, summary alarm rate, missing alarm rate and precision-recall curve with varying the neighborhood size $k$ are illustrated in Figs. 3–6. From the results described in Fig. 3, we can obviously observe that the threat score outperforms the other three methods for almost all values of the neighborhood size $k$. Taken one value of the neighborhood size $k$ as an example, our proposed approach achieves a threat score of 50.46% when the value of the neighborhood size $k$ is equal to 6. However, the other three algorithms, DWKNN, WKNN

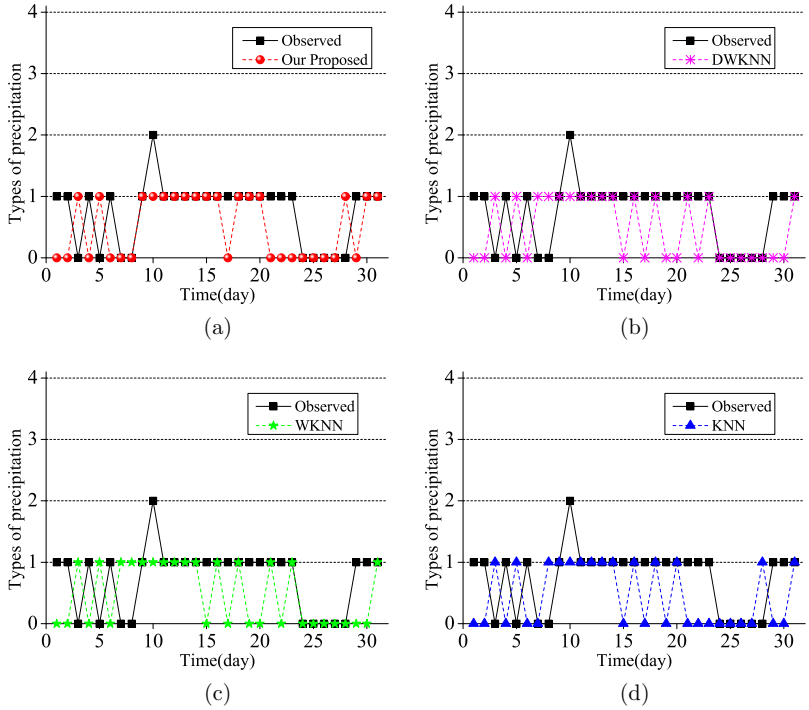


**Fig. 2.** Comparison between observed and predicted of forecast approaches in testing samples.

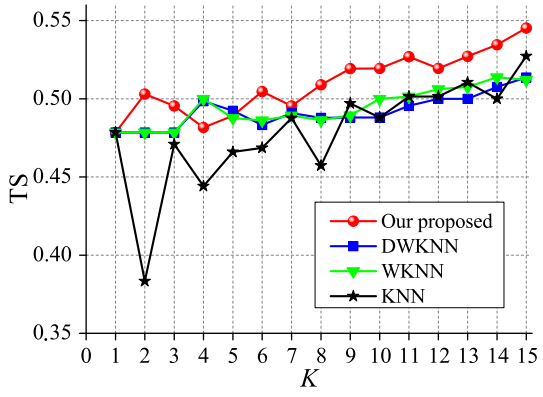


**Fig. 3.** The threat score of different methods under various values of the neighborhood size.

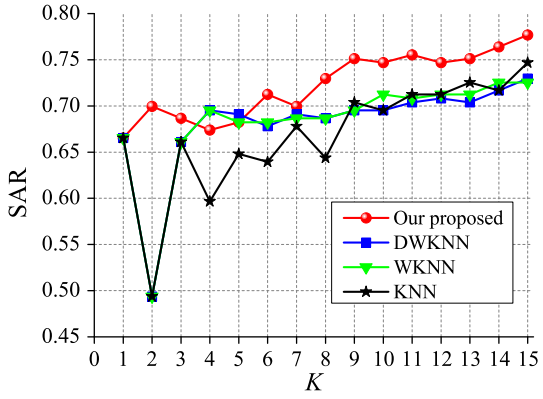


**Fig. 4.** The summary alarm rate of different methods under various values of the neighborhood size.

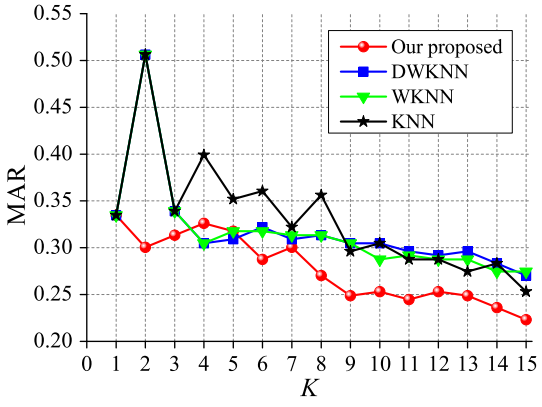


**Fig. 5.** The missing alarm rate of different methods under various values of the neighborhood size.

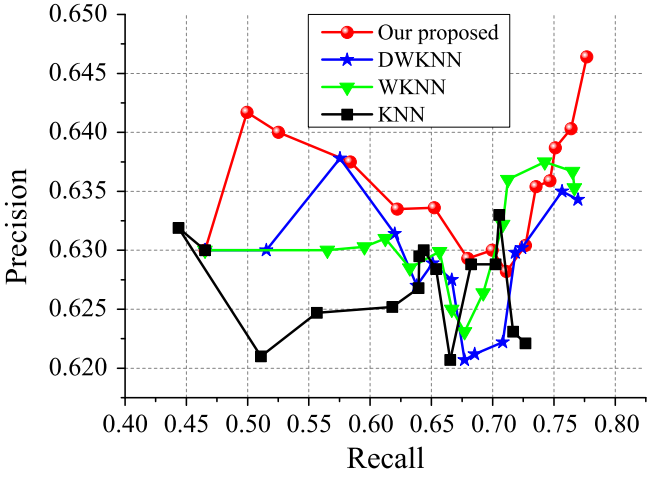


**Fig. 6.** The precision-recall curve of different methods under various values of the neighborhood size.

and KNN, only obtain a threat score of 48.32%, 48.62% and 46.88% respectively.

As shown in Fig. 4, our proposed precipitation forecast approach based on improved KNN algorithm almost achieves the best performance among the interval of the neighborhood size $k$, compared to the other three competing methods; while the proposed approach always performs significantly better than KNN. And the summary alarm rate of our proposed algorithm increases as the neighborhood size grows further.

Fig. 5 has shown the missing alarm rate of our proposed precipitation forecast approach based on improved KNN algorithm. It should be noted that the missing alarm rate is extremely crucial in precipitation forecast. And the smaller the missing alarm rate is, the better the forecast performance of our proposed approach is. As can be seen in Fig. 5, our proposed approach almost always performs significantly better than DWKNN, WKNN and KNN. Furthermore, it is noticeable that our proposed precipitation forecast approach based on improved KNN algorithm always outperforms KNN. Taken one value of the neighborhood size $k$ as an example, our proposed approach obtains a missing alarm rate of 30.04% when the neighborhood size is fixed as 2; while the other three algorithms, DWKNN, WKNN and KNN, only obtain a missing alarm rate of 50.64%, 50.64% and 50.64% respectively.

Fig. 6 has shown the precision-recall curve of our proposed precipitation forecast approach based on improved KNN algorithm. The precision-recall curve can be obtained by changing the neighborhood size $k$; and in PR space the goal is to be in the upper-right-hand corner [31]. As shown in Fig. 6, we can observe that the performance of our proposed precipitation forecast approach based on improved KNN algorithm in rain and no-rain forecast is somewhat better than the other three approaches including DWKNN, WKNN and KNN.

## 6. Conclusions

In this article, an improved KNN algorithm has first been proposed. It offers robustness against different choices of the neighborhood size $k$ particularly in the case of the precipitation dataset with an uneven distribution. Then based on the improved KNN algorithm, we introduce a new precipitation forecast approach. Extensive experimental results for grade forecast and rain and no-rain forecast demonstrate that the effectiveness of our proposed precipitation forecast approach based on improved KNN algorithm. In our future work, we are willing to design selecting methods of predictors to improve the performance of precipitation forecast.

## Acknowledgments

This article is supported by China Meteorological Administration (Grant No. CMAHX20160701), Beijing talents Fund (Grant No. 2015000057592G270), Beijing Natural Science Foundation (Grant No. 8174078) and research grant from Institute of Crustal Dynamics, China Earthquake Administration (Grant No. ZDJ2016-12). The authors would like to express their sincere appreciation to the anonymous reviewers for their insightful comments, which greatly helped them to improve the quality of the paper.

## References

[1] P.M. Jermey, R.J. Renshaw, Precipitation representation over a two year period in regional reanalysis, Quart. J. Roy. Meteorol. Soc. (2016).
[2] L. Dubus, Monthly and seasonal forecasts in the French power sector, 2012.
[3] I. Ross, L. Misson, S. Rambal, A. Arneth, R.L. Scott, A. Carrara, A. Cescatti, L. Genesio, How do more extreme rainfall regimes affect ecosystem fluxes in seasonally water-limited northern hemisphere temperate shrublands and forests?, Biogeosci Discuss. 8 (5) (2011) 9813–9845.
[4] B.B. Balana, A. Vinten, B. Slee, A review on cost-effectiveness analysis of agri-environmental measures related to the eu wfd: key issues, methods, and applications, Ecol. Econ. 70 (6) (2011) 1021–1031.
[5] X.X. Qiu, F.Q. Zhang, Prediction and predictability of a catastrophic local extreme precipitation event through cloud-resolving ensemble analysis and forecasting with doppler radar observations, Sci. China Earth Sci. 59 (3) (2016) 518–532.
[6] K.C. Luk, J.E. Ball, A. Sharma, An application of artificial neural networks for rainfall forecasting, Math. Comput. Modell. 33 (6-7) (2001) 683–693.
[7] K. Chen, L.s. Wang, Novel weighted nearest neighbor algorithm of precipitation forecast experiment, Comput. Simul. 31 (6) (2014) 325–328.
[8] H.X. Li, C.W. Li, Construction and application of fuzzy neural network model in precipitation forecast of Sanjiang Plain, China, 2008, pp. 1–3.
[9] D.J. Gagne Ii, A. Mcgovern, M. Xue, Machine learning enhancement of storm scale ensemble precipitation forecasts, in: Intelligent Data Understanding, 2011, pp. 39–46.
[10] M.C.V. Ramírez, H.F.D.C. Velho, N.J. Ferreira, Artificial neural network technique for rainfall forecasting applied to the sao paulo region, J. Hydrol. 301 (1) (2005) 146–162.
[11] B.T. Zavodsky, S.H. Chou, G.J. Jedlovec, Improved regional analyses and heavy precipitation forecasts with assimilation of atmospheric infrared sounder retrieved thermodynamic profiles, IEEE Trans. Geosci. Remote Sensing 50 (50) (2012) 4243–4251.
[12] J.I. Yano, B. Jakubiak, Wavelet-based verification of the quantitative precipitation forecast, Dynam. Atmos. Oceans 74 (2016) 14–29.
[13] Y. Huang, X. Cui, X. Li, A three-dimensional wrf-based precipitation equation and its application in the analysis of roles of surface evaporation in a torrential rainfall event, Atmos. Res. 169 (123) (2016) 54–64.
[14] J.H. Seo, H.L. Yong, Y.H. Kim, Feature selection for very short-term heavy rainfall prediction using evolutionary computation, Adv. Meteorol. 2014 (1C4) (2014) 1–15.
[15] Christopher, Allen, Ajay, KALRA, Sajjad, AHMAD, Long-range precipitation forecasts using paleoclimate reconstructions in the western united states, J. Mountain Sci. 13 (4) (2016) 614–632.
[16] L. Jiang, J. Wu, Hybrid pso and ga for neural network evolutionary in monthly rainfall forecasting, in: Asian Conference on Intelligent Information and Database Systems, 2013, pp. 79–88.
[17] W.C. Hong, Rainfall forecasting by technological machine learning models, Appl. Math. Comput. 200 (1) (2008) 41–57.
[18] Y. Di, W. Ding, Y. Mu, S. Small, David L.and Islam, N.-B. Chang, Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data, 2015, pp. 63–68.
[19] Z. Beheshti, M. Firouzi, S.M. Shamsuddin, M. Zibarzani, Z. Yusop, A new rainfall forecasting model using the capso algorithm and an artificial neural network, Neural Comput. Appl. (2015) 1–15.
[20] Z.X.S.M.W.S.L.H.A.S. School, L. University, Lanzhou, Forecasting precipitation experiment with knn based on crossing verification technology, J. Appl. Meteorol. Sci. 19 (4) (2008) 471–478.
[21] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.
[22] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. Mclachlan, A. Ng, B. Liu, P.S. Yu, Top 10 algorithms in data mining, Knowl. Inform. Syst. 14 (1) (2008) 1–37.
[23] P. Pudil, P. Somol, M. Haindl, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[24] J. Gou, Y. Zhan, Y. Rao, X. Shen, X. Wang, W. He, Improved pseudo nearest neighbor classification, Knowl.-Based Syst. 70 (C) (2014) 361–375.
[25] Z. Geler, V. Kurbalija, M. Radovanović, M. Ivanović, Comparison of different weighting schemes for the knn classifier on time-series data, Knowl. Inform. Syst. 48 (2) (2015) 1–48.
[26] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, IEEE Trans. Syst. Man Cybernet. SMC-6 (4) (1976) 325–327.
[27] J. Gou, L. Du, Y. Zhang, T. Xiong, A new distance-weighted k-nearest neighbor classifier, J. Inform. Comput. Sci. 9 (6) (2012).
[28] N. Li, H. Kong, Y. Ma, G. Gong, W. Huai, Human performance modeling for manufacturing based on an improved knn algorithm, Int. J. Adv. Manuf. Technol. 84 (1) (2016) 473–483.
[29] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, The ncep/ncar 40-year reanalysis project, Bull. Am. Meteorol. Soc. 77 (1996) (1995) 437–472.
[30] A. Shirvani, W.A. Lman, Seasonal precipitation forecast skill over iran, Int. J. Climatol. 241 (4865) (2015) 599–601.
[31] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: International Conference on Machine Learning, 2006, pp. 233–240.