

Research Article

A Quality Control Method Based on an Improved Kernel Regression Algorithm for Surface Air Temperature Observations

Xiaoling Ye ^{1,2}, Yajin Kan ¹, Xiong Xiong ¹, Yingchao Zhang,² and Xin Chen¹

¹School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China

²Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China

Correspondence should be addressed to Yajin Kan; kanyajin@126.com

Received 12 November 2019; Accepted 14 January 2020; Published 13 February 2020

Academic Editor: Giacomo Gerosa

Copyright © 2020 Xiaoling Ye et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An improved kernel regression (IKR) method based on an adaptive algorithm and particle swarm optimization is proposed. Considering the limitations of current quality control methods in different regions and on multiple time scales, the kernel regression algorithm is applied to the quality control of surface air temperature observations. Observations of 12 reference stations in Jiangsu from 1961 to 2008 and of 14 regions in China from 2010 to 2014 were selected. The analysis of surface air temperature observations was performed in terms of the mean absolute error (MAE), root mean square error (RMSE), consistency indicator (IOA), and Nash–Sutcliffe model efficiency coefficient (NSC). The results indicate that compared with the traditional IDW and SRT methods, the IKR method has a high error detection rate. Furthermore, the IKR method achieves better predictions and fitting in the single-station and multistation regression experiments in Jiangsu and in the national multistation regression prediction experiment.

1. Introduction

Compared with radiosonde stations, surface observation stations have a higher spatial resolution. A high-resolution model is sensitive to high-resolution initial fields. However, surface observations, except pressure, have not yet been assimilated in the numerical weather prediction system. The study shows that the forecast skill of numerical weather prediction will be improved if surface observations are assimilated properly. 2 m temperature observations (surface air temperature observations) have more noticeable impact on the model forecast, compared with other elements [1]. Surface air temperature is measured in surface observation stations at 2.0 m above the ground and represents the important energy interaction between the earth's surface and the atmospheric layer, and between the soil surface and underlayers [2, 3].

The goals of the quality control (QC) of surface air temperature observations are to review the observations, identify missing data and suspect data, and supplement and correct the observation data to ensure the maximum extent

of archived data [4]. Surface meteorological observations are the basic element in meteorological research and have important decision-making significance for data assimilation technology and numerical weather prediction technology [5]. The accuracy of numerical weather prediction (NWP), a key meteorological forecasting technology in the current era, is largely restricted by data assimilation technology, and the QC of surface air temperature observations in the process of data assimilation is the basis of research in this area. With the rapid improvement of the current social economy, the distribution of surface weather stations is becoming more systematic and refined, leading to geometric growth in the number of meteorological observations; thus, the QC of surface meteorological observations is a basic and necessary task of meteorological research experiments [6]. Therefore, the QC of surface air temperature observations is the basis of China's meteorological industry: only by ensuring the accuracy and rationality of surface air temperature observations, we can further complete data assimilation and improve the ability of NWP [7–11].

In general, foreign scholars' QC methods for surface air temperature observations follow single-station to multi-station approaches to finally form comprehensive QC methods. Common QC methods for single-station surface meteorological observations include internal consistency tests, extreme value tests, time consistency tests, limit value tests, and time-varying tests [12–14]. These methods can identify large errors in the observations and pave the way for subsequent high-precision QC methods. The spatial consistency test is a representative QC method for surface air temperature observations from multiple stations. The test overcomes the problem of poor detection of reasonable changes in observations, and the effect of regional QC methods is obviously better than that of the QC methods used for single stations. Common spatial consistency test methods include inverse distance weighting (IDW), polynomial interpolation, and the spatial regression test (SRT) [15–17]. Among these methods, IDW and SRT are widely used in foreign countries. The Euclidean distance between stations is typically used in IDW to estimate the weights of neighbouring stations, whereas the SRT mainly uses the RMSE to predict neighbouring stations. The core idea of the above spatial consistency test methods is to estimate the observations of the target station based on the observations of neighbouring stations to achieve the purpose of QC by judging whether the observations are acceptable or need to be corrected based on the difference between the estimated observations and the observed observations [18].

Domestic research on the QC of surface air temperature observations started late, so most scholars study and discuss the corresponding domestic situation on the basis of foreign research. QC methods for surface air temperature observations have been developed steadily in China due to the completion of a three-level QC system, which has improved the QC of surface air temperature observations [4]. Domestic scholars continue to promote the development of QC methods. For example, Ren et al. conducted comprehensive QC and analysis of observations of extreme anomalies, thus preventing correct extreme meteorological observations from being completely eliminated and further improving the QC effect [19]. Li et al. proposed a QC method based on the blackboard model and used the ontology analysis method to diagnose and analyse meteorological observations via co-operation between heterogeneous knowledge bodies [20]. Wang et al. conducted QC research on meteorological observations based on comprehensive time and spatial consistency and showed that the integrated consistency method can detect erroneous observations more effectively than a single method [21]. Ye et al. proposed a method based on intelligent equation fitting to construct QC equations to achieve the QC of surface air temperature observations. Furthermore, a QC method based on improved random forest had been proposed, which effectively improved the QC effect of surface air temperature observations [22, 23]. Zhang et al. considered the continuity and stability of surface air temperature observations in short time series and proposed an integrated learning method based on particle swarm optimization for phase space reconstruction and extreme learning machines [24]. Xu et al. proposed a new

QC method to analyse the observations of the new surface automatic weather station, which lacks historical observations. According to regional and climatic characteristics, the national ground observatory is divided into eight areas for QC methods [25]. Xiong et al. proposed a QC method of surface air temperature observations based on the difference of spatial observations [26].

All the above methods can be used for QC, but the existing models of surface air temperature observations have a single structure and fail to achieve high universality in the face of surface air temperature observations in different regions or different scales. Aiming to overcome this deficiency, this paper proposes an IKR method to achieve the QC of surface air temperature observations: the test results are compared with the QC effects of traditional methods to verify the superiority of the method. The specific objectives are to (1) use surface air temperature observations from different time scales in Jiangsu and China to test the feasibility of the KR algorithm, (2) use surface air temperature observations of 14 regions in China to test the superiority of the IKR algorithm compared with the KR algorithm, and (3) test the QC effects of the IKR and KR algorithms and the traditional IDW and SRT algorithms using Jiangsu and national surface air temperature observations, including the error detection rate, predictions, and fitting.

All the above methods lack physics basis, and multiple resource information [2, 3] has to be introduced for the better quality control (QC) of surface air temperature in the future. The AE (e.g., generalization error) in the studies [2, 3] can overcome deficiency of all the above methods and achieve high universality of surface air temperature observations in different regions or different scales or different topography and/or ground cover. However, the operation of those algorithms is not easy, and a powerful tool needs to be developed.

2. Data

The observations used in this paper are from the National Meteorological Center and include six-hour surface air temperature observations (02:00, 08:00, 14:00, and 20:00) from 1951 to 2009 in 75 districts of 13 cities in Jiangsu Province and surface daily average air temperature observations of all stations in 14 national regions from 2010 to 2014. All these observations passed traditional QC, and the obvious coarse error was eliminated. After screening, observations of 12 cities in Jiangsu Province from 1961 to 2008 were selected: Xuzhou (XZ), Suqian (SQ), Lianyungang (LYG), Huai'an (HA), Yangzhou (YZ), Nanjing (NJ), Zhenjiang (ZJ), Changzhou (CZ), Yancheng (YC), Wuxi (WX), Suzhou (SZ), and Nantong (NT). Data preprocessing was performed on the six-hour surface air temperature observations and monthly average observations, taking 12~2 months' observations as the average air temperature in winter, 3~5 months' observations as the spring average air temperature, 6~8 months' observations as the average air temperature in summer, and 9~11 months' observations as the average air temperature in autumn. The 14 regions selected from the national area are Beihai (BH), Chengdu

(CD), Guangzhou (GZ), Haikou (HK), and Hohhot (HHHT), Jinghong (JH), Lhasa (LS), Lanzhou (LZ), Miyun (MY), Mohe (MH), Nanjing (NJ), Taiyuan (TY), Urumqi (WLMQ), and Changchun (CC).

To test the performance of the QC method, we refer to Hubbard's multistation QC method and randomly generate error values in the original data to simulate possible erroneous observations [27]. The error values K_λ are generated by

$$K_\lambda = s_\lambda p_\lambda, \quad (1)$$

where p_λ is a uniformly distributed random number in the interval $[-3.5, 3.5]$, s_λ is the standard deviation of the original observations, λ is the position of the error, and K_λ is the original data.

3. Methods

3.1. Traditional Methods

3.1.1. Inverse Distance Weighting. IDW, one of the most commonly used spatial interpolation methods [28], was proposed by the US National Weather Service to interpolate the distance between a point to be inserted and the actual observed sample point. The closer the sample points are, the greater the weights are; that is, the weight contribution is inversely proportional to the distance. The calculation formula is

$$Z = \sum_{i=1}^m \lambda_i Z_i, \quad (2)$$

$$\lambda_i = \frac{d_i^{-n}}{\sum_{i=1}^m d_i^{-n}},$$

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2},$$

where Z is the estimated value of the point to be interpolated, Z_i is the measured value of the i -th sample point, d_i is the distance between the i -th sample point and the point to be inserted, m is the measured sample point of the participating calculation number, and n is a power exponent that controls the degree of decrease in the weight coefficient as the distance between the point to be inserted and the sample point increases. When n is larger, closer sample points have higher weights; when n is smaller, the weights are more evenly distributed to each sample point.

3.1.2. Spatial Regression. The SRT is a QC method that checks whether a variable is within the confidence interval formed by the surrounding station data in a period of length N . All stations (M) within a certain distance are selected, each station is paired with the relevant station, and linear regression is performed. For each surrounding station, an estimate is obtained based on the regression formula $x_i = a_i + b_i y_i$. The weighted estimate x' is then obtained based on the estimated standard errors, also known as the error in the process:

$$x' = \pm \left(\sum_{i=1}^N \pm x_i^2 s_i^{-2} \right)^{0.5} \left(\sum_{i=1}^N s_i^{-2} \right)^{-0.5}, \quad (3)$$

where N is the number of stations used in the estimation, which is selected by the user ($N \leq M$), considering that the observation system may change over time and that the data of the surrounding stations change within a day. The formula for the estimated weighted standard deviation s' is

$$\frac{1}{s'^2} N^{-1} \sum_{i=1}^N s_i^{-2}. \quad (4)$$

According to the above formula, the neighbouring station is selected according to the minimum value of the standard error between each station and the target station.

3.2. The IKR Method

3.2.1. Kernel Regression Algorithm. Kernel regression estimation is an important and commonly used method in nonparametric regression. Due to the particularity of air temperature observations and the multistation QC model that this paper considers, targeted improvements must be made to the proposed method.

Assuming X and Y for a given temperature sample observation $\{(X_i, Y_j)\}_{i=1}^n$, there is a regression model given by

$$Y_i = g(X_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (5)$$

where $g(X_i)$ is the regression function and ε_i is a sequence of random error variables that are independent of each other, which has a mean of 0 and a variance of σ^2 . The regression function $g(X_i)$ is the conditional expectation given $X = x$, namely,

$$g(x) = E(Y | X = x) = \int y f(y | x) dy = \int \frac{y f(x, y)}{f_X(x)} dy, \quad (6)$$

where $f(x, y)$ is the joint density function of (X, Y) and $f_X(x)$ is the edge density function of X .

The kernel estimator of $f(x)$ is defined by the formula according to the kernel density estimation algorithm as $\hat{f}_X(x) = (1/nh) \sum_{i=1}^n k((x - X_i)/h)$, the kernel estimator of $f(x, y)$ is $\hat{f}(x, y) = (1/nh_0h) \sum_{i=1}^n k((x - X_i)/h)k((y - Y_i)/h_0)$, and h and h_0 are the window widths of X and Y . The estimator $\hat{g}(x) = \int (y \hat{f}(x, y) / \hat{f}_X(x)) dy$ of the regression function can then be written in the form of a nonparametric kernel estimator as

$$\hat{g}_{NW}(x) = \frac{\sum_{i=1}^n k((x - X_i)/h) Y_i}{\sum_{i=1}^n k((x - X_i)/h)} = \sum_{i=1}^n w_i Y_i. \quad (7)$$

From formula (7), it is seen that if there is no specific functional relationship between the response variable Y and the explanatory variable X , then there is no relationship between the data types of surface meteorological observations such as temperature, humidity, rainfall, and wind speed. This process involves the study of only surface air

temperature observations, and the feasibility of the algorithm requires further verification. The following is an explanation of formula (7):

$$\hat{g}_{\text{NW}}(x) = \frac{\sum_{i=1}^n k((x - X_i)/h)Y_i}{\sum_{i=1}^n k((x - X_i)/h)} = \frac{k((x - X_1)/h)Y_1 + k((x - X_2)/h)Y_2 + \dots + k((x - X_n)/h)Y_n}{\sum_{i=1}^n k((x - X_i)/h)} = \sum_{i=1}^n w_i Y_i. \quad (8)$$

Formula (8) can be interpreted in the form of matrix

multiplication as $\sum_{i=1}^n w_i Y_i = [w_1, w_2, \dots, w_n] \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, where

$w_i = (k((x - X_i)/h))/(\sum_{i=1}^n k((x - X_i)/h))$. The value of $\sum_{i=1}^n k((x - X_i)/h)$ is related to the explanatory variable X itself and how x is selected in $\hat{g}_{\text{NW}}(x)$. Then, for formula (5), surface air temperature observations are introduced for interpretation. If station 1 is the estimated station, its data

form is $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$. Station 2 provides the explanatory variables, and its data form is $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$. Then, formula (5) can

be modified to

$$Y_i = g(X_i) + \varepsilon_i = \hat{g}_{\text{NW}}(x)_j + \varepsilon_i$$

$$= \sum_{i=1}^n w_{ij} Y_i + \varepsilon_i = [w_1, w_2, \dots, w_n]_j \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (9)$$

Therefore, each X_i has a corresponding weight coefficient matrix $[w_1, w_2, \dots, w_n]_j$, and the standard format of formula (9) can be modified as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g(X_1) \\ g(X_2) \\ \vdots \\ g(X_n) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \hat{g}_{\text{NW}}(x)_1 \\ \hat{g}_{\text{NW}}(x)_2 \\ \vdots \\ \hat{g}_{\text{NW}}(x)_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \sum_{i=1}^n w_{ji} Y_i + \varepsilon_i = \begin{bmatrix} w_{11}, w_{12}, \dots, w_{1n} \\ w_{21}, w_{22}, \dots, w_{2n} \\ \vdots \\ w_{n1}, w_{n2}, \dots, w_{nn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (10)$$

The above is the interpretation of the calculation process of the regression estimator of surface air temperature observations of the central site for a neighbouring site.

3.2.2. Multistation Kernel Regression Algorithm. The formula for kernel regression based on multiple stations to a central station is

$$Y_i^k = g(X_i^k) + \varepsilon_i = \hat{g}_{\text{NW}}(x^k)_j + \varepsilon_i = \sum_{i=1}^n w_{ji}^k Y_i + \varepsilon_i \quad (11)$$

$$= \sum_{i=1}^n W_k Y_i + \varepsilon_i.$$

The explanation of formula (11) is as follows: for each group X^k , there is a weight coefficient matrix

$$W_k = \begin{bmatrix} w_{11}, w_{12}, \dots, w_{1n} \\ w_{21}, w_{22}, \dots, w_{2n} \\ \vdots \\ w_{n1}, w_{n2}, \dots, w_{nn} \end{bmatrix}_k, \text{ so we can obtain a set of cen-}$$

tral-site estimator matrix data $Y_i^k = \begin{bmatrix} Y_{11}, Y_{12}, \dots, Y_{1k} \\ Y_{21}, Y_{22}, \dots, Y_{2k} \\ \vdots \\ Y_{n1}, Y_{n2}, \dots, Y_{nk} \end{bmatrix}$,

where each column is a set of estimates for the central station.

However, the above formula increases the amount of calculated data from single-column data to matrix data, and the corresponding result is also matrix data, which does not represent the central station directly. Therefore, this paper draws on the idea of a multidimensional regression formula and improves it into a multistation kernel regression formula, as follows:

$$\hat{g}_{\text{NW}}(x, y) = \frac{\sum_{i=1}^n k((x - X_i)/h_x)k((y - Y_i)/h_y)Z_i}{\sum_{i=1}^n k((x - X_i)/h_x)k((y - Y_i)/h_y)} = \sum_{i=1}^n w_i Z_i, \quad (12)$$

where X_i, Y_i , and Z_i are different types of multidimensional data, such as temperature and humidity, and the following is drawn from formula (12):

$$\hat{g}_{\text{NW}}(x_1, x_2, \dots, x_s) = \frac{\sum_{i=1}^n k((x_1 - X_{1i})/h_{x_1})k((x_2 - X_{2i})/h_{x_2}) \dots k((x_s - X_{si})/h_{x_s})Y_i}{\sum_{i=1}^n k((x_1 - X_{1i})/h_{x_1})k((x_2 - X_{2i})/h_{x_2}) \dots k((x_s - X_{si})/h_{x_s})} = \sum_{i=1}^n w_i Y_i, \quad (13)$$

where X_{si} is the surface air temperature observations of multiple neighbouring stations, Y_i is the surface air temperature observations of the central station, and \hat{g}_{NW} is the central station observations of the regression prediction.

3.2.3. Window Width Improved by PSO and Adaptive Algorithm. Since a fixed window width cannot effectively reflect the influence of the sparseness of the data, the width must be improved by the adaptive algorithm. Simultaneously, based on the improvement of the adaptive algorithm, particle swarm optimization (PSO) is proposed to further improve the method.

The specific design method is as follows: according to the kernel density estimates $f(x, y)$ and $f_X(x)$ in equation (6), the window width coefficient is designed on the basis of h being proportional to $f^{-1/2}$ [29]; $\lambda = [f/g]^{-\alpha}$, where g is the arithmetic mean of f , $g = (1/n) \sum_{i=1}^n f$, the effect of which is better than that of the geometric mean [30], and α is a sensitivity parameter that satisfies $0 \leq \alpha \leq 1$. Studies have shown that the practical effect is best when α is 0.5. Therefore, the adaptive window width is $h^* = \lambda h$, and h in equation (7) can be substituted to obtain the adaptive kernel regression algorithm formula as

$$\hat{g}_{NW}^*(x) = \frac{\sum_{i=1}^n k((x - X_i)/h^*) Y_i}{\sum_{i=1}^n k((x - X_i)/h^*)} = \sum_{i=1}^n w_i^* Y_i. \quad (14)$$

On this basis, a new window width formula is $\hat{h} = c \hat{\sigma} n^{-\alpha}$, where c and α are parameters to be determined and optimized by the adjusted PSO algorithm. Taking the kernel density estimation function as the objective function and assuming the N -dimensional space, the air temperature observations of multiple stations constitute a particle population $X = (X_1, X_2, \dots, X_d)$, where the i -th particle data $X_i = (x_{i1}, x_{i2}, \dots, x_{iN})^T$ are calculated by the objective function \hat{f} to obtain a set of potential solutions of the kernel density estimation function $F(X_i) = (\hat{f}_{i1}, \hat{f}_{i2}, \dots, \hat{f}_{iN})$. The RMSE is used as the fitness function: in the initial solution, the parameters c and α are set to 1.06 and 0.2, the speed parameter V is adjusted to the two change factors ω and μ , and the position parameter X is adjusted to the window width \hat{h} . The new window width formula is then $\hat{h} = (c + \omega) \hat{\sigma} n^{-(\alpha + \mu)}$, which is combined with the window width coefficient λ and substituted into equation (7) to obtain the following equation:

$$\hat{g}_{NW}^{**}(x) = \frac{\sum_{i=1}^n k((x - X_i)/\lambda(c + \omega) \hat{\sigma} n^{-(\alpha + \mu)}) Y_i}{\sum_{i=1}^n k((x - X_i)/\lambda(c + \omega) \hat{\sigma} n^{-(\alpha + \mu)})} = \sum_{i=1}^n w_i^{**} Y_i. \quad (15)$$

The window width improvement in the multistation kernel regression formula (13) is the same.

3.3. Evaluation of Model Performance. Commonly used evaluation parameters are the MAE, RMSE, IOA, and NSC. The MAE and RMSE represent the prediction accuracy: the smaller the value is, the higher the accuracy is. The NSC and IOA measure the goodness of fit: the larger and closer to 1 the value is, the better the goodness of fit is.

$$\begin{aligned} \text{MAE} &= \frac{1}{n} \sum_{i=1}^n |y_{\text{est}} - y_{\text{obs}}|, \\ \text{RMSE} &= \sqrt{\frac{\sum_{i=1}^n (y_{\text{obs}} - y_{\text{est}})^2}{n}}, \\ \text{NSC} &= 1 - \frac{\sum_{i=1}^n (y_{\text{obs}} - y_{\text{est}})^2}{\sum_{i=1}^n (y_{\text{obs}} - \bar{y})^2}, \\ \text{IOA} &= 1 - \frac{\sum_{i=1}^n (y_{\text{est}} - y_{\text{obs}})^2}{\sum_{i=1}^n (|y_{\text{obs}} - \bar{y}| + |y_{\text{est}} - \bar{y}|)^2}. \end{aligned} \quad (16)$$

4. Results and Discussion

4.1. Results and Analysis of Jiangsu Single-Station Test and Multistation Test. The surface air temperature observations of Jiangsu Province from 1961 to 2008 were selected to meet the conditions of large samples. First, the single-station kernel regression test in Jiangsu was conducted: the Nanjing station was the central station, and the other stations were used for regression predictions of the neighbouring stations. On this basis, four time scales—year, quarter, month, and day—were tested separately, and the four indicators (MAE, RMSE, NSC, and IOA) were calculated. The test results are as follows.

4.1.1. Single-Station Basic Test at Different Time Scales.

Figure 1 shows that, overall, the neighbouring stations with shorter distances to the central station (NJ), such as YZ and SZ, have better regression effects, but this conclusion is not absolute. For example, when the neighbouring station is ZJ, the effect of the regression test is not as good as that when the neighbouring stations are LYG and SQ. Compared to the number of observations in the kernel regression test at the annual time scale, that at the quarterly, monthly, and daily time scales is greatly increased. According to the kernel regression evaluation indicators in Figure 1, on the quarterly time scale, with the Nanjing station as the central station, the kernel regression effect of the whole region is better when the city goes southeast. On the monthly and daily time scales, the overall effect is similar to that of the annual time scale: the shorter the Euclidean distance between the neighbouring station and the central station is, the better the kernel regression effect is. In addition, since the analysis uses a single-neighbour kernel regression test, even the daily average air temperature observations are acceptable in terms of computational efficiency.

The analysis is performed based on comparison charts at different time scales under different indicators (Figure 1). From the perspective of the prediction accuracy indicators MAE and RMSE, the prediction effects on the annual time scale are the best. On the daily time scale, the effectiveness of a few stations that are close to the central station is similar to that on the quarterly and monthly time scales, and the other effects are not satisfactory. The prediction effects on the

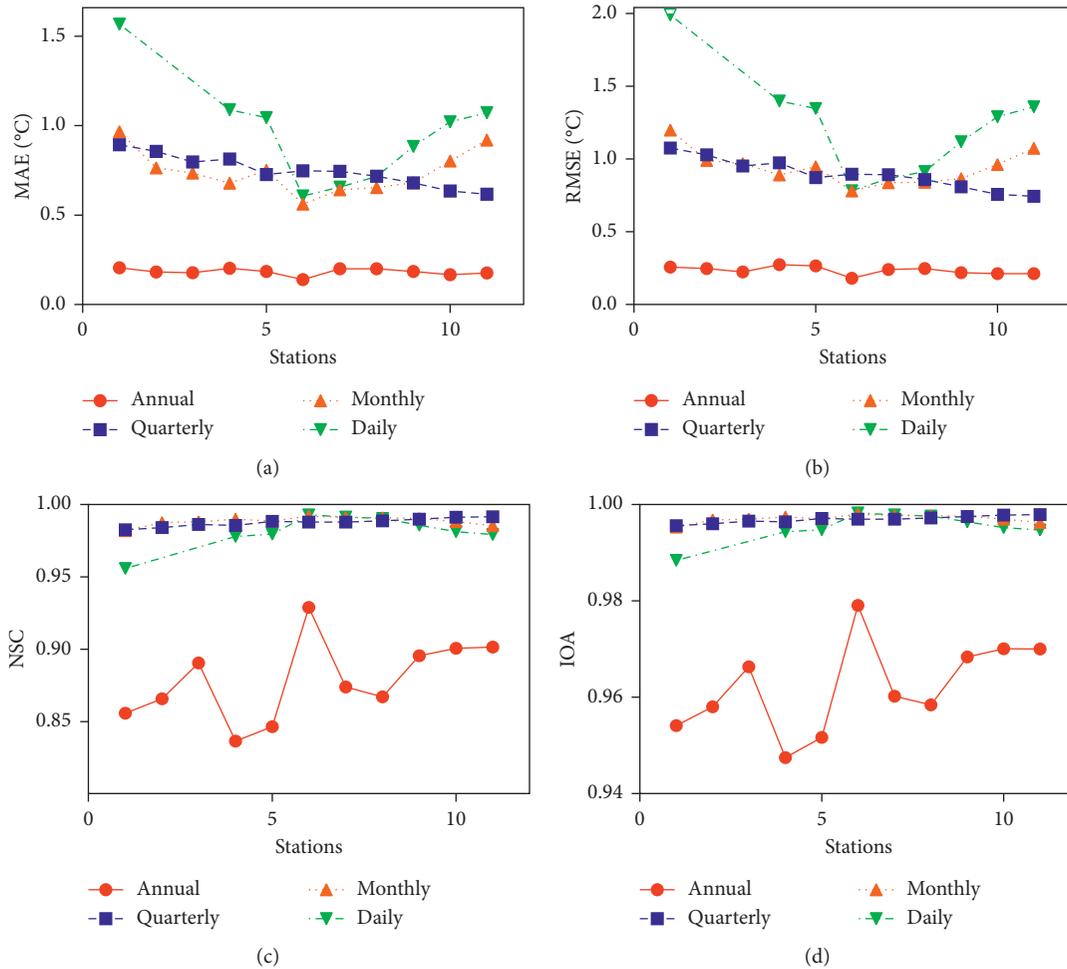


FIGURE 1: Comparison of single-station kernel regression indicators at multiple time scales: (a) MAE; (b) RMSE; (c) NSC; (d) IOA.

quarterly and monthly time scales also do not reach the prediction accuracy standard (RMSE value below 0.6) of the conventional regression algorithm. From the perspective of the fitting accuracy indicators NSC and IOA, except for the low effectiveness of the daily time scale (in terms of the numerical value, the fitting effect is good) far from the central station in the northern part of Jiangsu Province, the fitting accuracy of the quarterly, monthly, and daily time scales is very good. The fitting effect on the annual time scale is not good: the overall indicator value is not high, the numerical value is unstable, and there is no clear relation.

Jiangsu is located in the plain area in the southeast where the climate is pleasant and no obvious air temperature differences exist throughout the whole area. Therefore, when conducting the experiments, the Euclidean distance will be affected by factors such as the geographical location of different urban areas. These influencing factors are more obvious when the overall air temperature changes only slightly, so different regions in the country must be selected for verification experiments.

4.1.2. Multistation Test at Different Time Scales. Based on the abovementioned experiments, the multistation test of

kernel regression was conducted. The Jiangsu area was taken as an example, and the number of neighbouring stations was increased from 2 to 11. To solve the problem of irregularity in station selection when increasing the number of neighbouring stations, this paper calculates the indicators by traversing all the neighbouring stations in turn and taking the average. The specific test results are as follows.

According to Figure 2, from the perspective of the prediction accuracy indicators MAE and RMSE, as the number of neighbouring stations increases gradually, the values of the regression indicators for the central station gradually decrease and become stable. When the number of neighbouring stations is 5 or more, the MAE and RMSE can reach 0.6 or less on three time scales. When the number of neighbouring stations reaches 8, the values stabilize. The regression effects on the three time scales are ranked as follows: year > month > quarter. From the perspective of the fitting accuracy indicators NSC and IOA, as the number of neighbouring stations increases, the values of the indicators for the central station gradually increase and tend to 1. Moreover, the indicators on the quarterly and monthly time scales stabilize at approximately 6 neighbouring stations. In addition, although the curves on the three time scales do not look ideal, in terms of the numerical value, the fitting

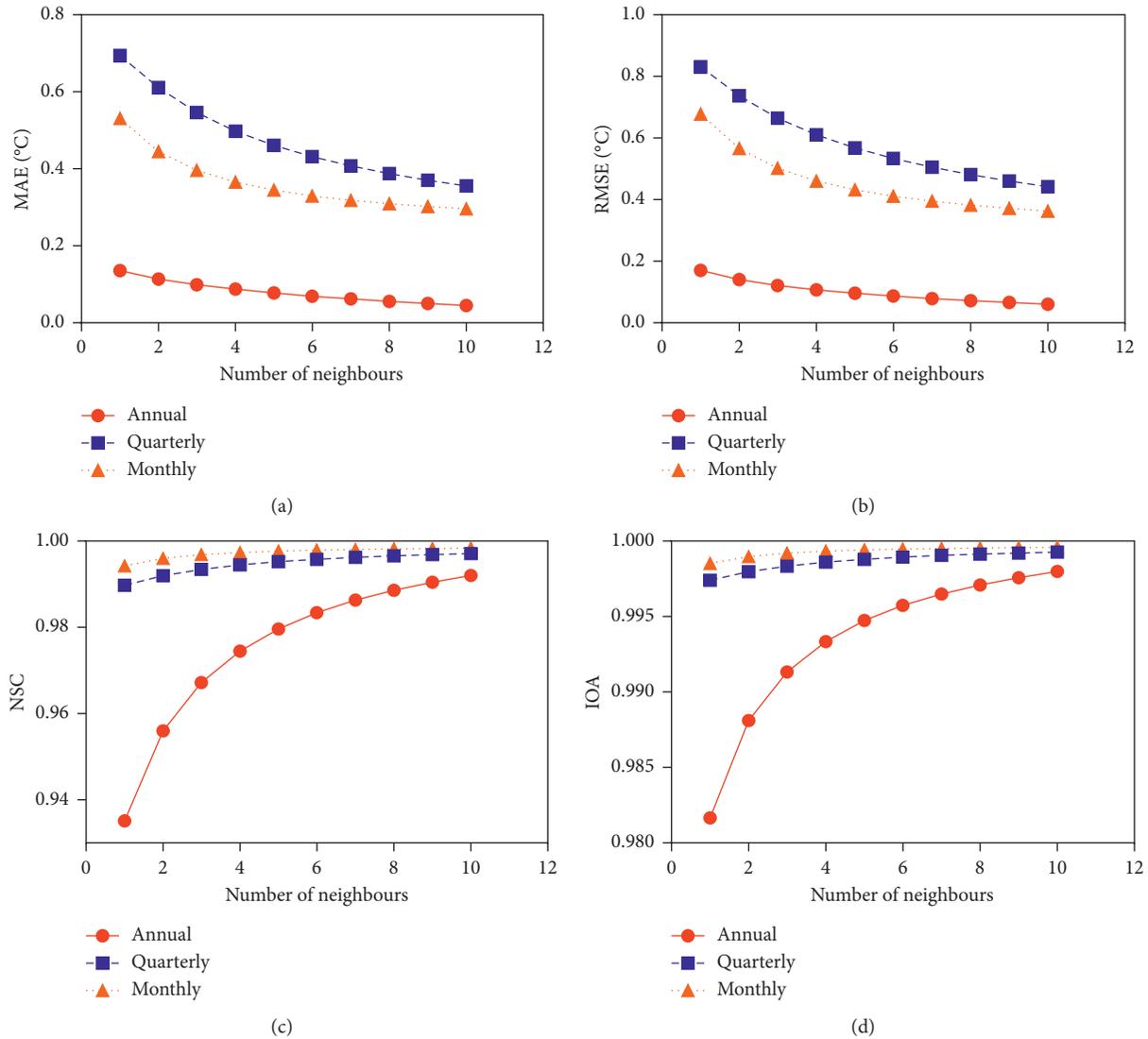


FIGURE 2: Comparison of kernel regression evaluation indicators of multiple neighbouring stations in the Jiangsu area: (a) MAE; (b) RMSE; (c) NSC; (d) IOA.

indicators are very good. On the daily time scale, for which the quantity of data is approximately 30 times that of the monthly data, the time required for the calculation is nearly 40 times greater and the calculation efficiency is linearly decreased. Therefore, for the daily average air temperature data, this paper does not consider the data of Jiangsu Province because the temperature is relatively stable and the calculation efficiency is low. In Section 4.2, this paper selects the daily average air temperature observations of fourteen regions in China from 2010 to 2014 for the kernel regression test on a daily time scale.

4.2. Results and Analysis of the National Multistation Test. The national average daily air temperature observations from 2010 to 2014 are selected for part of the experiment. Due to the inconsistent number of stations in the 14 regions across the country, there is no comparability between

different regions when gradually increasing the stations; therefore, the central station is considered to be the center, and the selection radius is gradually expanded from 20 to 200 km. That is, the concept of gradually increasing the number of neighbours (n) is replaced with the concept of gradually expanding the range of adjacent stations. The test is performed as follows.

Figure 3 shows that the kernel regression effect for multiple stations in China is similar to that in Figure 2 (Jiangsu multistation kernel regression). The MAE and RMSE of 11 regions (excluding JH, LS, and MH, where the sites are sparse) continue to decrease and eventually stabilize as the radius increases. When the radius reaches 160 km and above, regions reach unique stable values, and the improvement in prediction accuracy obtained by further increasing the radius is not substantial. The NSC and IOA gradually stabilize and approach 1 as the radius increases. When the radius reaches 120 km and above, the fitting

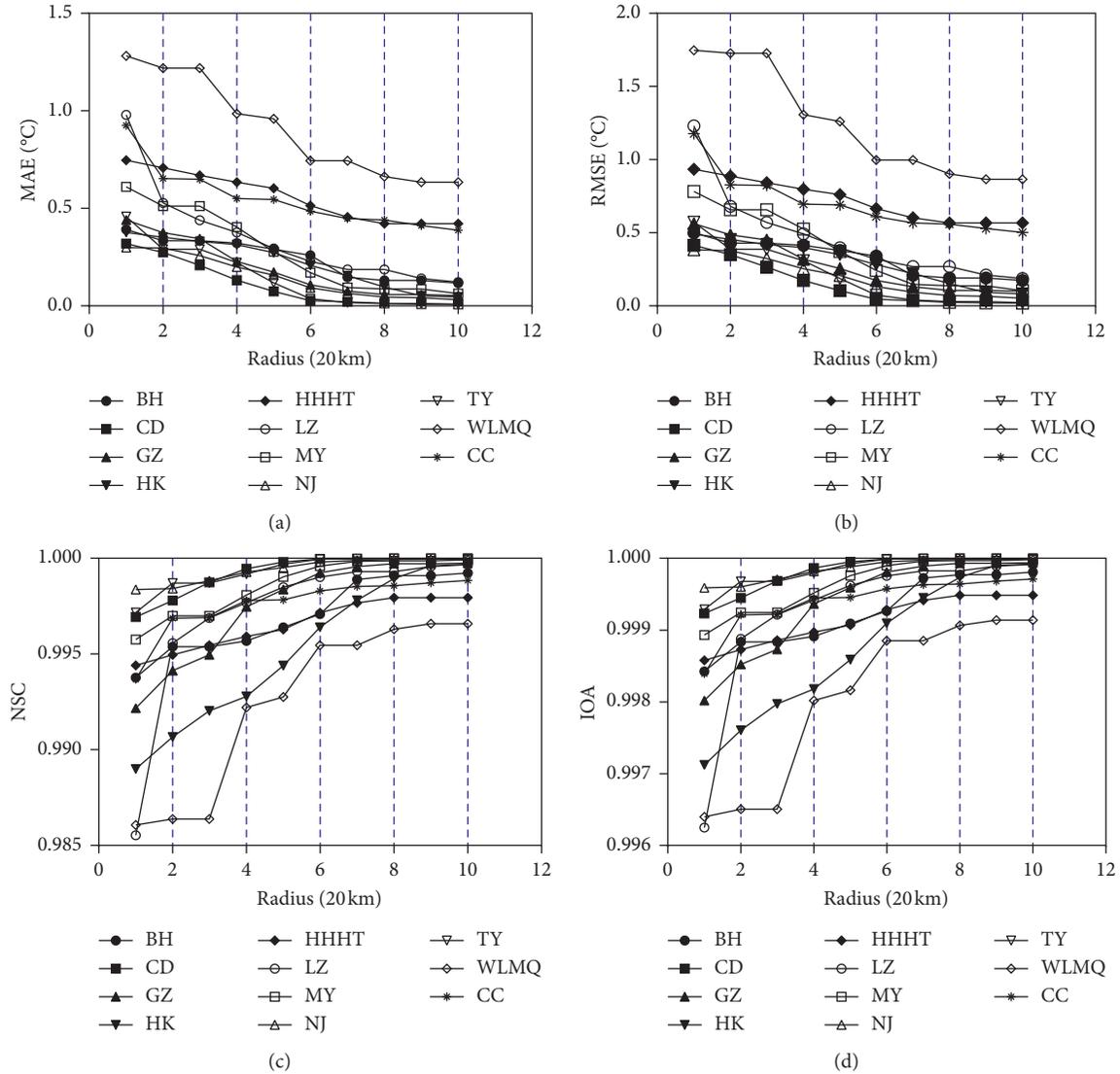


FIGURE 3: Comparison of various indicators with the increasing radius in different regions: (a) MAE; (b) RMSE; (c) NSC; (d) IOA.

indicator values are already close to 1, and the fitting accuracy cannot be further improved by increasing the radius. In addition, the effect of the kernel regression method is excellent in terms of the numerical values of the fitting indicators: even in the regression test with a radius of 20 km, the fitting result is quite good.

Furthermore, to improve the universality of the kernel regression method in different regions, we use the data from all stations within 200 km of the target station in the 14 regions (including JH, LS, and MH areas, where the stations are sparse) to conduct a comparative experiment to assess the prediction accuracy and fitting accuracy of the IKR method and KR method in different regions. The comparison chart is as follows.

As shown in Figure 4, in terms of the MAE and RMSE, the improvement of the IKR method relative to the KR method in the radius range of 20 km to 60 km is not obvious, and the prediction accuracy is only slightly increased. When the radius range is expanded to 100 km, the prediction

accuracy of the IKR method is improved substantially, and when the radius is expanded to 200 km, the IKR method is qualitatively better than the KR method: the improvement gradually stabilizes in the range of 160–200 km. In terms of the fitting accuracy indicator NSC, because the KR method already achieves good results, the IKR method shows little improvement compared to the KR method. In summary, the IKR method achieves good better universality and robustness and is suitable for QC research on surface air temperature observations in different regions of the country.

4.3. Comparative Test. The regression predictions of the IDW, SRT, KR, and IKR methods are compared for the single-station and multistation tests in Jiangsu and the multistation tests in different regions of the country. The evaluation selects three indicators: MAE, RMSE, and NSC. The abscissas 1, 2, and 3 are the annual, quarterly, and monthly time scales.

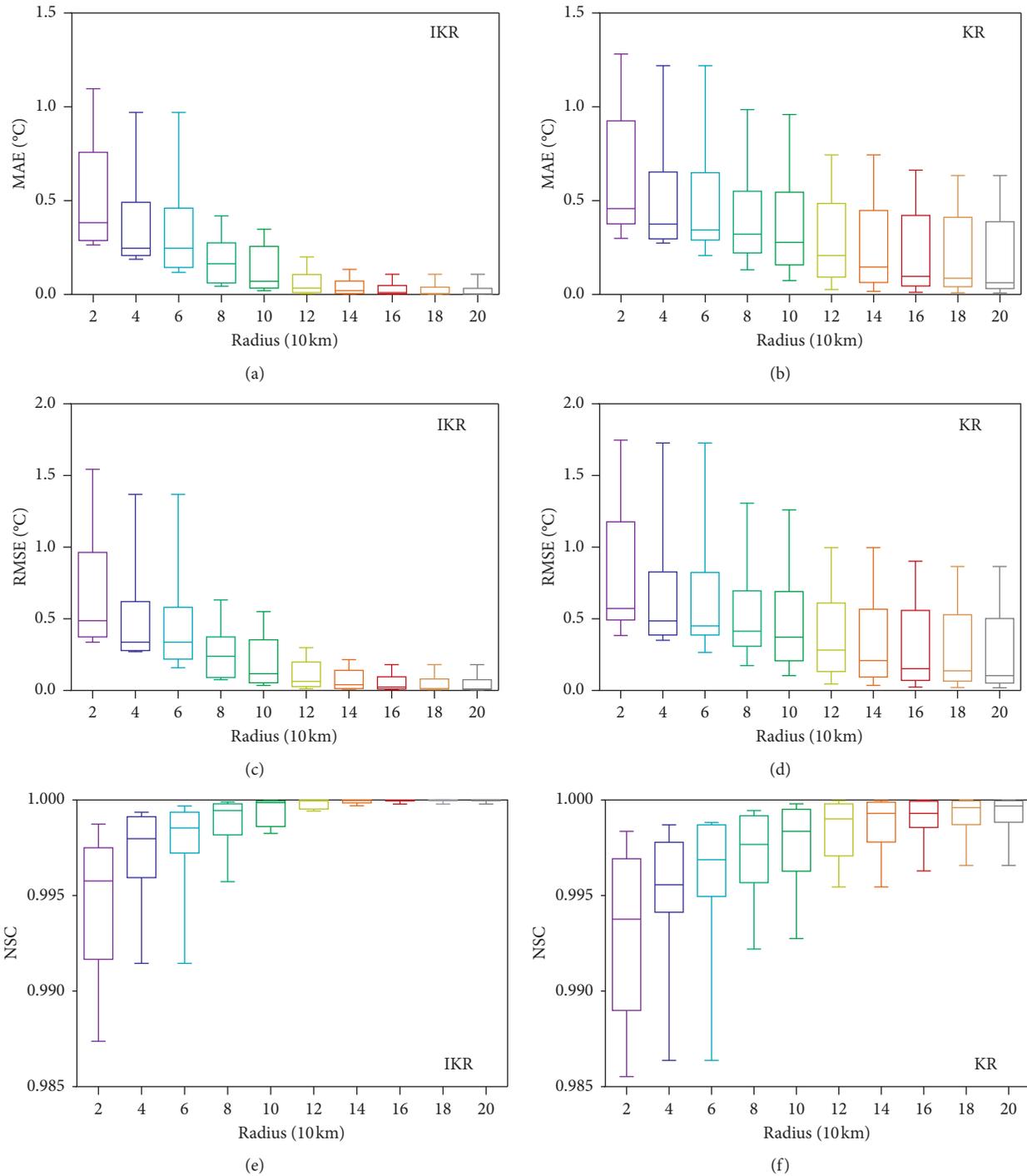


FIGURE 4: Comparison of the IKR and KR methods in terms of different indicators: (a, b) MAE; (c, d) RMSE; (e, f) NSC.

4.3.1. *Jiangsu Single-Station Comparison Test.* Figure 5 shows the optimal effect comparison of the four methods in the Jiangsu single-station regression. The MAE and RMSE of the IKR method are better than those of the other three methods on the annual, quarterly, and monthly time scales. The traditional IDW and SRT methods are second only to the IKR method on the annual, quarterly, and monthly time scales, and the SRT method is slightly better than the IDW method. Moreover, the effect of the KR method is between

those of the IDW and SRT methods on the annual time scale, but its prediction accuracy on the quarterly and monthly time scales is far worse than that of the other three methods. Therefore, the KR method does not achieve high universality in the single-station regression prediction in Jiangsu. The NSC indicates that the KR and IDW methods have lower fitting effects than the IKR and SRT methods on the annual time scale, and the IDW method is slightly lower than the KR method. On the quarterly and monthly time scales, the

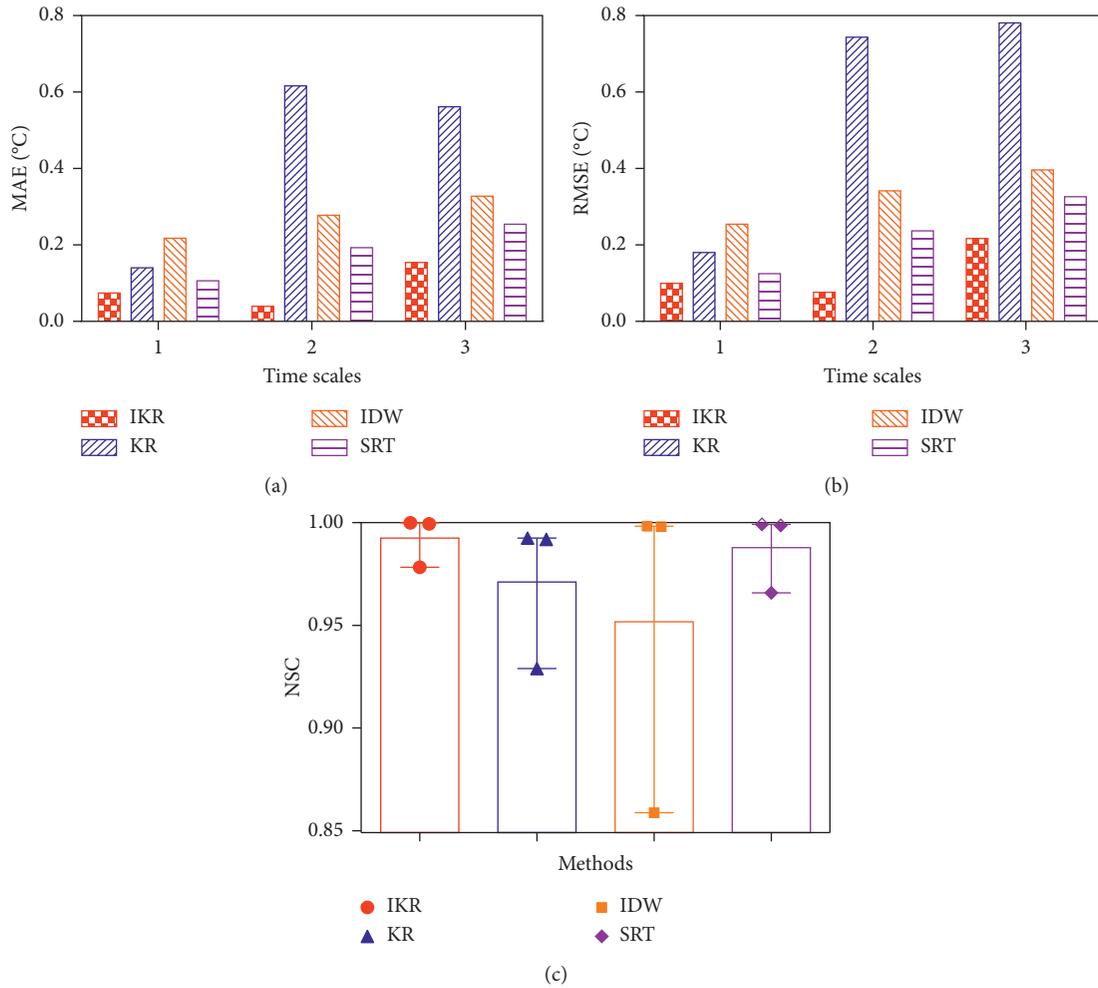


FIGURE 5: Comparison of the performance indicators of the four methods on different time scales for Jiangsu single-station kernel regression: (a) MAE; (b) RMSE; (c) NSC.

effects of the four methods are similar and the NSC values are close to 1.

and KR methods achieve better fitting effects than the traditional IDW and SRT methods on the annual time scale.

4.3.2. *Jiangsu Multistation Comparison Test.* Figure 6 presents the optimal results of the four methods in the Jiangsu multistation regression prediction experiment. In terms of the MAE and RMSE, the traditional IDW and SRT methods show little difference in prediction accuracy from the single-station regression prediction and achieve good prediction results. The prediction effect of the KR method on the quarterly and monthly time scales is obviously improved and approaches the prediction effect of the traditional method. On the annual time scale, the KR method is far better than the traditional methods. Overall, the KR method is more suitable for multistation regression prediction experiments than for single-station experiments. The prediction effect of the IKR algorithm on the annual and monthly time scales has been improved, but that on the quarterly time scale has been reduced. However, the IKR method is the best among the four methods. The overall trend of the NSC is similar to that in the Jiangsu single-station test, but the IKR

4.3.3. *National Multistation Comparison Test.* In Figure 7, abscissas 1–14 represent BH, CD, GZ, HK, HHHT, JH, LS, LZ, MY, MH, NJ, TY, WLMQ, and CC, respectively. The MAE and RMSE indicate that the IKR method has excellent prediction effects in the 14 regions of the country, whereas the other three methods have disadvantages in terms of universality and robustness. For example, in terms of the MAE, although the KR method has a good predictive effect in most areas, the effect in the MH region is not as good as that of the traditional IDW and SRT methods. Furthermore, the IDW and SRT methods provide good predictions in station-intensive areas, but the prediction accuracy in areas such as JH, LS, LZ, MH, and WLMQ is reduced or even unsatisfactory. The overall trend of the RMSE is similar to that of the MAE, but the prediction accuracy of IDW in HHHT and CC is not good. In terms of the NSC, the fitting effect of the IKR, KR, and SRT methods is very good in the 14 regions of the country, but the fitting effect of the IDW

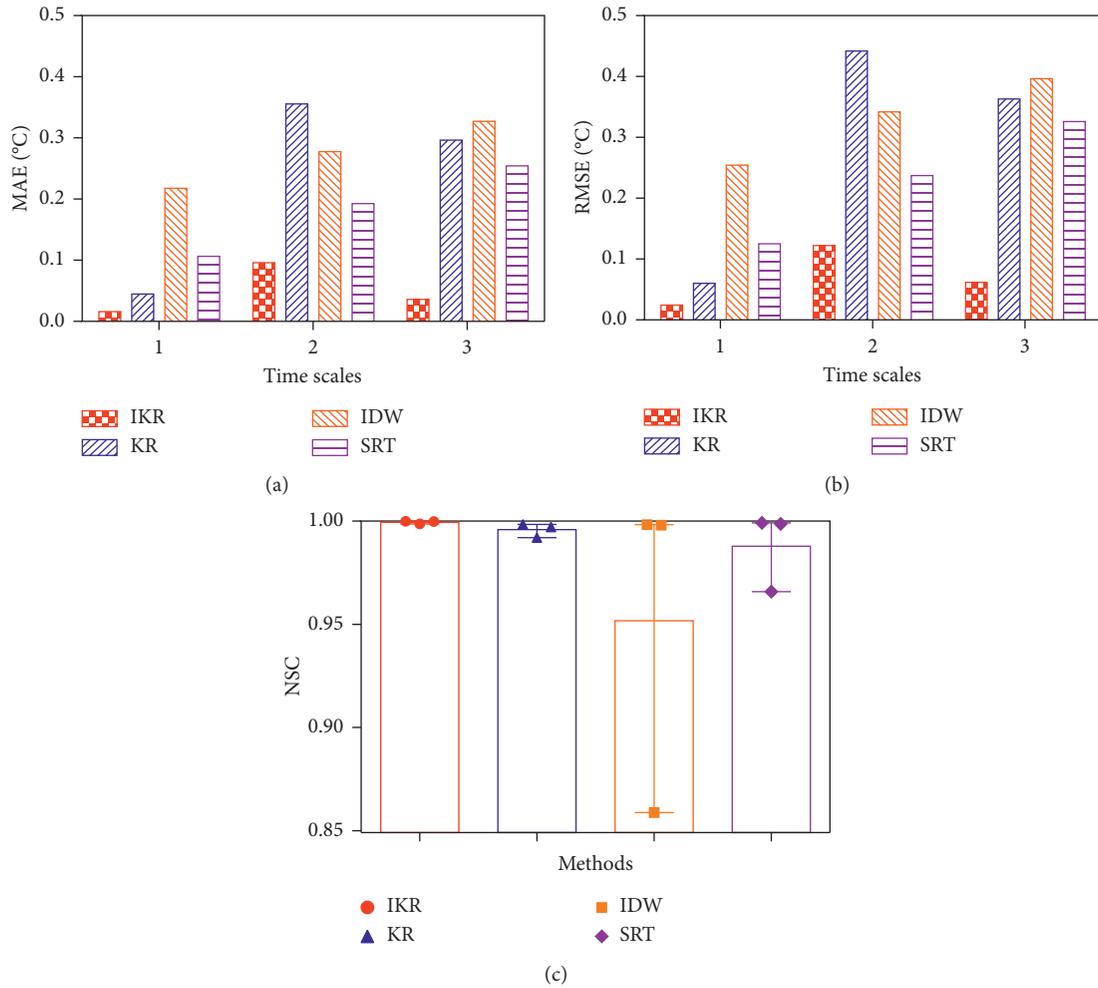


FIGURE 6: Comparison of the indicators of the four methods on different time scales in the Jiangsu multistation kernel regression experiment: (a) MAE; (b) RMSE; (c) NSC.

method is not good in JH, LS, and LZ and is especially poor in JH.

4.3.4. Error Detection Rate. When testing a hypothesis, we always want to obtain a correct judgement, but according to probability statistics, it is impossible to do so completely. We may make an incorrect judgement when we accept or reject a hypothesis.

There are two types of errors, one of which is called “the first kind of error” in mathematical statistics, that is, to “take the truth as false.” In statistical QC testing, if the data are correct and rejected, the “first-type error” occurs. The second type of error is to “pass the false as true”; that is, the data are wrong and accepted, resulting in “the second kind of error.” Table 1 shows the matrix of the relationship between the research conclusion and the actual situation.

The first type of error is also called an α error, which refers to rejecting H_0 when the null hypothesis H_0 is correct. In this case, the researcher’s conclusion is incorrect; that is, a treatment effect that does not actually exist is observed. Possible causes of this type of error are extreme values in the

sample and the adoption of relatively loose decision criteria. The second type of error, also known as a β error, refers to the situation in which the null hypothesis H_0 is accepted when the null hypothesis H_0 is wrong; that is, a treatment effect that actually exists is not observed.

The harm of making a type I error is substantial. Since a phenomenon that does not actually exist is reported, the harm of follow-up research and applications derived from the phenomenon will be inestimable. The potential harm of a type II error is relatively small. Therefore, on the basis of controlling the type I error, we should select appropriate QC parameters to minimize type II errors. According to this principle, the ratio of the number of errors detected by the QC method to the total number of inserted errors is calculated and referred to as the error detection rate. The performance of different QC models is evaluated by comparing the error detection rates. In this paper, the QC parameter values are set to those that produce an equivalent number of type I and type II errors.

Figure 8 shows the first- and second-error sizes of the IKR, SRT, and IDW algorithms under different QC parameter settings (four regions, BH, HK, LS, and MH, are

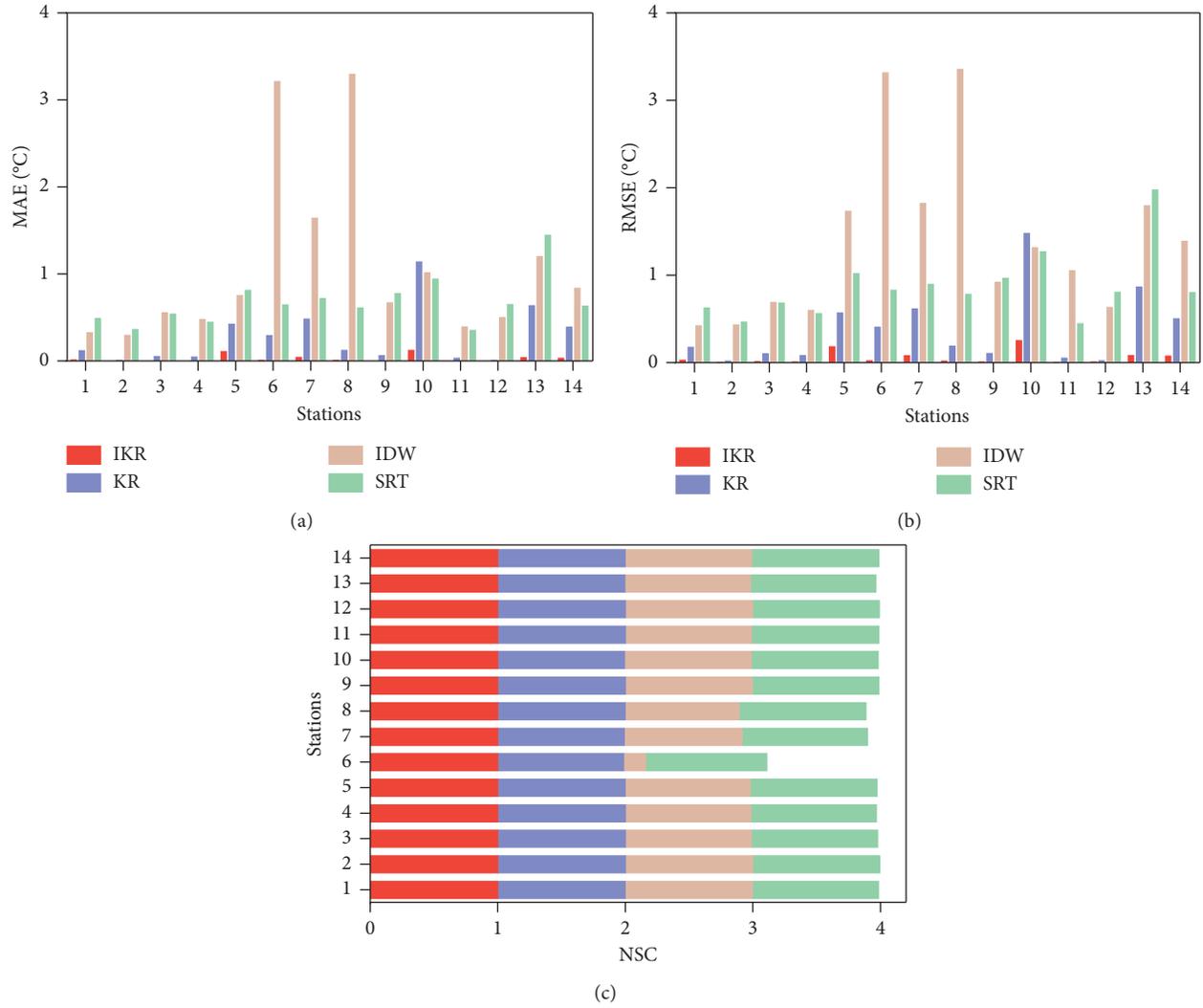


FIGURE 7: Comparison of the indicators of the four methods on different time scales in the national multistation kernel regression experiment: (a) MAE; (b) RMSE; (c) NSC.

TABLE 1: Relationship between two types of statistical errors.

		H_0 is correct	H_0 is incorrect
Research conclusion	Reject H_0	Type I error	Correct
	Accept H_0	Correct	Type II error

selected for display). Figure 8 shows that as the values increase, the first type of error decreases and the second type of error increases. Generally, the intersection of the curves of the first type of error and the second type of error is taken as the best QC parameter.

In Figures 8(a)–8(l), when selecting the best value, the relationship between the type I and type II error rates is as follows: $IKR < IDW < SRT$. Therefore, the IKR algorithm can effectively reduce the type I and type II error rates.

Figure 8 shows the change in the rates of the two types of errors for the IKR, IDW, and SRT algorithms with the change in QC parameters in four regions. Tables 2 and 3 show the best

QC parameters and corresponding error detection rates of the three algorithms in the 14 regions of China.

According to the data in Tables 2 and 3, in most regions, the effectiveness of the IKR algorithm in selecting the best QC parameters and maximizing the error detection rate is better than that of the IDW and SRT algorithms. In addition, in terms of selecting the best QC parameters, the IKR algorithm is between the IDW and SRT algorithms in the HHHHT and JH areas. In the MH area, the error detection rate of the IKR algorithm is slightly lower (2%) than that of the other two methods. Combined with the comprehensive analysis in Figure 9, these results show that the IKR

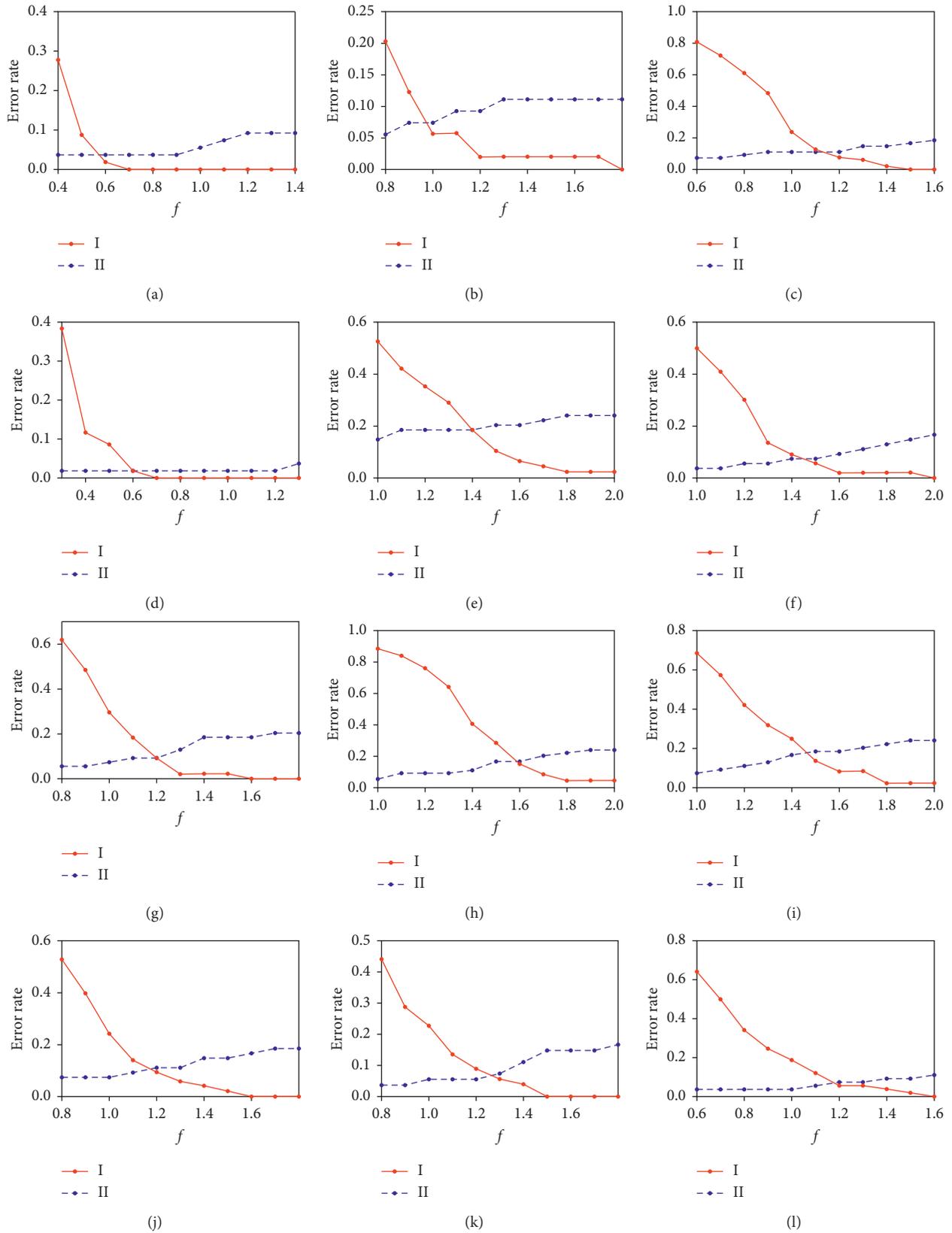


FIGURE 8: Comparison of the two types of errors under different f values, taking (a) IKR-BH, (b) IDW-BH, (c) SRT-BH, (d) IKR-HK, (e) IDW-HK, (f) SRT-HK, (g) IKR-LS, (h) IDW-LS, (i) SRT-LS, (j) IKR-MH, (k) IDW-MH, and (l) SRT-MH as examples.

TABLE 2: The best QC parameters of the three algorithms in the 14 regions of the country.

Region	IKR	IDW	SRT	Region	IKR	IDW	SRT
BH	0.53	0.88	1.16	CD	0.41	0.79	0.75
GZ	0.48	1.25	1.12	HK	0.45	1.41	1.45
HHHT	0.81	0.73	0.99	JH	1.51	1.43	1.97
LS	1.09	1.53	1.41	LZ	0.43	1.11	0.95
MY	0.43	1.10	1.02	MH	1.02	1.04	1.07
NJ	0.45	1.29	0.69	TY	0.49	0.74	0.91
WLMQ	0.92	1.65	1.77	CC	0.51	1.07	0.79

TABLE 3: Error detection rates under the best QC parameters in the 14 regions of the country.

Region	IKR	IDW	SRT	Region	IKR	IDW	SRT
BH	92%	83%	80%	CD	93%	82%	90%
GZ	91%	82%	80%	HK	92%	79%	84%
HHHT	90%	84%	84%	JH	74%	50%	69%
LS	89%	75%	82%	LZ	91%	84%	81%
MY	93%	84%	84%	MH	82%	86%	84%
NJ	93%	78%	91%	TY	93%	80%	87%
WLMQ	87%	77%	71%	CC	92%	78%	87%

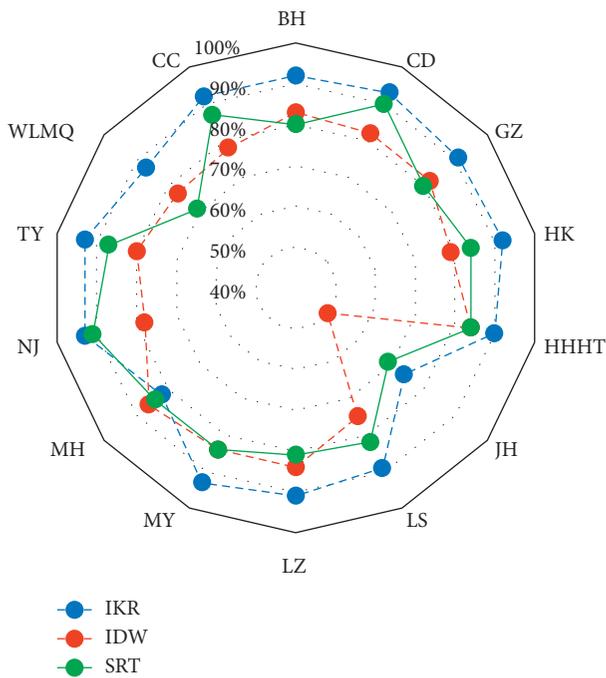


FIGURE 9: Error detection rate under the best QC parameters in the 14 regions of the country.

algorithm has good universality in terms of the error detection rate for different regions.

5. Conclusions

The IKR method for the QC of surface air temperature observations is introduced in this paper. The results show that the IKR method has better prediction accuracy and

fitting accuracy than the traditional QC methods. Moreover, the IKR method can effectively identify problems in surface air temperature observations. The specific conclusions are as follows:

- (1) Surface air temperature observations of the Jiangsu area are selected to verify the KR method. Whether the KR method can be applied to the regression prediction of surface air temperature observations is investigated, and the effectiveness of single-station and multistation predictions is determined. Finally, different regions of the country are selected to validate the feasibility of the KR method. The results show that application of the KR method in the regression prediction of surface air temperature observations is feasible, but its prediction accuracy and stability must be improved, and it cannot be applied to all regions.
- (2) The surface air temperature observations of 14 regions in the country are selected to verify the prediction effect of the IKR method, which is improved via PSO and the adaptive algorithm, compared with the KR method. The results show that the IKR method has higher prediction accuracy, fitting accuracy, universality, and robustness in different regions.
- (3) The surface air temperature observations of Jiangsu and 14 regions across the country are selected to compare the four methods considered in this paper. The error detection rate, prediction accuracy, and fitting accuracy demonstrate the superiority of the IKR method. Furthermore, the results show that the IKR method has good universality and robustness in terms of prediction accuracy, fitting accuracy, and error detection rate and is suitable for QC research in different regions.

Data Availability

The meteorological data used to support the findings of this study have not been made available because the data from the National Meteorological Center are confidential.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant no. 41675156, Talent Startup Project of Nanjing University of Information Science and Technology under Grant no. 2243141701053, General Program of Natural Science Research in Jiangsu Province (19KJB170004), and Key Scientific Research Projects of China State Railway Group Co., Ltd. (N2019T003).

References

- [1] Z. Xu, Yi Wang, and G. Fan, "A two-stage quality control method for 2 m temperature observations using biweight means and a progressive EOF analysis," *Monthly Weather Review*, vol. 141, no. 2, 2013.
- [2] J.-L. Qin, X.-H. Yang, H. Fu, and X.-F. Lei, "A novel nonlinear algorithm for area-wide near surface air temperature retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 3283–3296, 2016.
- [3] J.-L. Qin, X.-H. Yang, J.-T. Luo et al., "An improved novel nonlinear algorithm of area-wide near-surface air temperature retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 830–844, 2018.
- [4] Z. Ren, Z. Zhang, C. Sun et al., "Development of three-level quality control system for real-time observations of automatic weather stations nationwide," *Meteorology*, vol. 41, no. 10, pp. 1268–1277, 2015.
- [5] H. Zhao, *Research on Quality Control Methods of Surface Observations and Its Application in Numerical Model*, Nanjing University of Information Science and Technology, Nanjing, China, 2015.
- [6] Z. Ren, Xu Song, H. Sun, and Q. Zhang, "Quality inspection and analysis of historical data of global ground weather report," *Quarterly Journal of Applied Meteorology*, vol. 17, no. 4, pp. 412–420, 2006.
- [7] V. Lakshmanan, C. Karstens, J. Krause, and L. Tang, "Quality control of weather radar data using polarimetric variables," *Journal of Atmospheric and Oceanic Technology*, vol. 31, no. 6, pp. 1234–1249, 2014.
- [8] U. Schneider, A. Becker, P. Finger et al., "GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle," *Theoretical and Applied Climatology*, vol. 115, no. 1–2, pp. 15–40, 2014.
- [9] G.-L. Oh, S.-J. Lee, B.-C. Choi et al., "Quality control of agrometeorological data measured at suwon weather station of korea meteorological administration," *Korean Journal of Agricultural and Forest Meteorology*, vol. 17, no. 1, pp. 25–34, 2015.
- [10] İ Sönmez, B. B. Üstündağ, S. Bağış et al., "Agro-meteorological data quality control system design for turkey's agricultural monitoring and information system (TARBIL)," in *Proceeding of the International Conference on Agro-Geoinformatics*, IEEE, Istanbul, Turkey, July 2015.
- [11] A. J. Coop, K. G. Hubbard, M. D. Shulski, J. You, and D. B. Marx, "Spatial accuracy of climate networks: a case study in Nebraska," *Journal of Applied Meteorology and Climatology*, vol. 53, no. 8, pp. 1932–1942, 2014.
- [12] M. A. Shafer, C. A. Fiebrich, D. S. Arndt et al., "Quality assurance procedures in the Oklahoma mesonet network," *Journal of Atmospheric and Oceanic Technology*, vol. 17, no. 4, 2000.
- [13] P. Kubecka, "A possible world record maximum natural ground surface temperature," *Weather*, vol. 56, no. 7, pp. 218–221, 2001.
- [14] B. E. Gleason, *Global Daily Climatology Network*, National Climatic Data Center, Asheville, NC, USA, 2002.
- [15] K. G. Hubbard and J. You, "Sensitivity analysis of quality assurance using the spatial regression approach—a case study of the maximum/minimum air temperature," *Journal of Atmospheric and Oceanic Technology*, vol. 22, no. 10, pp. 1520–1530, 2005.
- [16] C. Yin, Z. Jiang, X. Wu, and X. Ju, "Application of spatial difference test method in quality control of surface meteorological observations," *Climatic and Environmental Research*, vol. 15, no. 3, pp. 229–236, 2010.
- [17] X. Liu, X. Ju, and S. Fan, "Application of spatial regression test method in quality inspection of meteorological observations," *Quarterly Journal of Applied Meteorology*, vol. 17, no. 1, pp. 37–43, 2006.
- [18] A. Xiong, "Quality control of meteorological observations in Northern Europe," *Meteorological Science and Technology*, vol. 31, no. 5, pp. 314–320, 2003.
- [19] Z. Ren, X. Liu, and W. Yang, "Comprehensive quality control and analysis of extreme anomalous meteorological observations," *Acta Meteorologica Sinica*, vol. 63, no. 4, pp. 526–533, 2005.
- [20] L. Li, H. Wang, J. Liu, and S. Song, "Quality control of surface meteorological observations based on blackboard model," *Meteorological Science and Technology*, vol. 34, no. 2, pp. 199–204, 2006.
- [21] H. Wang, Q. Yan, F. Xiang, and W. Pan, "Design of the threshold value checking algorithm for quality control of hourly temperature," *Plateau Meteorology*, vol. 33, no. 6, pp. 1722–1729, 2014.
- [22] X. Ye, J. Zhou, and X. Xiong, "A quality control method for surface air temperature observations based on GEP," *Journal of Tropical Meteorology*, vol. 30, no. 6, pp. 1196–1200, 2014.
- [23] X. Ye, X. Yang, X. Xiong, Y. Shen, M. Hao, and R. Gu, "A quality control method based on an improved random forest algorithm for surface air temperature observations," *Advances in Meteorology*, vol. 2017, Article ID 8601296, 15 pages, 2017.
- [24] Y. Zhang, R. Yao, X. Xiong, and Y. Shen, "Application of PSO-PSR-ELM integrated learning algorithm in quality control of surface air temperature observations," *Climatic and Environmental Research*, vol. 22, no. 1, pp. 59–70, 2017.
- [25] Z. Xu, X. Chen, and W. Wang, "Design of data quality control method for new surface meteorological automatic station," *Scientia Meteorologica Sinica*, vol. 33, no. 1, pp. 26–36, 2013.
- [26] X. Xiong, X. Ye, Y. Zhang, N. Sun, H. Deng, and Z. Jiang, "Study on quality control algorithm of surface air temperature observations based on spatial observations' difference," *Chinese Journal of Geophysics*, vol. 60, no. 3, pp. 912–923, 2017.
- [27] K. G. Hubbard, N. B. Guttman, J. You, and Z. Chen, "An improved QC process for temperature in the daily cooperative weather observations," *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 2, pp. 206–213, 2007.
- [28] B. Wang, L. Shi, and Y. Lu, "A space-time inverse distance weighted interpolation method considering elevation," *Surveying and Spatial Geography Information*, vol. 41, no. 10, pp. 82–85, 2018.
- [29] N. Yang, Y. Huang, D. Ye et al., "Modeling of multi-wind farm output correlation based on adaptive multivariable non-parametric kernel density estimation," *Proceedings of the CSEE*, vol. 38, no. 13, 2018.
- [30] Y. Zhang, "Improvement of adaptive N-W kernel regression estimator," *Statistics and Decision*, vol. 34, no. 5, pp. 16–19, 2018.