Precipitation quality control method based on SFA and phase space local prediction

X Chen¹, H S Tang², Y J Kan¹ and X Xiong¹

¹School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China
²Jiangsu Meteorological Bureau, Nanjing 210000, China
18751903109@163.com, achendian1@163.com, kanyajin@126.com, nxgxiong@163.com

Correspondence should be addressed to X Chen, E-mail: 18751903109@163.com,

Abstract. Most of the quality control applications of hydrological data are based on basic quality control methods such as logical detection, extreme value check and spatial consistency check. Although these methods can detect problem data with large errors, this makes the data lack credibility. Therefore, a single station data quality control method, SFA-WZLM, is proposed in this paper. This method uses slow feature analysis (SFA) to extract external forcing factors for embedding in chaotic local prediction models. Observations from January 1987 to October 2015 was used as the train set, and observations from December 2015 to October 2017 were used as the test set. The results indicate that the method has higher prediction accuracy than the prediction model without embedded external forcing factors, weighted first-order local prediction model (WFLM) and weighted first-order local prediction model (SFA-WFLM) including external forcing factors and exhibited the best quality control error detection. In addition, the method shows good stability in 6 different climates and different terrain stations across the country in China.

1. Introduction

As part of the weather system, precipitation is the main factor of the climate. It not only directly affects the hydrological process, but also plays a vital role in climate regulation [1]. Precipitation provides basic data in numerical weather simulation, hydrology and water resources management and weather forecasting. Domestic scholars have proposed many quality management and evaluation schemes for surface precipitation data. Ren Z H et al. [2] analyzed the existence of erroneous data in depth, and developed a national automatic station hourly precipitation data control scheme that made hourly precipitation data a major breakthrough in the "national level" automatic control method. The current QC methods applied to precipitation data in actual meteorological operations are mainly traditional QC methods such as extreme value inspection, time consistency and space consistency inspection. However, it is difficult to provide more accurate precipitation data through traditional QC methods alone [3].

In order to solve the problems of the lack of accuracy of the current preliminary precipitation QC methods and the unsatisfactory interpolation effect in sparse stations. Based on the slow feature analysis, this paper extracts the external forcing factors of the solar precipitation time series, reconstructs the phase space of the original sequence and the external forcing factors, and compares
the two methods of local space phase prediction including external forcing factors. The method with good prediction effect was selected and tested. The stability and applicability of this method to the QC of single station precipitation in the country.

2. Data
The data in this article are from the National Meteorological Center, including daily observations of precipitation at 8:00 to 20:00 from 1987 to 2017 at more than 2,000 surface observation stations across the country. In this paper, six stations with different geographical environments and climatic characteristics are selected as target stations for experiment. The six stations are Nanjing (NJ), Yuxi (YX), Baoshan (BS), Mohe (MH), Chang'an (CA) and Nagqu (NQ).

3. Method
3.1. Slow Feature Analysis
The Slow Feature Analysis (SFA) method based on singular spectrum analysis by Wiskott and Sejnowski can extract linear factors of invariant features from a known non-stationary time series [4]. The specific expression of the slow feature is as follows [5]:

For a given time series of \( x(t) = [x_1(t), x_2(t), \ldots, x_L(t)]^T \), the linear and quadratic terms in \( x(t) \) are used to expand it non-linearly to construct a \( k \)-dimensional function space:

\[
H(t) = (h_1(t), h_2(t), \ldots, h_k(t))
\]

Where \( k = L + L(L+1)/2 \), Pre-whitening \( H(t) \) to obtain a whitening sequence \( Z(t) \):

\[
Z(t) = \{z_1(t), z_2(t), \ldots, z_L(t)\}_{t=t_1,t_2,\cdots,t_L}
\]

The obtained \( Z(t) \) satisfies \( ZZ^T = 1 \), \( \overline{Z} = 0 \). At this time, each component of \( Z(t) \) can be represented by a linear combination of \( Z_j \), and the first derivative is represented by \( \dot{z}_j(t_i) = z_j(t_{i+1}) - z_j(t_i) \) to construct a derivative space of \( Z(t) \):

\[
\dot{Z}(t) = \{\dot{z}_1(t), \dot{z}_2(t), \ldots, \dot{z}_L(t)\}_{t=t_1,t_2,\cdots,t_L}
\]

Finally, the principal component analysis is performed on the matrix \( <\dot{Z}\dot{Z}^T> \) to obtain the eigenvalues \( \lambda_j \) and eigenvectors of \( <\dot{Z}\dot{Z}^T> \) to obtain the weight vector \( w_1, \ldots, w_j \) after normalization. The output signal is as follow:

\[
g_j(t) = w_j \dot{Z}(t)
\]

Where \( g_j(x(t)) \) is the external forcing factors.

3.2. Phase space reconstruction
The phase space reconstruction theory proposed by Packard in 1980 reconstructs time series with chaotic characteristics. The constructed high-dimensional space can be used to mine the evolution laws of dynamic systems. Takens proved the embedding theorem in 1981 [6]. As long as the dimension \( m \geq 2D+1 \) ( \( D \) is the dimension of the dynamic system) is embedded, the phase space can be reconstructed. Univariate time series \( x = \{x(t), t = 1, 2, \ldots, N\} \) with length \( N \) and the reconstructed phase space \( X(t) \) is as follows:

\[
X(t) = [x(t), x(t + \tau), \ldots, x(t + (m-1)\tau)], \quad t = 1, 2, \ldots, M
\]

Where \( m \) is the embedding dimension, \( \tau \) is the delay time, and the length of the reconstructed sequence is \( M \) and \( M = N - (m-1)\tau \).
3.3. Chaotic local prediction model with external forcing factors

Two algorithms commonly used in chaotic local prediction models are the weighted zero-order local prediction method (WZLM) and the weighted first-order local prediction method (WFLM). The two methods are based on several points with the smallest distance between the center points as the correlation points, find the correlation between these points and the center point, fit the next phase point, and then extract the required prediction value.

3.3.1. Weighted zero-order local prediction algorithm with external forcing factors (SFA-WZLM).

Weighted zero-order local method (WZLM) adds weights to the zero-order prediction method and has a certain noise reduction ability. Let the center point be \( M \), find \( f \) neighboring points \( X_{Mi} \), \( i=1,2,\cdots,f, f=m+1 \), with the smallest Euclidean distance from the center point, the distance is expressed by \( d_i \), and \( d_m \) is the minimum value of \( d_i \), then the weight can be obtained by the following formula:

\[
P_i = \frac{e^{-c(d_i-d_m)}}{\sum_{i=1}^{f} e^{-c(d_i-d_m)}}
\]

Where \( c \) usually takes 1.

The weighted zero-order local prediction algorithm (SFA-WZLM) with external forcing factors is an external forcing of precipitation sequences extracted by the SFA method into the WZLM model. Considering the precipitation sequence \( \{x(t), t=1,2,\cdots,N\} \) and the external forcing sequence \( \{\alpha_i, i=1,2,\cdots,N\} \), the two sequences are simultaneously reconstructed and normalized in phase space to obtain the reconstructed sequence \( Y_n \):

\[
\{Y_n\} = \{x_1, x_2, \cdots, x_n; \alpha_1, \alpha_2, \cdots, \alpha_n\}
\]

Let \( Y_M \) be the center point of the sequence \( \{Y_n\} \), find the \( f \) points with the smallest distance from the center point in the phase points, and calculate \( Y_{M+1} \) according to the weight given by Equation 5. The \( Y_{M+1} \) is calculated using (8).

\[
Y_{M+1} = p_i Y_{Mi} = \sum_{i=1}^{f} \frac{Y_{Mi} e^{-c(d_i-d_m)}}{\sum_{i=1}^{f} e^{-c(d_i-d_m)}}
\]

Anti-normalization \( Y_{M+1} \) to get the predicted value \( X_{M+1} \), and the \( m \)-th component is the predicted result \( \hat{x}_{N+1} \) at the next moment.

3.3.2. Weighted zero-order local prediction algorithm with external forcing factors (SFA-WFLM).

The weighted first-order local method (WFLM) is to establish a first-order linear fitting relationship \( X_{M+1} = aE + bX_M \) between the neighbouring point \( X_M \) of the center point \( X_M \) and the phase point set \( X_{M+1} \) of the neighbouring step one step, where \( E = (1,1,\cdots,1)^T \) is a \( m \times 1 \) order matrix, \( a \) \( b \) is the fitting coefficient, which can be obtained by the method of least squares. Considering the precipitation sequence \( \{x(t), t=1,2,\cdots,N\} \) and the external forcing sequence \( \{\alpha_i, i=1,2,\cdots,N\} \), the two sequences are simultaneously reconstructed and normalized in phase space to obtain the reconstructed sequence of formula (7), which is abbreviated as (8).

\[
\{Y_n\} = \{y_1, y_2, \cdots, y_n\}
\]
Find the adjacent point \( Y_{M_i} \) of the central point \( Y_M \) of the sequence \( \{Y_n\} \), \( i = 1, 2, \cdots, f \), then use the first-order local linear fitting to obtain the predictive model:

\[
Y_{M_{i+1}} = aE + bY_{M_i}
\]

According to the weighted least squares method, the optimal objective function as follows:

\[
\min z = \sum_{i=1}^{f} \sum_{j=1}^{f} (y_{M_{i+1}} - y_{M_j})^2
\]

Think of Equation 11 as a binary function on \( a \) and \( b \), with partial derivatives on both sides and making it equal to 0 to solve \( a \) and \( b \). Bring \( a \) and \( b \) into formula (10) to get \( \hat{Y}_{M_{i+1}} \):

\[
\hat{Y}_{M_{i+1}} = aE + bY_{M_i}
\]

Anti-normalization \( \hat{Y}_{M_{i+1}} \) to get the predicted value \( \hat{X}_{M_{i+1}} \), and the \( m \)-th component is the predicted result \( \hat{x}_{N_{i+1}} \) at the next moment.

4. Results and Discussion

In this paper, Nanjing Station is selected as the case of precipitation QC for analysis. The daily precipitation values from January 1987 to October 2017 are divided according to solar terms, and the precipitation values of each solar term are summed to form a precipitation sequence \( \{x(t), t = 1, 2, \cdots, N\} \) with a total sample length \( N = 734 \). SFA method for slow feature analysis of precipitation series to extract external forcing factors \( \{a_i, i = 1, 2, \cdots, N\} \), is shown in Figure 1. The phase space reconstruction parameters are set by Kim H S's C-C method based on correlation integral, which is based on statistical methods to estimate delay time and delay time window \( \tau \) simultaneously [7-8]. As is shown in Figure 2, according to the statistics \( \Delta S(t) \) and \( \Delta S_{\text{kor}} \) calculated by the C-C method, the optimal embedding time of the precipitation time series is 3 and the embedding dimension is 4. Then the phase sequence reconstruction of the precipitation sequence containing the external forcing factors is used to form \( M = N - (m - 1)\tau = 725 \) phase points. The first 663 phase points are extracted as samples and the last 72 phase points are used as verification samples. Finally, the SFA-WZLM method and SFA-WFLM method are used for prediction, and the prediction results are compared with the weighted zero-order local prediction method and the weighted first-order local prediction method.

![Figure 1. External forcing factors extracted from the solar station precipitation series in Nanjing Station.](image1)

![Figure 2. Using C-C method to determine the effect of embedded window on the precipitation time series of Nanjing Station.](image2)

4.1. Analysis of prediction results by four methods

Figure 3 shows the comparison of the prediction results of the four methods WZLM, SFA-WZLM, WFLM and SFA-WFLM with the original sequence. Figure 3 (a) (b) shows that SFA-WZLM's
prediction of precipitation is the closest to the true value, and WFLM’s prediction of precipitation is the most different from the true value. It can be seen from Figure 3 (c) (d) that the SFA-WZLM algorithm with external forcing factor has the smallest oscillation amplitude of prediction error, the maximum value is only 41.51mm, and the prediction error is relatively stable. The WZLM and WFLM methods have large fluctuations in prediction errors. It can be seen from Figure 3 (c) that the forecasting error of the SFA-WZLM algorithm with external forcing factors is small, the maximum value is 41.51mm, the forecasting error is relatively stable, and WZLM has a large fluctuation in forecasting error.

In order to test the predictive performance of the above four quality control methods, the three absolute indicators of quality control, mean absolute error (MAE), root mean square error (RMSE) and Nash coefficient (NSC), were selected for evaluation:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| 
\]

\[
RMSE = \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)^{1/2} 
\]

\[
NSC = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} 
\]

Figure 4 is a comparison of the effects of four methods at six stations: NJ, YX, BS, MH, CA, and NQ. Among them, (a) and (b) have similar change curves, and the precipitation characteristics of the 6 stations are different. It can be seen that of the four methods, the SFA-WZLM method performs best at
6 stations, MAE, The RMSE evaluation index is the lowest value, and the NSC index is the highest value, indicating that the method has obvious superiority in prediction accuracy. Among them, the MAE values of MH and NQ are lower than 5, and the RMSE values of MH and NQ below 5, NSC values of NJ, YX, MH, CA and NQ Station are all higher than 0.95. As shown in Figure 4 (a), before the external forcing factors is added, the MAE value of the WFLM method of each station is the largest. As shown in Figure 4(c); compared with the NSC index of the test estimate and the real value, the original WFLM method has improved to different degrees in the six stations, and the increase rates are 127.61%, 39.51%, 38.01%, 101.74%, 127.08%, and 46.35%.

![Figure 4. Comparison of the effects of four QC methods at six stations (a) MAE (b) RMSE (c) NSC](image)

4.2. Error detection rate analysis
QC model is used to predict the target stations of the precipitation observations, and the predicted values and observed values are tested via the threshold test. The artificial implantation error method proposed by Hubbard et al. Is introduced to judge the quality of the quality control method by examining the results of the incorrect data implanted in the original sequence.

3% artificial error was implanted in the NJ solar term precipitation test set and error detection analysis was performed (see Figure 5). The error detection effects of the four methods are SFA-WZLM, WZLM, SFA-WFLM, and WFLM in order from good to bad. Figure 5 shows that, when \( f = 1.2 \), the WFLM error detection rate is 0.7; when \( f = 0.77 \), the SFA-WFLM error detection rate is 0.73; when \( f = 1.17 \), the WZLM error detection rate is 0.75; when \( f = 0.47 \), the SFA-WZLM error detection rate is 0.82. The SFA-WFLM and SFA-WZLM QC methods have 0.03 and 0.07 higher error detection rates than the WFLM and WZLM methods, respectively, and the setting of QC parameters is reduced by 0.43 and 0.7 respectively compared to the original method. The above shows that after the introduction of external forcing factors, the SFA-WFLM and SFA-WZLM methods have improved error detection rates, and have better sensitivity to erroneous data.
Figure 5. Relationship between QC parameters and two types of errors of the four QC methods at Nanjing Station (a) WFLM; (b) SFA-WFLM; (c) WZLM; (d) SFA-WZLM.

The average error detection rate of 50 artificial errors was implanted for each of the four methods of NJ, YX, BS, MH, CA, and NQ. Figure 6 (a) is the error detection rate radar charts of the WFLM and SFA-WFLM methods at six stations and the area of the WFLM method's error detection area in the circled area is less than half that of SFA-WFLM. Figure 6 (b) is the error detection rate radar charts of the WZLM and SFA-WZLM methods at six stations and the WZLM method has a significantly smaller error detection area in the circled area than SFA-WZLM, with values between 0.7 and 0.85. The SFA-WZLM method maintains a high error detection rate. The error rate is 0.84, YX is 0.91, and other stations are between 0.85 and 0.9. Comparing the two graphs in Figures 6 (a) (b), the average error detection rate of the SFA-WZLM method at six stations is higher than that of other methods, and all are higher than 0.8, indicating that this method has strong stability and applicability.

Figure 6. Comparison of error detection rates at six stations (a) WFLM and SFA-WFLM; (b) WZLM and SFA-WZLM
5. Conclusion
This paper proposes a new single-station precipitation quality control method based on SFA and chaotic local prediction model. The following conclusions are obtained through rating index analysis and error detection rate analysis: (1) The external forcing factors is embedded in the phase space local prediction model of solar terms, which has greatly improved the accuracy of the original prediction model and expanded the prediction model of non-stationary time series. (2) Based on the selection of QC parameters, the SFA-WZLM method has higher sensitivity than the WZLM, SFA-WFLM, and WFLM methods, and can better mark suspicious data. (3) The error detection rate of the four methods at six stations in the country is analysed. Compared with other methods, the SFA-WZLM method has the highest error detection rate and the error detection rate is higher than 80%. Some of the selected stations are distributed on plateaus, mountains, and cold regions, indicating that they have strong stability and applicability.

Acknowledgments
This work was supported by 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). Science and technology major project of China State Shanghai Railway Group Co., Ltd(201904).

References
[1] Lakshmanan V, Karstens C, Krause J and Tang L 2014 Quality control of weather radar data using polarimetric variables J. Atoms. Ocean. Technol. 31 1234-1249.
[2] Ren Z H, Zhao P, Zhang Q, Zhang Z F, Cao L J, Yang Y R, Zhou F L, Zhao Y F, Zhao H M and Chen Z 2010 Quality control method applicable to the hourly precipitation data of automatic stations in China Meteorol. 36 123-32.
[3] Zhao Q, Zhu Y L, Ding S W and Yu Y F 2018 Research on the Data-Driven Quality Control Method of Hydrological Time Series Data Water. 10.
[4] Wiskott L and Sejnowski T J 2002 Slow Feature Analysis: Unsupervised Learning of Invariances Neural. Comput. 14 715-70.
[5] Shi T, Fei M, Lu J X, Lu J J, Pan Y, Zhou C, Wu J Z and Zheng J Y 2019 Phase Space Reconstruction Algorithm and Deep Learning-Based Very Short-Term Bus Load Forecasting Energies 12 4349.
[6] Floris Tankens 1981 Detecting strange attractors in turbulence Dyn. Sys. Turbul. 898 366-381.
[7] Kim H S, Eykholt R and Salas J D 1999 Nonlinear dynamics, delay times and embedding windows Phys. D 127 48-60.
[8] Wang Y, Wang J Z and Li Z W 2020 A novel hybrid air quality early-warning system based on phase-space reconstruction and multi-objective optimization: A case study in China. J. Cleaner. Prod. 260 121027.
[9] Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models part I—Adiscussion of principles J. Hydrol. 10 282-290.
[10] Hubbard K G and You J 2005 Sensitivity analysis of quality assurance using the spatial regression approach—A case study of the maximum/minimum air temperature J. Atoms. Ocean. Technol. 22 1520-1530.