IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 3, 2019, accepted November 27, 2019, date of publication December 4, 2019,
date of current version December 17, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2957602*

# An Adaptive Outlier Detection and Processing Approach Towards Time Series Sensor Data

**MINGHU ZHANG[1,4], (Student Member, IEEE), XIN LI[2,3,4], (Senior Member, IEEE), AND LILI WANG[5]**

[1]Key Laboratory of Remote Sensing of Gansu Province, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
[2]National Tibetan Plateau Data Center, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
[3]CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing 100101, China
[4]Northwest Institute of Eco-Environment and Resources, University of Chinese Academy of Sciences, Beijing 100049, China
[5]College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China

Corresponding author: Xin Li (lixin@lzb.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0500106, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA20100104, and in part by the 13th Five-Year Informatization Plan of the Chinese Academy of Sciences under Grant XXH13505-06.

**ABSTRACT** The intelligent environment monitoring network, as the foundation of ecosystem research, has rapidly developed with the ever-growing Internet of Things (IoT). IoT-networked sensors deployed to monitor ecosystems generate copious sensor data characterized by nonstationarity and nonlinearity such that outlier detection remains a source of concern. Most outlier detection models involve hypothesis tests based on setting outlier threshold values. However, signal decomposition describes stationary and nonstationary relationships sensor data. Therefore, this paper proposes a three-level hybrid model based on the median filter (MF), empirical mode decomposition (EMD), classification and regression tree (CART), autoregression (AR) and exponential weighted moving average (EWMA) methods called MF-EMD-CART-AR-EWMA to detect outliers in sensor data. The first-level performance is compared to that of the Butterworth filter, FIR filter, moving average filter, wavelet filter and Wiener filter. The second-level prediction performance is compared to support vector regression (SVR), K-nearest neighbor (KNN), CART, complementary ensemble EEMD with CART and AR (EEMD-CART-AR) and ensemble CEEMD with CART and AR (CEEMD-CART-AR) methods. Finally, EWMA is compared to Cumulative Sum Control Chart (CUSUM) and Shewhart control charts. The proposed hybrid model was evaluated with a real dataset from the hydrometeorological observation network in the Heihe River Basin, yielding experimental results with better generalization ability and higher accuracy than the compared models, and providing extremely effective detection of minor outliers in predicted values. This paper provides valuable insight and a promising reference for outlier detection involving sensor data and presents a new perspective for detecting outliers.

**INDEX TERMS** Environmental monitoring, sensor data, outlier detection, integrated model, statistical analysis.

## I. INTRODUCTION

The intelligent environment monitoring network consists of numerous sensor devices that form a ubiquitous, reliable and distributed internet of things (IoT) network for sensing and communicating which is gradually driving the evolution of ecosystem research, and massive amounts of time series sensor data have been collected [1]–[3]. According to a published report, the total amount of global Earth monitoring data is increasing exponentially each year and the IDC report predicts that global data might be excepted to reach 163 ZB approximately by Sen and Jayawardena [4]. Outlier detection in bulk sensor-collected data has been a matter of great concern and major challenge. In particular, devices deployed in high altitude and harsh regions often generate spatiotemporal variations in networked sensor data [5], [6]. In addition, the sensed data are largely affected by the environment of the underlying surface of the atmosphere in cold and arid regions, where high uncertainty can be caused by local climate change with non-stationary and nonlinear characteristics [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Heng Wang.

Data outliers have posed a considerable challenge for scientific research. It is of practical significance and importance to develop a suitable outlier detection approach for sensor data.

A sensor data outlier is defined as an observed value that is far from others. Outlier detection focuses on the process of discovering data deviations [8]–[10]. In fact, outlier detection and processing play vital roles in identifying abnormal patterns and have been applied in many different fields, such as process control [11], environmental monitoring [12] and traffic monitoring [13]. Many existing detection methods based on hypothesis tests setting the threshold values of outliers have been proposed to identify outliers through uniformly inspecting the main characteristics of a set of objects [14], including distance-based methods [15], K-nearest neighbor (KNN) methods and prediction-based methods [16]. The autoregressive (AR) model, autoregressive moving average (ARMA) model, and autoregressive integrated moving average (ARIMA) model based on statistics were used to detect outliers in complicated multivariate sensor data involving single-variable time series [17]–[19].

Similarly, the physical, statistical, and machine learning models that have been developed to detect outliers are not sufficiently capable of analyzing non-stationary data [20]–[23]. Signal decomposition is a processing method that describes stationary and non-stationary relationships. This approach decomposes non-stationary sensor data into stationary data and retains the structure of the raw data. Therefore, to solve the problems noted above, some signal processing methods, such as empirical mode decomposition (EMD), complementary ensemble EMD (CEEMD), ensemble EMD (EEMD), variational mode decomposition (VMD) and wavelet transform (WT), have been widely applied to recursively decompose data into different intrinsic modes and improve the effectiveness of outlier detection [24]–[26]. To a large extent, signal processing methods have a limited capacity to improve the performance and accuracy of a detection model. Therefore, researchers have extended these methods, for instance, a hybrid model based on EMD and AR aimed at transforming data from the time domain to the frequency domain was successfully applied for outlier detection to assess the construction and precisely track the frequency of signals [27]. WT provides a high temporal resolution in the high-frequency range of a time series signal. However, WT in outlier detection has led to shortcomings in analyses of big data, and WT is time consuming compared to existing models [28].Similarly, researchers have extended the applications of EMD to process sensor data with non-stationary due to its prominent advantages [29].

Although single, hybrid and combined methods have achieved some success, the existing approaches have not achieved exceptional performance. Considering the above shortcomings, a high-level outlier detection model called the MF-EMD-CART-AR-EWMA model is presented for outlier detection in this paper. Of this model, a three-level ensemble method is leveraged, where MF is used as the preprocessor to preliminarily screen a series data that contains outliers, such as large sudden changes. EMD is chosen due to its flexibility in processing nonstationary data. CART with the AR method are employed as the base learner for the prediction task, and then we use the EWMA control chart to detect outliers. The proposed outlier detection model is designed with a black-box scenario in mind. We define that outliers are deviation-based or significant changes in time-series sensor data. Specifically, the outliers that deviate from the upper (UCL) and lower (LCL) control limits of EWMA can be addressed for further investigation, while implementing replacement with the prediction value.

The ultimate objective of the proposed approach is to provide a highly accurate and robust outlier detection model to overcome the challenges of large-scale sensor data. The model proposed in this paper aims at not only detecting outliers but also processing the outliers so that an improved dataset is obtained. To investigate and evaluate the performance of the model, the proposed method was thoroughly evaluated and benchmarked based on real sensor data from the hydrometeorological observation network in the Heihe River Basin.

The primary contributions of the proposed model are summarized as follows.

(a) One-step-ahead preprocessing for identifiable outliers

Preprocessing is the first level of the proposed model for original data series analysis. In this step, the original data with obvious outliers, such as sudden extremes, are be processed. We aim to address various real-world sensor data outlier challenges using MF, thereby eliminating these patterns before the outlier analysis and modeling steps.

(b) Developing the EMD-CART-AR approach for second-level prediction.

EMD is used to decompose the preprocessed data into new and stationary intrinsic mode functions (IMFs) with different features, and the CART and AR models are employed considering the characteristic scales of decomposed subsequences, which can promote the accuracy of the prediction model.

(c) Using an EWMA control chart to detect outliers in predicted data.

EWMA is introduced as the last model level based on the aforementioned first two levels for identifying the minor outliers in the predicted data. Taking advantage of the control parameters, the entire iterative process of the model can be effectively regulated.

(d) Applying comprehensive statistical indicators to evaluate the performance of the proposed model.

The proposed approach is applied to real-world data sets, and the results are evaluated with statistical indicators. The test includes four sets of data from the hydrometeorological observation network dataset. The results are also compared to those of other models, including SVR, KNN, CART, CEEMD-CART-AR, and EEMD-CART-AR, to assess the preprocessing and prediction performance of the proposed model.

The paper is organized as follows. In section 2, the framework, main implementation steps and employed

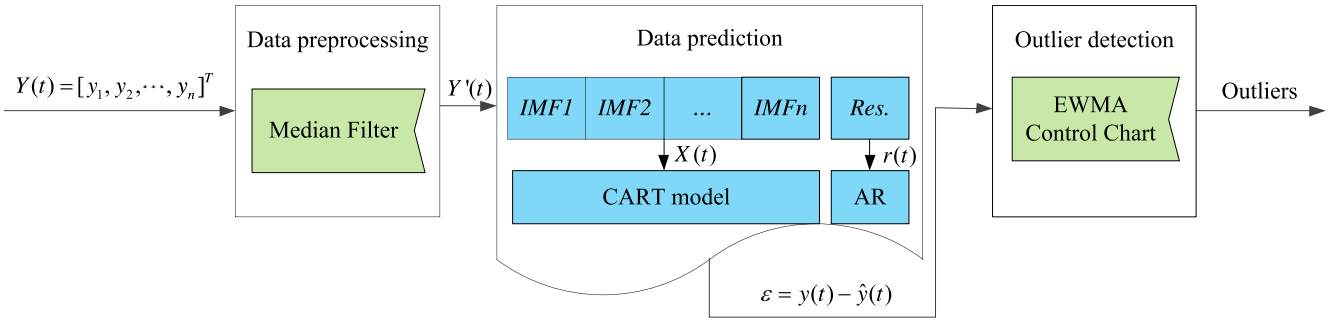


**FIGURE 1.** Schematics of the three-level hybrid model.

methodology are given. In the section 3, the data description and analysis are presented. In section 4, the evaluation criteria used in this paper and the experimental results and discussion are introduced. Finally, a brief conclusion is made.

## II. IMPLEMENTATION METHOD AND SCHEMATICS

### A. SCHEMATICS OF THE THREE-LEVEL HYBRID MODEL

In this section, the adaptive outlier detection modeling approach is established for outlier detection in real-world data set, and the schematics of the three-level hybrid model are shown in Fig.1. The three levels are placed at different positions and have specific functions. The preprocessing level is the first level, and it preprocesses the original data that may be influenced by obvious outliers, such as large or small sudden variational patterns. The EMD-CART-AR level, as the second level located between the preprocessing and outlier detection levels, is a predictive model that provides input data for outlier detection. The final level, the EWMA detector, identifies possible minor outliers in the predictive output and is used to adjust the iterative procedure of the model.

The main steps of the model are as follows.

Step one: Conduct a preliminary data test on $Y(t) = [y_1, y_2, \cdots, y_n]^T$, and then preprocess the result with MF. The preprocessed data are recorded as $Y'(t)$.

Step two: Decompose the preprocessed data $Y'(t)$ into $X(t)$ and $r(t)$ with EMD and record it the result as $Y'(t) = X(t) + r(t)$, where $X(t) = [x_1, x_2, \cdots, x_n]^T$ represents the high-frequency terms, $r(t) = [r_1, r_2, \cdots, r_n]^T$ is the trend term, and $n$ is the sample size.

Step three: Predict $X(t)$ and $r(t)$ with the CART and AR models, respectively. Predict the high-frequency terms with the CART model, and record the result $\hat{x}(t) = [x_1, x_2, \cdots, x_n]^T$. The trend term, $r(t)$, is predicted with the AR model, and the result is recorded as $\hat{r}(t) = [r_1, r_2, \cdots, r_n]^T$, where $n$ is the sample size. The final predicted value is denoted as $\hat{y}(t) = \hat{x}(t) + \hat{r}(t)$.

Step four: Compare the real and predicted values, and calculate the residual sequence, namely $\varepsilon = y(t) - \hat{y}(t)$.

Step five: Detect the outliers with the EWMA control chart, which is also used to control the entire iterative process of the model.

Last step: Process the outlier data with the proposed model and obtain clean data through the iteration and reconstruction of the proposed model.

### B. METHODS

#### 1) MEDIAN FILTER (MF)

The MF is an algorithm based on statistical theory to suppress noise in nonlinear signal processing [30]. The basic principle of this algorithm is to replace the value of a point in a sequence with the median value of each point in the neighborhood to eliminate the isolated noisy points. Suppose data series $X(m)$ is a signal written as $X(m) = [x_1, x_2, \ldots, x_m]$,where $m$ is the size of the series.The time window length of the MF is defined as $n$. The process for the $j_{th}$ point is to take $n$ samples centered on the $j_{th}$ point as the input values, reorder them by size, and generate a new data sequence($X_{j-\frac{n-1}{2}}, \cdots X_j, \cdots, X_{j+\frac{n-1}{2}}$).The median value $X_j$ is selected as the output of the filter. n is typically an odd number, and if n is an even number,the output value will be the mean of the two sample values at the middle position.

#### 2) EMPIRICAL MODE DECOMPOSITION (EMD)

EMD was proposed by Huang et al. and is a new signal processing method for decomposing a signal into IMFs [31], [32]. The algorithm refers to the smooth processing of a signal and subsequent decomposition of a non-stationary signal into a stationary series with functions of different characteristic scales, each called an IMF [33]. The IMF must satisfy two conditions. First, in the whole data series, the number of extreme points must be the same as the number of zero-crossing instances, or the difference between these two values must be not greater than 1. Second, the data series must be locally symmetric about the time axis, namely, the local mean is zero at any time point.

The main processing steps in the model are as follows.

Step one: Find all the maximum and minimum points in$X(t)$(the original signal), and fit two envelope curves with the cubic spline interpolation function method.

Step two: Find the mean $m(t)$ of the upper envelope and the lower envelope.

Step three: Subtract the $m(t)$ mean by the original series to obtain the new series $c(t)$,namely $c(t) = X(t) - m(t)$.

Step four: Determine whether $c(t)$ meets the IMF conditions: if the conditions are met, separate $c(t)$ and obtain the remainder $r(t)$, namely, $r(t) = X(t) - c(t)$; if the conditions are not met, take $c(t)$ as the new signal, and repeat Step one to Step three until the conditions are met.

Step five: Take the obtained $r(t)$ as the new original series and repeat Step one to Step four. Finally, obtain finite IMF components and a trend component.

After the process above is implemented, the signal with random non-stationarity is decomposed into several stationary IMF components and a trend component, as shown in Eq. (1).

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{1}$$

In Eq. (1), $c_i(t)$ refers to the $i_{th}$ IMF component, representing the signal components with different characteristic scales in the original signal$x(t)$, and $r$ refers to the trend component, reflecting the trend of the original signal $x(t)$. Therefore, the signal $x(t)$ can be decomposed into $n$ stationary components (IMFs) with different characteristic scales and a trend term.

### 3) CLASSIFICATION AND REGRESSION TREE (CART)

As a typical classification algorithm, the CART method is a supervised non-parametric classification method that creates a binary tree based on a simple model and easily implemented extraction rules to obtain predictions [34]. The CART algorithm has been widely applied in classification and prediction tasks [35]. The properties of the root node of the data are first found according to the Gini index, and a tree is created from the top to the bottom in a recursive manner until every sample established after division is pure. The leaf nodes of the decision tree represent the categories of information associated with the sample, and each path along a branch from the root node to the leaf node represents a rule. A complete binary tree refers to a rule set. Essentially, the decision tree classifies data with a series of rules. The main decision trees are binary branched trees and multibranch trees, and the former is used in this research because of its search flexibility.

The following concepts were used to construct the CART. For all the sample data, a tree with many levels and leaf nodes is created to fully reflect the relations among the data (at this moment, the data relations reflected by the tree are often influenced by overtraining). Through trimming the tree, a series of subtrees is created, from which the trees of appropriate size are selected to classify the data.

The main process of the model is as follows.

Step one: Input the training dataset D.

Step two: Output the CART $f(x)$.

In the input space of the training dataset, divide every region into two subregions recursively and determine the output value of each subregion to create the corresponding binary decision tree.

1) Choose the optimal segmentation variable $j$ and segmentation point $s$, and solve Eq.(2) as follows.

$$\min_{js}[\min_{c_1} Loss(y_i, c_1) + \min_{c_2} Loss(y_i, c_2)] \tag{2}$$

Traverse $j$ and scan $s$ for the fixed segmentation variable $j$; then, obtain the minimum pair $(j, s)$ through Eq.(2).

2) Divide the region with the chosen $(j, s)$, and determine the corresponding output value, as shown in Eq.(3),

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i, \quad x_i \in R_m, = 1, 2 \tag{3}$$

3) Continue to repeat Steps one and two until the stopping condition is met.

4) Divide the input space into $M$ regions, $R_1, R_2, \cdots R_m$, and generate the decision tree, as shown in Eq.(4).

$$f(x) = \sum_{m=1}^{M} \hat{c}_m I(x \in R_m) \tag{4}$$

### 4) EXPONENTIAL WEIGHTED MOVING AVERAGE (EWMA) CONTROL CHART

The EWMA control chart as a prediction-based detector is introduced in this work, and it presents a robustness for detecting minor outliers compared with the traditional control chart, e.g., Shewhart control chart and Cumulative Sum Control Chart (CUSUM) control chart. The EWMA chart proposed by Roberts in 1959 assigns the maximum weight to the nearest observed value [36]. Due to the flexibility and reliability of the EWMA control chart for monitoring the small shifts in parameters, this control chart has been applied widely [37]. The Shewhart control chart yields omission of minor outliers among slight fluctuations aspects. The CUSUM control chart has better performance than the Shewhart control chart in terms of detecting slight fluctuations. However, for CUSUM, the two adjacent statistics have a strong correlation, in fact, there is only one sample difference. When the mean and variance of the sample cannot be accurately estimated, the analysis effect is weakened. However, the EWMA control chart is flexible and has a strong detection ability for small fluctuations and gradual drifts. Compared with traditional outlier detection methods, the EWMA control chart provides excellent performance in identifying small fluctuations and slow shift processes; therefore, it is highly suitable for outlier detection based on prediction [38].In particular, the outlier detection was driven by the desire to present a robustness as much as possible and to allow accurate detection in time-series sensor data [39]. Therefore, we proposed an adaptive outlier detection tightly coupled to the prediction-based estimator to detect minor outliers and close the detection iterations.The EWMA control chart employed in the MF-EMD-CART-AR model is to detect possible minor outliers in prediction process, while is used to regulate model iterations.

The EWMA control chart can be expressed as shown in Eq.(5),

$$Z_i = \lambda X_i + (1-\lambda)Z_{i-1} \tag{5}$$

where the $\lambda$ is a constant constrained by $0 < \lambda \leq 1$ and $X_1, X_2, \cdots, X_n$ compose a sample of observed values. The target value of the process is usually taken as the initial value $Z_0 = \mu$. Alternatively, the mean of the initial data can serve as the initial value, namely, $Z_0 = \bar{X}$.

If the observed value $X_i$ is an independent random variable with the same variance $\sigma^2$, then the variance of $Z_i$ is as shown in Eq.(6).

$$\sigma_{Z_i}^2 = \sigma^2 \frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}] \tag{6}$$

Therefore, the EWMA control chart is constructed with a monitoring index based on the $Z_i$ statistics, and the upper and lower control limits are shown in Eq. (7).

$$(UCL, LCL) = \mu \pm L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]} \tag{7}$$

where $L$ refers to the regulatory factor selected to ensure that the expected $ARL_0$ can be achieved. As $i$ increases, the control limits will converge to $\mu \pm L\sigma\sqrt{\frac{\lambda}{2-\lambda}}$. The process parameters of the EWMA control chart are $L$ and $\lambda$. Hence, detailed research has been conducted on the $ARL$ properties of the EWMA control chart with different design parameters. Generally, when $0.05 \leq \lambda \leq 0.25$ [37], the EWMA control chart provides excellent detection performance. According to practical experience, the value of $\lambda$ is generally relatively small to make the control chart flexible and effective.

## III. EXPERIMENT AND ANALYSIS

### A. DATASETS

In this section, the sensor data from the hydrometeorological observation network in the Heihe River Basin, an endorheic basin located in the arid and semiarid regions of Northwest China [40]–[42], are used to verify the accuracy and robustness of the proposed model. The hydrometeorological observation network currently transmits approximately 200,000 recorded values per day collected from sensor devices, such as temperature and humidity sensors, wind speed and direction sensors and soil moisture sensors. Moreover, changing seasonal factors result in non-stationary and nonlinear characteristics in the sensor data. Therefore, we employ four sets of data from different sites and with different collection times and sample sizes. These datasets are independently used to evaluate the proposed approach.

To evaluate the generality of the proposed prediction model, for each experimental case, we evaluated two kinds of sample sizes; 7-day temperature and humidity sensor data samples (1008 data points) from the Daman superstation and 10-day data (1440 data points) from the Arou superstation were obtained for different time periods. Then 80 % of the data are randomly selected for training the EMD-CART-AR model. The remainder of the data is used as test sets to evaluate the performance of the proposed model. The locations of the Daman and Arou superstations can be seen in Fig.2.

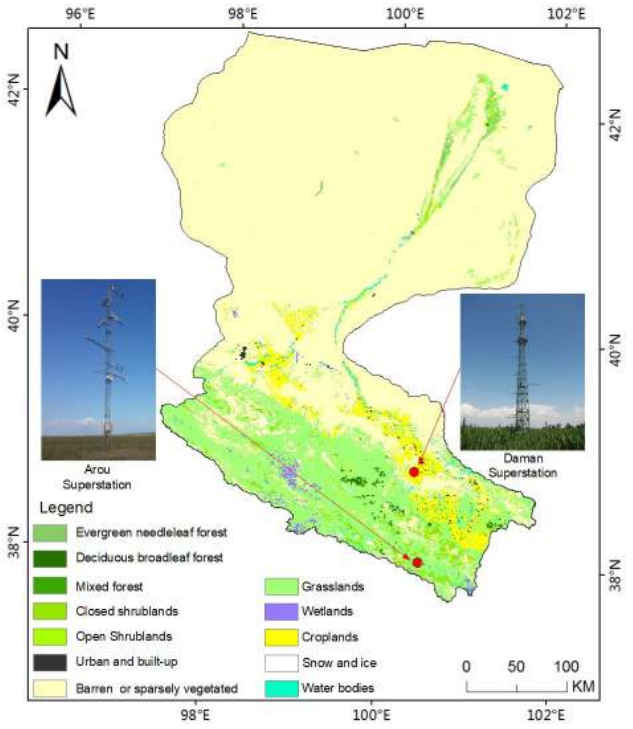


**FIGURE 2.** Locations of the Daman and Arou superstations.

#### 1) DAMAN SUPERSTATION DATASET

Daman superstation (Altitude is 1556 m; 100.3722$E$, 38.8555$N$) is located in the Dagan Irrigation District of Wuxing Village, Xiaoman Town, Zhangye City, Gansu Province, China, and consists of a meteorological element gradient observation system, an eddy-covariance system, 2 large-aperture scintillometers, a lysimeter, a cosmic-ray soil moisture observation system and nine soil moisture wireless sensor network nodes. The temperature and humidity data from Daman superstation dataset were selected from May 5 to 11, 2018, and from December 31, 2016, to January 6, 2017.

The recurrence plot (RP) is an important method to analyze the periodicity, chaos and nonstationarity of time-series data. Specifically, RP depicts black and white points on the time plane of the square, where the black points represent the occurrence of recursion in the corresponding state of the horizontal and vertical axis on the coordinate, while the white point indicates that no recursion occurs [43].Therefore, the RP can be used to analyze the nonstationary and nonlinear characteristics of a time-series data. For a stationary time series, the corresponding RP is uniformly distributed, and the RP of a nonstationary time series is nonuniformly distributed. The RP of temperature and humidity sensor data of Daman superstation dataset is given in Fig.3, According the figure, Fig. 3-a and Fig. 3-b exhibit a significant difference. It can

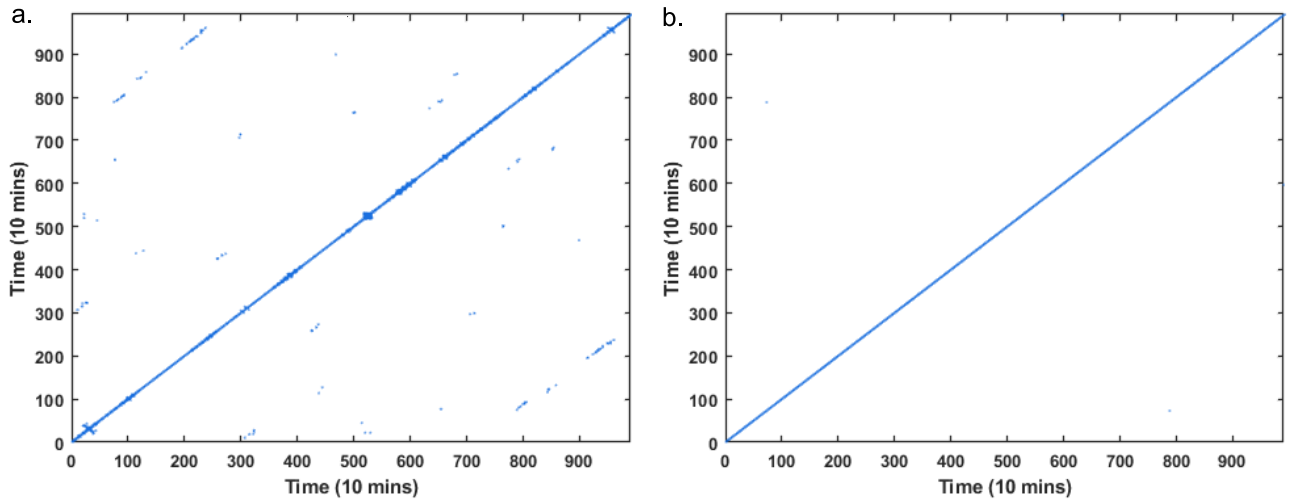


**FIGURE 3.** Recurrence plots of sensor data from Daman superstation: a. temperature data and b. humidity data.

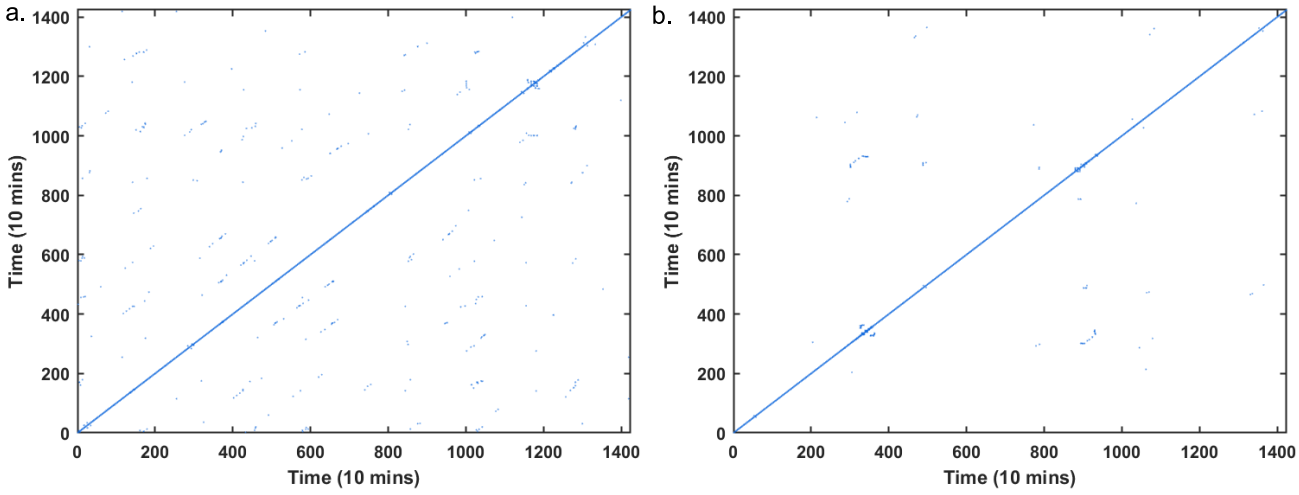


**FIGURE 4.** Recurrence plots of sensor data from Arou superstation: a. temperature data and b. humidity data.

be found in Fig. 3-a that RP has large white or blue points, which indicate that the time-series data has a large mutation during this period, and the data are in a relatively stable state of a period of time before and after the sudden change, that is, a stable state. In Fig. 3-b, the nonuniform characteristics of the data are relatively weak with respect to Fig. 3-a.

### 2) AROU SUPERSTATION DATASET

Arou superstation (Altitude is 3033 m; 100.4572$E$, 38.0384$N$) is located in Arou Village, Qilian County, Qinghai Province, China (Che et al., 2019), and consists of a meteorological element gradient observation system, an eddy-covariance system, 2 large-aperture scintillometers, a weighing-type rain gauge, a vegetation phenology observation system, a cosmic-ray soil moisture observation system and 16 soil moisture wireless sensor network nodes. Due to the high altitude of the location, low average annual temperature and poor observation conditions at Arou, outliers are common in the sensor data collected from Arou superstation. To further verify the robustness and applicability of the model, an experiment was conducted on the temperature and humidity data collected from Arou, and the samples were selected from November 1 to 10, 2017, and from September 6 to 16, 2017.

Similarly, Fig.4 shows the RPs of temperature and humidity at the Arou superstation. According to the figure, the nonstationarity of the temperature data is obvious, and the humidity data are weakly nonstationary. The nonuniform distributions of the temperature and humidity data RPs further reflect the nonstationary characteristics of the sensor data. In general, from the RPs analysis, it suggested that the experimental data has obvious nonstationary characteristics.

The MF is used first to preprocess the obviously visible outliers in the raw data, and the EMD-CART-AR model is

then employed for prediction. Finally, the EWMA method is used to detect the outliers. The detailed data outlier detection results are presented in the next section.

### B. EXPERIMENTAL RESULTS

#### 1) PARAMETER SETTING

In each experiment, all the data are first preprocessed with the MF. The filter window length of the MF for preprocessing needs to be adjusted according to the characteristics of the data. Here, we chose filter windows with different lengths to assess the performance of data preprocessing [44].Moreover, the scheme of partition of time-series data over a sequence of temporal windows via a time window is shown in Fig 5. It can be seen that 1 to k-1 from the first subset can be chosen to train the model, 2 to k from the same subset are selected for prediction by using the trained model. After several adaptive iteration processes of the model, the model can mitigate interference and noise effects and became sufficiently stable. The parameters of EMD are obtained by employing the stopping criteria [45]. Grid searches are adopted to optimize the CART parameters and provide maximum prediction accuracy [46]. The parameters of the AR model are defined according to the autocorrelation coefficient and partial correlation coefficient of the sample data. The $(\lambda, L)$ values of EWMA are considered based on a confidence level of 99.97% [37].For all the methods, detailed parameter settings are described in Table 1.

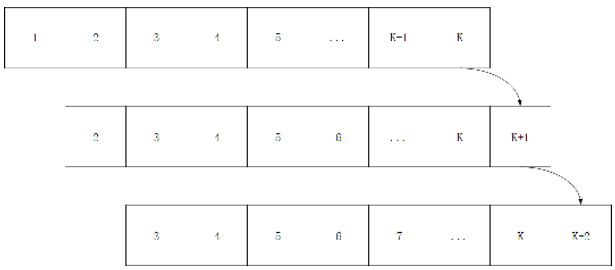

**FIGURE 5.** The time-window scheme of training dataset and testing dataset selection.

**TABLE 1.** Experimental parameter of all methods.

| Parameters | Daman superstation | | Arou superstation | |
|---|---|---|---|---|
| | Temperature | Humidity | Temperature | Humidity |
| W | 5 | 5 | 5 | 5 |
| IMF | 6 | 7 | 9 | 7 |
| Max_dept | 5 | 5 | 7 | 7 |
| p | 4 | 5 | 4 | 5 |
| $(\lambda, L)$ | (0.25, 9) | (0.25, 9) | (0.25, 9) | (0.25, 9) |

#### 2) MF PREPROCESSING RESULTS

This section presents the proposed preprocessing procedure focusing on the first level of an outlier detection model, with the aim of preliminarily screening a series data that contains outliers, such as large sudden changes. To address various real-world data outlier challenges, these outlier data should be eliminated before outlier analysis and modeling. In this context, the MF is used to preprocess the visual outlier data in $Y_{DamT}(t)$, $Y_{DamH}(t)$, $Y_{ArouT}(t)$ and $Y_{ArouH}(t)$, where these data series are selected from the temperature and humidity datasets from Daman superstation and Arou superstation. To highlight the advantages of the MF in processing non-stationary data, several outliers are randomly added to the historical temperature and humidity data. In practical applications, unprocessed historical data can also be assessed by outlier detection models.

The results of the data preprocessed by the MF are shown in Fig.6 and Fig.7, in which the red curve refers to the preprocessed data,the blue curve refers to the raw data and the hollow circles are outliers. The obvious discernible outliers that are too high or too low are processed, and the red curve almost coincides with the blue curve. The results confirm that the scheme used in this paper yields high accuracy. This finding suggests that the MF is suitable for the outlier processing of sensor data with the capability for fusing, denoising and smoothing to a certain extent. Notably, the MF is a nonlinear smoothing technique with a selection adjustment scheme based on a filter window, and the value of each data point is set as the median of all data points in a certain neighborhood window for that data point. As a result, the outlier value in a data series is replaced by the median value of the neighborhood window.

The preprocessed data are recorded as $Y'_{DamT}(t)$, $Y'_{DamH}(t)$, $Y'_{ArouT}(t)$ and $Y'_{ArouT}(t)$.

#### 3) EMD DECOMPOSITION RESULTS

In this section, the temperature and humidity sensor data can be regarded as a time series signal, and the EMD method is introduced to decompose the preprocessed data series, i.e. $Y'_{DamT}(t)$, $Y'_{DamH}(t)$, $Y'_{ArouT}(t)$ and $Y'_{ArouT}(t)$. Fig.8 shows that $Y'_{DamT}(t)$ decomposed by EMD comprises 6 IMF components $X_{DamT}(t) = IMF_i(i = 1, 2, \cdots, 6)$ and a trend term $r_DamT(t)$. To obtain relatively stationary original data and a locally stationary trend, the $IMF_i$ can be reconstructed by $X(t)_{DamT} = \sum_{i=1}^{6} IMF_i.X_{DamT}(t)$ displays an undulation trend similar to that of $Y'_{DamT}(t)$. Similarly, the high-frequency term $X(t)_{DamH} = \sum_{i=1}^{7} IMF_i$ and the trend term $r_{DamH}(t)$ are obtained by reconstructing the decomposed humidity data with the EMD method. $Y'_{ArouT}(t)$ and $Y'_{ArouH}(t)$ are also decomposed by the EMD model. $Y'_{ArouT}(t)$ is decomposed into 10 components, including 9 IMFs, namely, $IMF_i(i = 1, 2, \cdots, 9)$, and one trend term $r_{ArouT}(t)$. Similarly, $Y'_{ArouH}(t)$ is decomposed into 10 components, including 9 IMFs, namely, $IMF_i(i = 1, 2, \cdots, 9)$, and 1 trend term $r_{ArouH}(t)$, which are given in Fig.9 To obtain relatively stable original data and the partial stationary trend, the IMFs in $IMF_i$ are reconstructed. $X_{ArouT}(t) = \sum_{i=1}^{7} IMF_i$ and $X_{ArouH}(t) =$

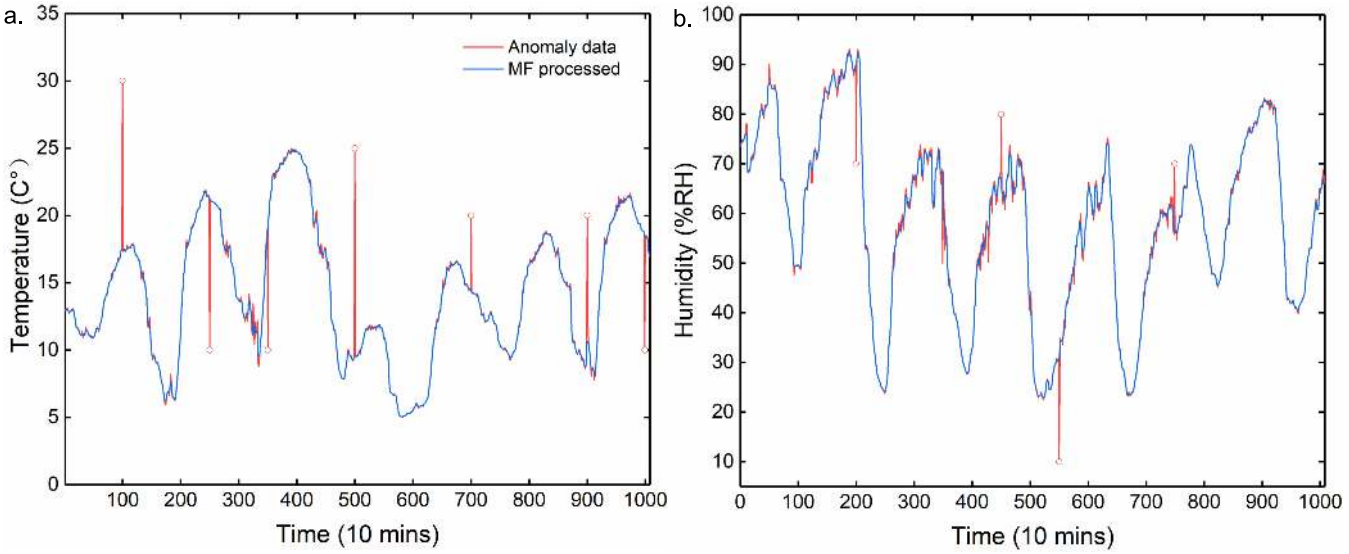


**FIGURE 6.** Results for the temperature and humidity data from Daman superstation preprocessed by the MF.

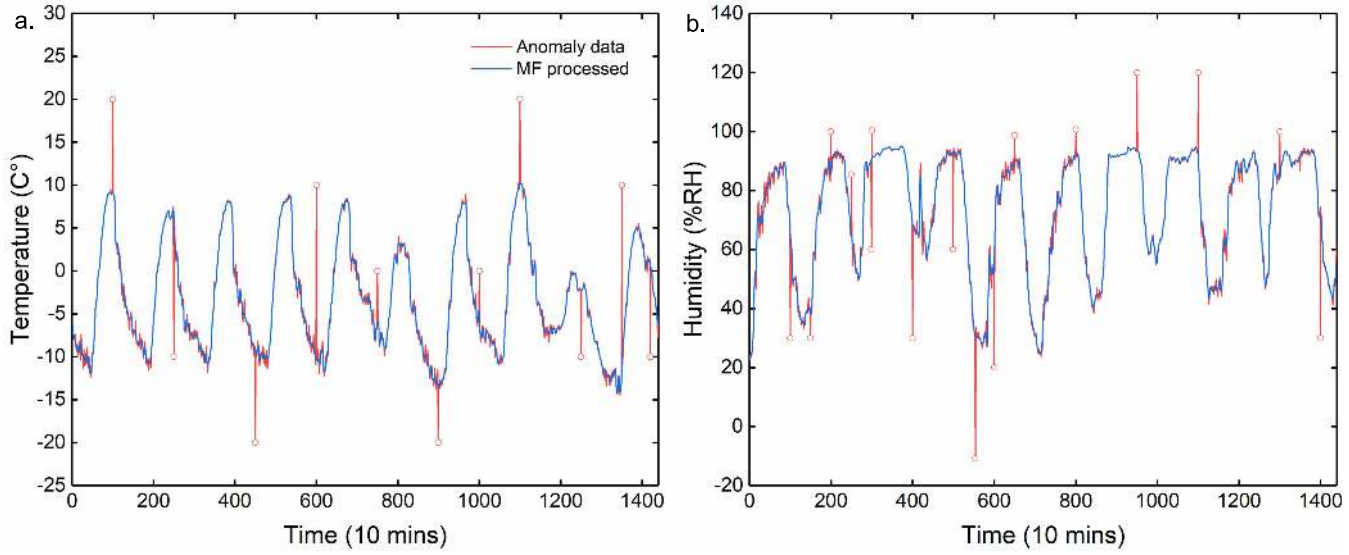


**FIGURE 7.** Results for the temperature and humidity data from Arou superstation preprocessed by the MF.

$\sum_{i=1}^{7} IMF_i$ are recorded as the high-frequency terms, and $r_{ArouT}(t)$ and $r_{ArouH}(t)$ are recorded as the trend terms.

The basic concept of employing EMD for predictions involves decomposing sequence data into IMF components and trend terms. The separated trend terms at different scales can reduce the complexity of the time series, and the divided IMF components are able to maintain the unique physical meaning and stationarity of the data [47]. Thus, EMD is able to improve the prediction accuracy in specific time horizons based on this approach.

Processing small-sample time series data with the CART model is effective. Therefore, the CART model is used to predict the high-frequency terms $X_{DamT}(t)$, $X_{DamH}(t)$, $X_{ArouT}(t)$ and $X_{ArouH}(t)$. As the most common analysis model for time series, the AR model, which is characterized by simplicity and high accuracy, is ideally qualified for predictions involving locally stationary data, such as $r_{DamT}(t)$, $r_{DamH}(t)$, $r_{ArouT}(t)$ and $r_{ArouH}(t)$. The detailed data processing results are presented in the next section.

#### 4) EMD-CART-AR PREDICTION RESULTS

In the CART prediction model, the curve smoothness and error degree are taken into consideration for the prediction of the nonlinear data series [48].Therefore, the high-frequency terms $X_{DamT}(t), X_{DamH}(t), X_{ArouT}(t)$ and $X_{ArouH}(t)$ are predicted by the CART model. The AR model can be used for time series prediction and analysis to the trends

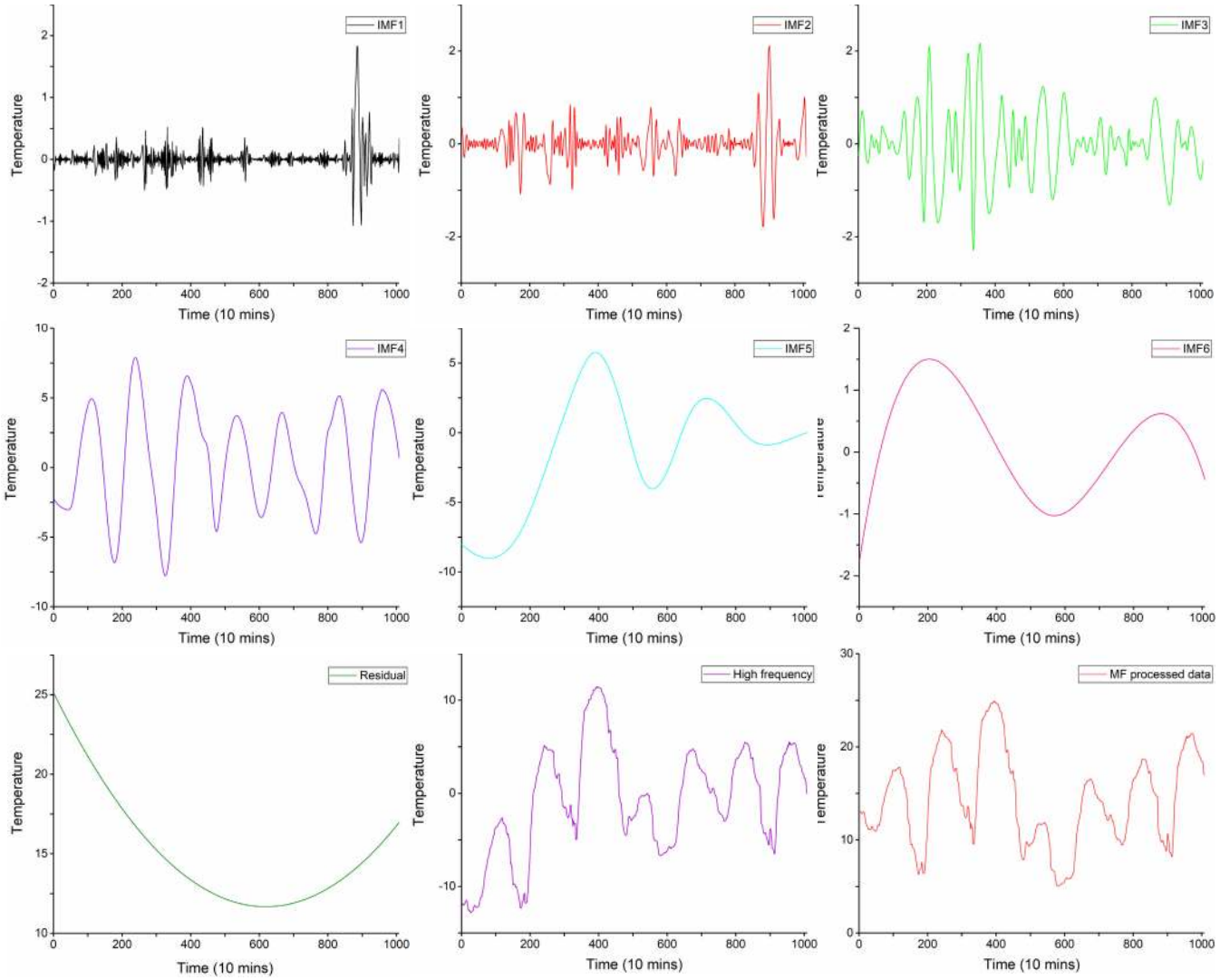


**FIGURE 8.** Results for temperature data from Daman superstation decomposed by EMD.

of dynamic data. The model quantitatively analyses linear data correlations and predicts future values [49]. Therefore, the trend terms $r_{DamT}(t)$, $r_{DamH}(t)$, $r_{ArouT}(t)$ and $r_{ArouH}(t)$ are predicted by the AR model. Performance comparisons of EMD-CART-AR based on the temperature and humidity data from the Daman and Arou superstations are presented in Fig.10 and Fig.11, in which the red curve represents the predicted values of EMD-CART-AR and the blue curve represents the preprocessed data from the MF. The comparison shows that the predicted and real values are almost coincident. The minor disagreement between the real and predicted values is reasonable. To address the performance of the proposed model, the Pearson correlation coefficients calculated between the predicted and real values of temperature and humidity at Daman superstation are 0.9995 and 0.9996, respectively; similarly, the results for the temperature and humidity at Arou superstation are 0.9995 and 0.9996, respectively.

The experimental results indicate that the EMD-CART-AR hybrid model proposed in this paper reduces the prediction error effectively and demonstrates excellent prediction ability in terms of processing the non-stationary time series problem.

### 5) EWMA OUTLIER DETECTION AND PROCESSING RESULTS

As noted earlier, the EWMA control chart is not affected by the mean value of a dataset and is widely used in the processing of time series data; additionally, the random error conforms to a normal distribution with a mean value of and variance of $\delta^2$ [50]. The robust EWMA control chart employed in this section involves detecting outliers and identifying minor errors in the residual series. The proposed detection model architecture and EMWA approach considered in this paper aim not only to control the reasonable error range but also to effectively regulate the entire iterative process of the model and achieve continuous detection and processing.

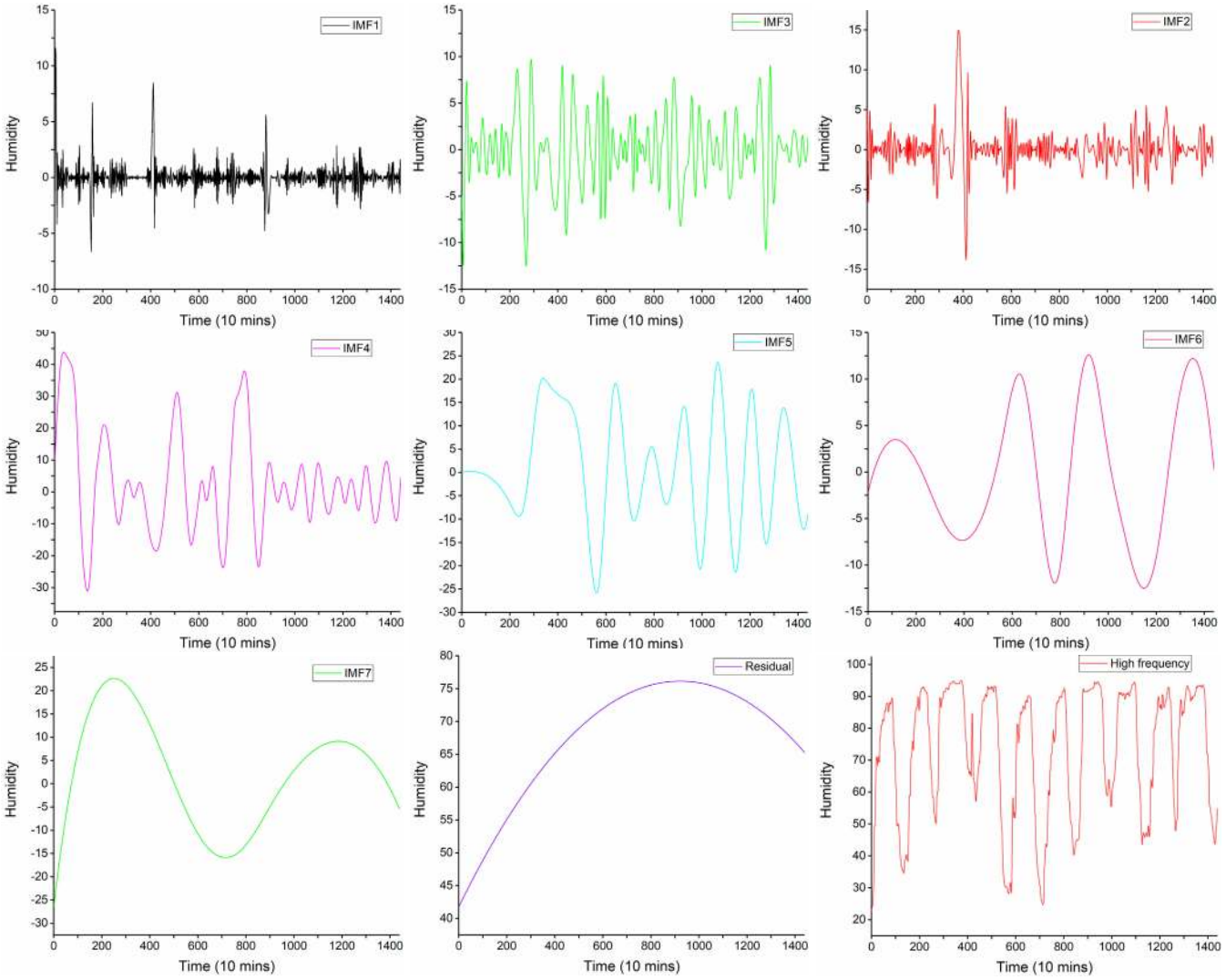


**FIGURE 9.** Results for humidity data from Arou superstation decomposed by EMD.

At the outlier detection and processing stage, the UCL and LCL control limits of the EWMA control chart of the four groups of experimental test data are calculated at the confidence level of 99.73% ($3\delta$), and the $(\lambda, L)$ values of EWMA are presented in Table 1.

The detection results for the Daman and Arou superstation data set are presented in Fig.12 and 13. Fig.12-a shows that the upper and lower limits of the EWMA control chart are approximately 0.6125 and $-0.6125$, respectively. According to the figure, the residual error obtained from the predicted and real values is within the upper and lower limits. As a result, the error range between the predicted and real values are 0.5. Similarly, in Fig.12-b, the upper and lower limits of the EWMA control chart are 1.6 and $-1.4$, respectively, and the error range of the humidity data is 2.5.

The results for Arou superstation data are shown in Fig.13. The upper and lower limits of temperature are approximately 1 and $-1$, respectively, in Fig.13-a, and the error almost zero. As shown in Fig. 13-b, the upper and lower limits of the humidity data are approximately 5 and $-5$, respectively, and several outliers are clearly marked, but these values are not shown in Fig.9-b. For instance, from 0 to 125, 3 obvious abnormal points between the predicted and real values are present, especially the 121th point, with an error that reaches 9. This point was not marked in Fig.12-b but was detected by the EWMA model.

These results suggest that the introduced approach, as an outlier detector, is effective in detecting outliers in time series and predicted values. In the meantime, the proposed model also targets processing outliers. Specifically, the obvious outlier can be preprocessed through the first-level of the proposed three-level adaptive detection system. Additionally, we can analyze conditions of actual values preliminarily based on the confidence level, and for the detected outliers that deviate from the UCL and LCL control limits of the EWMA control chart, we replace them using the prediction

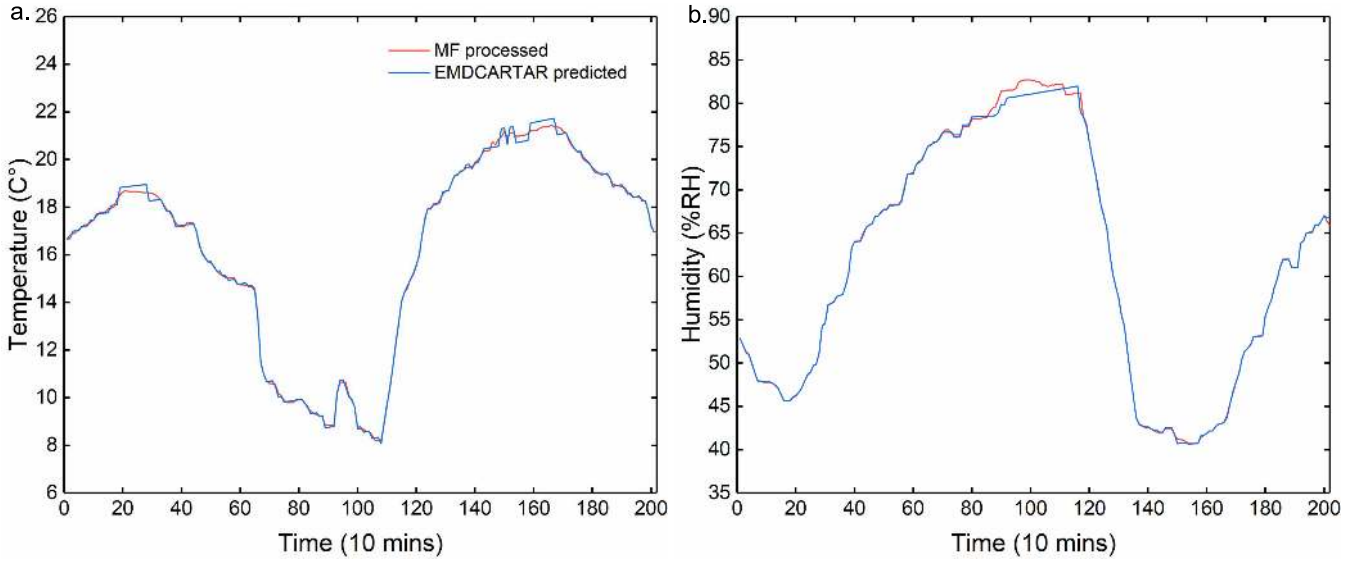


**FIGURE 10.** Results for the temperature and humidity data from Daman superstation predicted by EMD-CART-AR.

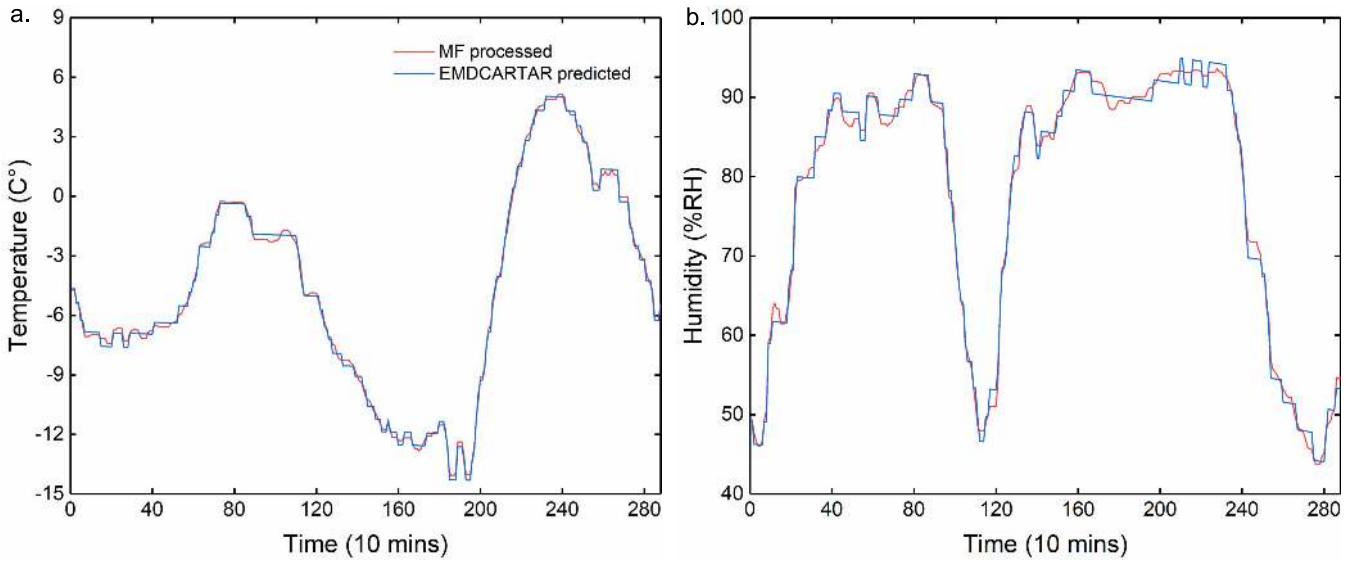


**FIGURE 11.** Results for the temperature and humidity data from Arou superstation predicted by EMD-CART-AR.

value, while addressing them for further analysis. Moreover, outliers may still be triggered from systematic noise and sensor faults. As a result, the role of the preprocessor and detector, e.g., the first level, for preprocessing some obvious outliers and minor outliers that are detected by the EWMA control chart, can also be deemed an alarm for some special applications.

## IV. DISCUSSION

This section mainly discusses and analyzes the performance of the MF-EMD-CART-AR-EWMA model proposed in this paper, which involves three levels, as shown in Fig.1. The MF, which is a single model, is used to preprocess outlier data in the first level. The second level includes a hybrid model, the EMD-CART-AR model, which is used to establish a prediction model. The last level identifies outliers based on detection with the EWMA control chart. The first two levels mainly improve the accuracy and robustness of the proposed model, and the third provides effective outlier detection and iterative control for regulating the allowable error range by adjusting the parameters of the EWMA control chart. Therefore, the performance of the three levels of the prototype model proposed in this paper are analyzed and evaluated.

To clearly illustrate the performance, stability and robustness of the MF-EMD-CART-AR-EWMA model, the temperature and humidity data collected at the Daman and Arou superstations were chosen. A comparison of the MF and other signal processing methods, e.g., WT, the Butterworth filter,

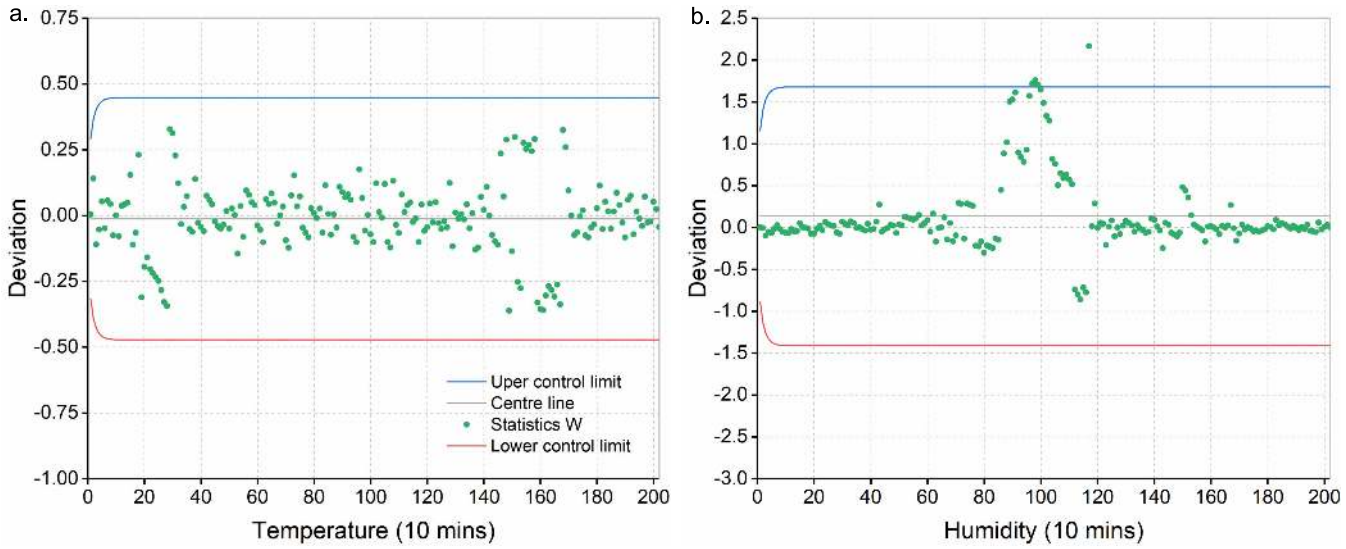


**FIGURE 12.** Results obtained by the EWMA control chart for Daman weather station data.

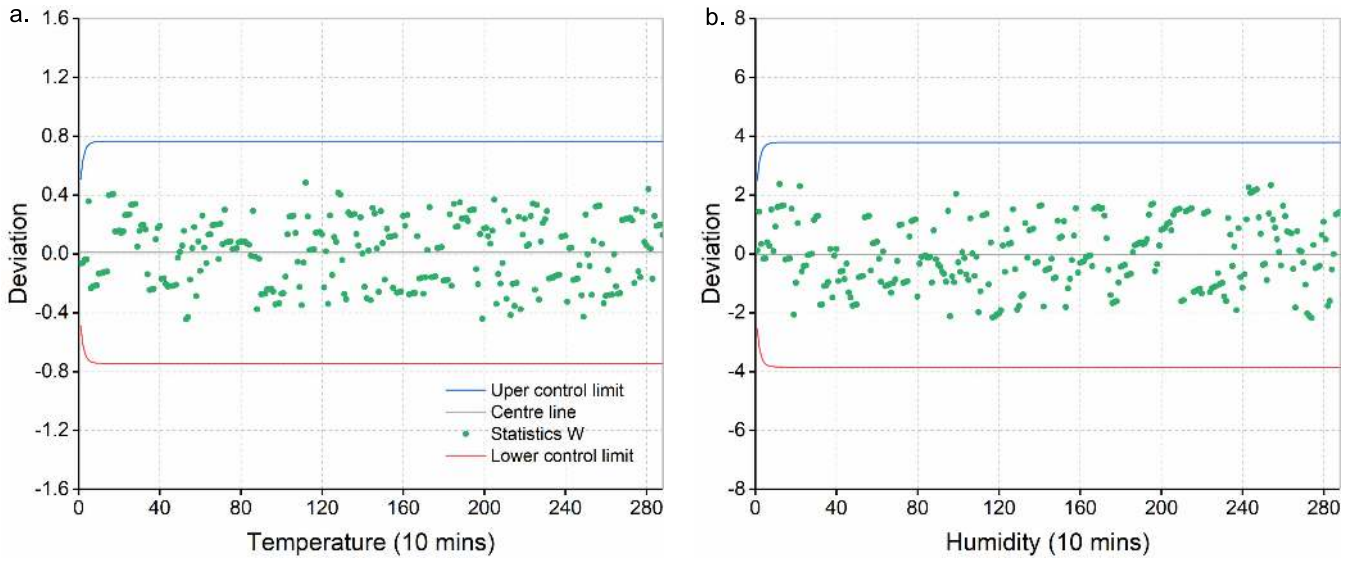


**FIGURE 13.** Result obtained by the EWMA control chart for Arou superstation data.

and others, was performed to demonstrate the advantages and performance of the MF in processing sensor data outliers. In addition, to evaluate the prediction ability of the EMD-CART-AR model, other comparisons are made involved the model and the SVR, KNN, CART, CEEMD-CART-AR and EEMD-CART-AR models.Finally, we assessed the performance of the EWMA control chart with the CUSUM control chart and the Shewhart control chart in terms of outlier detection. Figs. 14-20 illustrate the performance of the preprocessing model, prediction model and outlier detection model, and the results are presented in Tables 2-9.

### A. EVALUATION METHODOLOGY

To evaluate the preprocessing and prediction ability of the model, three different statistical indicators, namely, the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), were used [51]. The preprocessing and prediction accuracies reflect the consistency between the processed results and actual values, and these accuracies are usually reflected by error indicators. Therefore, the larger the error is, the lower the accuracy. The error is defined as $\varepsilon = y(t) - \hat{y}(t)$, where $y(t)$ is the actual value and $\hat{y}(t)$ is the preprocessed or predicted value. When $\varepsilon > 0$, $\hat{y}(t)$ is a poorly predicted value; conversely, when $\varepsilon < 0$, the prediction accuracy is high. The metrics are shown in Eq. (8), Eq. (9), and Eq. (10).

$$MAE = \frac{\sum_{t=1}^{n} |y(t) - \hat{y}(t)|}{n} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y(t) - \hat{y}(t))^2}{n}} \tag{9}$$

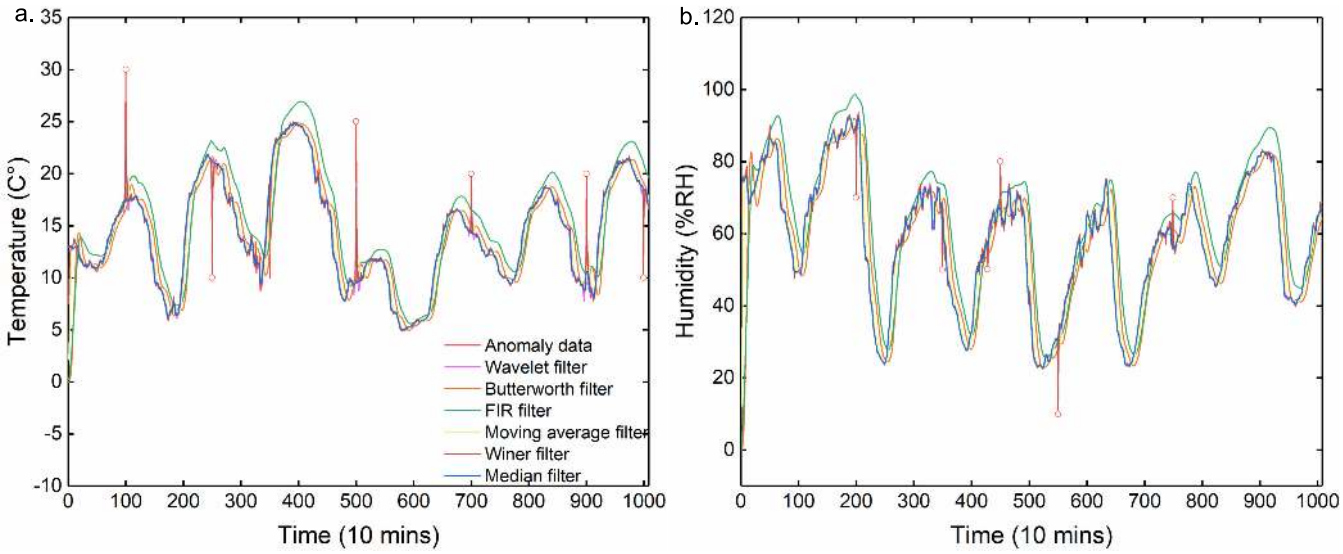


**FIGURE 14.** Comparison of the outliers in the raw data and outliers preprocessed by the MF, the Butterworth filter, the FIR filter, the moving average filter, wavelet transform, and the Wiener filter for temperature and humidity datasets from Daman superstation.

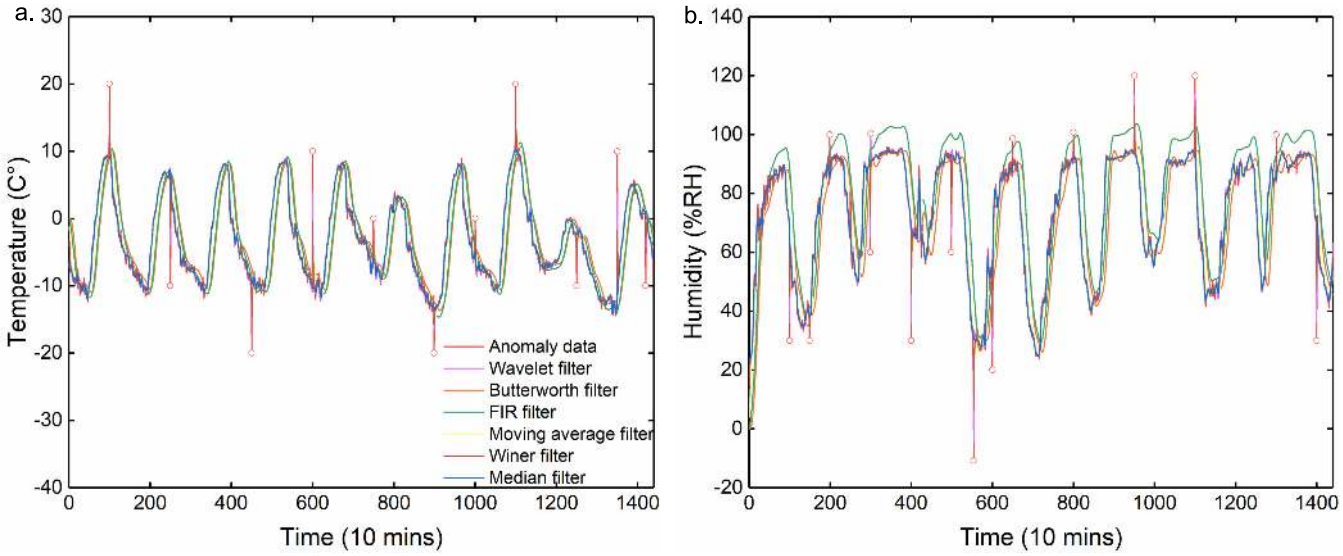


**FIGURE 15.** Comparison of the outliers in the raw data and outliers preprocessed by the MF, the Butterworth filter, the FIR filter, the moving average filter, wavelet transform, and the Wiener filter for temperature and humidity datasets from Arou superstation.

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{y(t)-\hat{y}(t)}{\hat{y}(t)} \right|}{n} * 100 \qquad (10)$$

### B. RESULTS AND DISCUSSION OF THE PREPROCESSING MODEL

Performance comparisons based on the MF and other filter methods for the raw temperature and humidity data from the Daman and Arou superstations are given in Fig.14 and Fig.15. The results of the MF are compared to those of the Butterworth filter, the FIR filter, the moving average filter, WT, and the Wiener filter. Based on the preprocessing scheme used in this paper, the preprocessed data are close to the real data because the MF is a nonlinear smoothing technique that sets the value of a given data point as the median of all data values in the corresponding neighborhood window. As a result, some obvious outlier points are processed. The results suggest that the MF outperforms the other methods for both specific points and the whole dataset in terms of processing the series outliers. The statistical evaluation criteria for several filter methods, such as the MAE, RMSE, and MAPE, are shown in Table 2 and 3. Similarly, the MF performs better than other filters in processing the data outliers. The MAE, RMSE and MAPE of the data processed by the MF are smaller than those for the Butterworth filter, FIR filter, moving average filter and Wiener filter for both the temperature and humidity datasets from the Daman and Arou superstations.

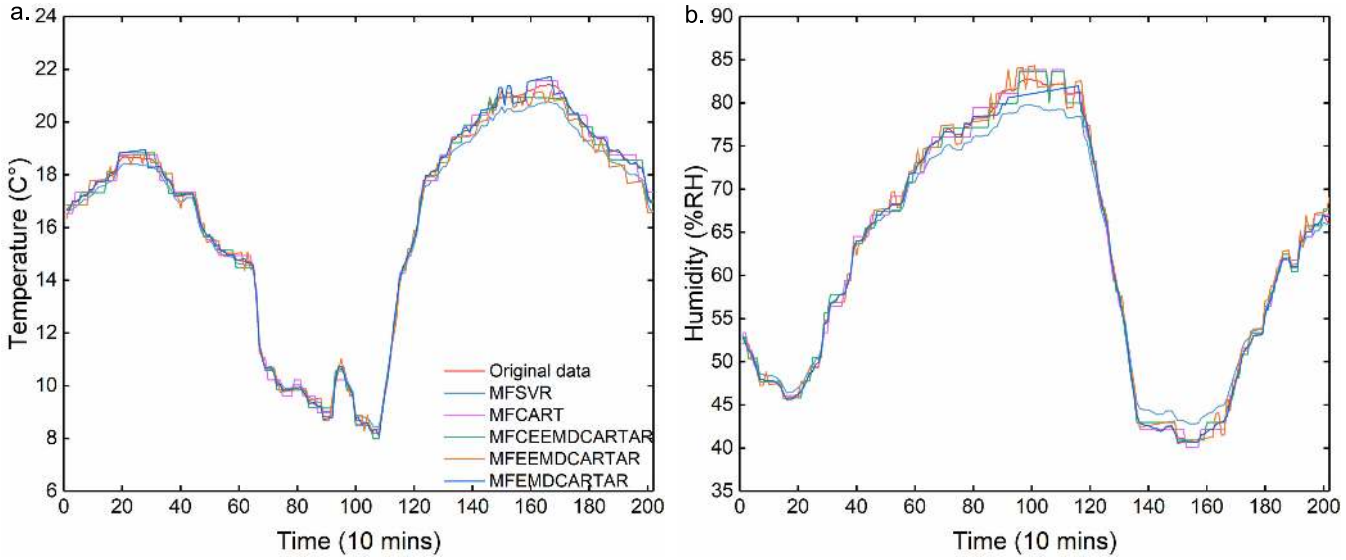


**FIGURE 16.** Comparison of raw data and values predicted with the MF-SVR, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR, and MF-EMD-CART-AR modes for temperature and humidity data from Daman superstation.

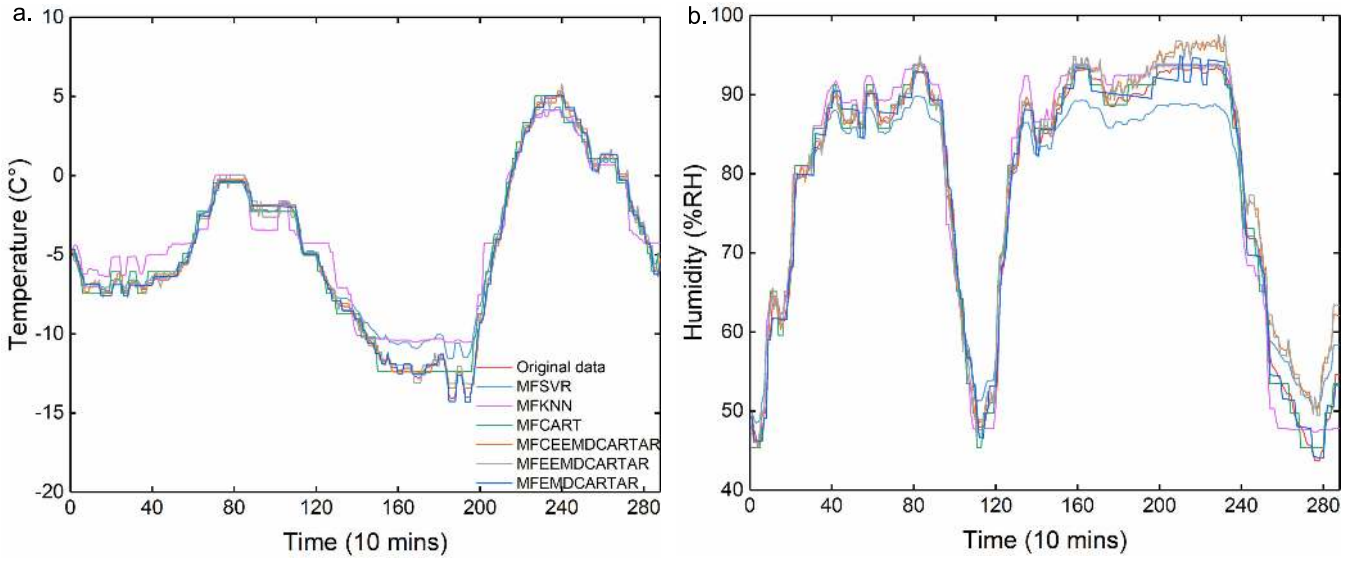


**FIGURE 17.** Comparison of the raw data and values predicted by the MF-SVR, MF-KNN, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR, and MF-EMD-CART-AR models for temperature and humidity data from Arou superstation.

For the Daman superstation test set, the MF yields the maximum observed improvement over the FIR filtering results based on the temperature data, with MAE, RMSE and RMSE values of approximately 96.7%, 99.5% and 96.9%, respectively. Additionally, the observed improvements in the MAE, RMSE and RMSE were approximately 97%, 99.7% and 97.4%, respectively, for the humidity data.

For the Arou superstation test set, compared with Butterworth filter, the MF yielded a 93.3% improvement in MAE, a 99.6% improvement in RMSE and a 95.8% improvement in MAPE for the temperature data. Similarly, improvements of approximately 93.4% in MAE, 99.6% in RMSE and 93.3% in MAPE were obtained for the humidity data.

Outliers influence the estimation of the parameters of prediction model; therefore, to effectively detect data outliers, data preprocessing is emphasized in this paper to improve the accuracy of the prediction model. According to Table 2 and 3, the MF has clear advantages in processing obvious outliers compared to the other methods assessed and displays stronger generalization ability and robustness.

### C. RESULTS AND DISCUSSION OF THE PREDICTION MODEL

For the Daman superstation test sets, several models, e.g., MF-SVR, MF-CART, MF-EEMD-CART-AR, and MF-CEEMD-CART-AR, were assessed to evaluate the

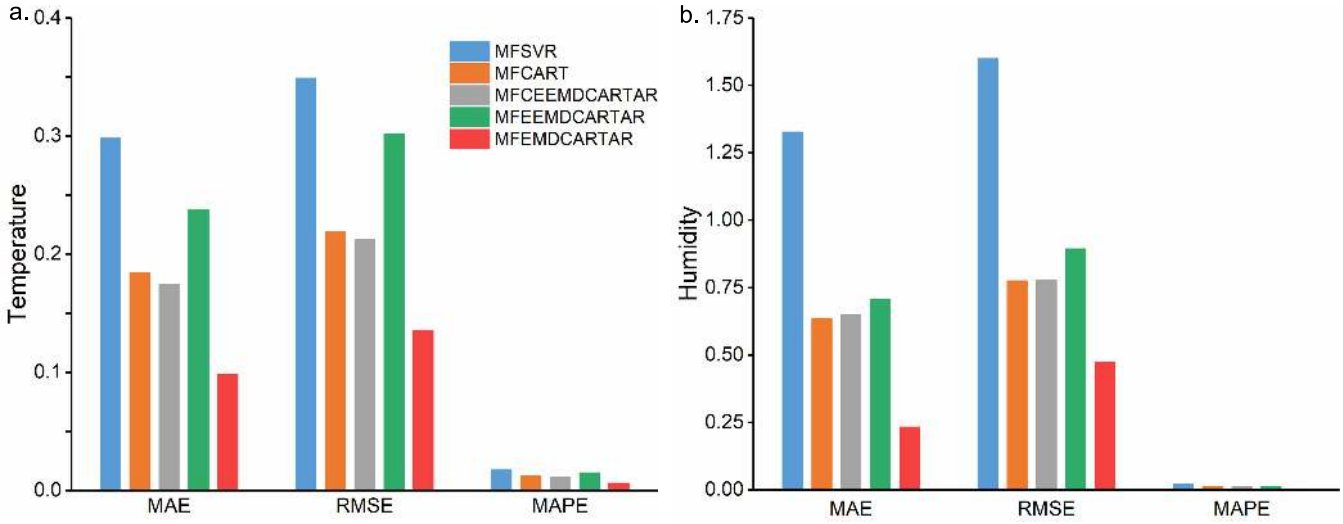


**FIGURE 18.** Comparison of the MAE, RMSE and MAPE for the MF-SVR, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR, and MF-EMD-CART-AR models based on temperature and humidity data from Daman superstation.

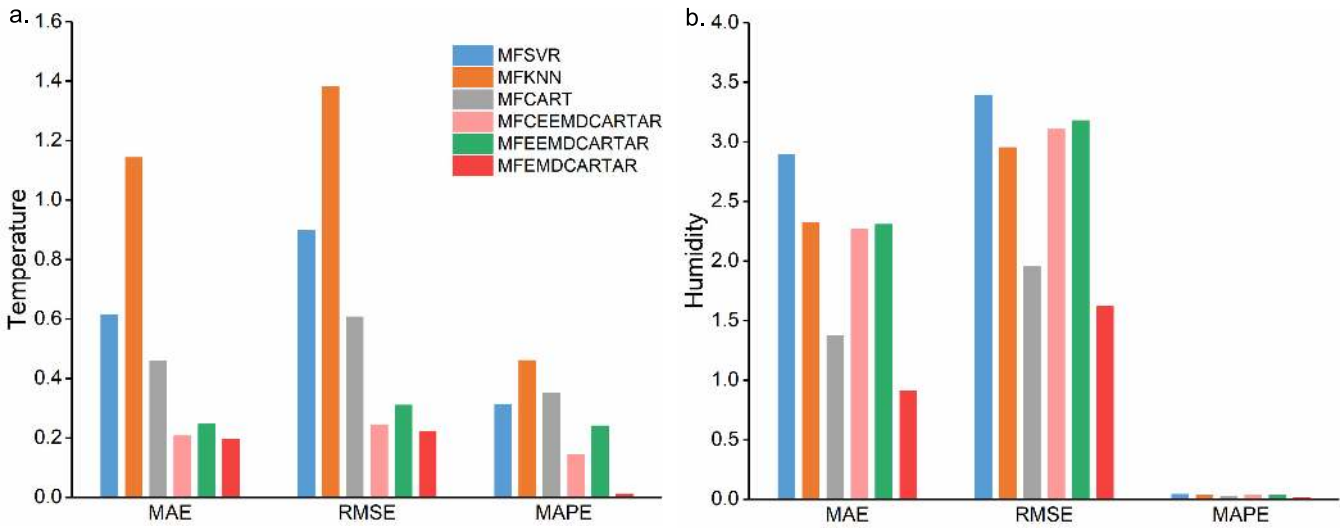


**FIGURE 19.** Comparison of the MAE, RMSE and MAPE for the MF-SVR, MF-KNN, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR, and MF-EMD-CART-AR models based on temperature and humidity data from Arou superstation.

**TABLE 2.** Preprocessing result of temperature and humidity data at Daman superstation.

| Preprocessing method | Temperature | | | | Humidity | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | PCCs | MAE | RMSE | MAPE | PCCs |
| Butterworth filter | 1.3815 | 4.3990 | 0.1079 | 0.9141 | 6.2564 | 95.8295 | 0.1215 | 0.8532 |
| FIR filter | 2.0367 | 6.6354 | 0.1580 | 0.9024 | 7.8559 | 133.6087 | 0.1593 | 0.8348 |
| Moving average filter | 0.5134 | 0.7791 | 0.0404 | 0.9844 | 2.2802 | 16.0058 | 0.0426 | 0.9746 |
| Wiener filter | 0.2168 | 0.3833 | 0.0168 | 0.9923 | 0.2709 | 1.4256 | 0.0049 | 0.9977 |
| Wavelet filter | 0.0981 | 0.0187 | 0.0060 | 0.9931 | 0.3168 | 0.0054 | 0.6568 | 0.9972 |
| **Median filter** | **0.0654** | **0.0307** | **0.0048** | **0.9995** | **0.2322** | **0.2235** | **0.0032** | **0.9992** |

performance of the proposed scheme, the MF-EMD-CART-AR hybrid model [52]. The results predicted by the employed models and the original data are presented in Fig.16. Notably, compared with the single models, such as MF-SVR and MF-CART, the hybrid models, such as MF-EMD-CART-AR, MF-EEMD-CART-AR, and MF-CEEMD-CART-AR, yield higher accuracy and better performance. For example, compared with MF-SVR, MF-CEEMD-CART-AR yielded

**TABLE 3.** Preprocessing result of temperature and humidity data at Arou superstation.

| Preprocessing method | Temperature | | | | Humidity | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | PCCs | MAE | RMSE | MAPE | PCCs |
| Butterworth filter | 2.9280 | 14.1629 | 3.1274 | 0.8218 | 8.8279 | 153.9130 | 0.1455 | 0.8169 |
| FIR filter | 2.4488 | 10.4660 | 2.6953 | 0.8766 | 9.9245 | 179.9066 | 0.1583 | 0.8726 |
| Moving average filter | 0.8420 | 1.2477 | 0.7113 | 0.9843 | 2.5553 | 14.4844 | 0.0415 | 0.9821 |
| Wiener filter | 0.4037 | 0.5645 | 0.2481 | 0.9930 | 1.0357 | 5.2479 | 0.0164 | 0.9935 |
| Wavelet filter | 0.2188 | 0.1058 | 0.1558 | 0.9902 | 0.9069 | 2.6217 | 0.0133 | 0.9907 |
| **Median filter** | **0.1954** | **0.0492** | **0.1320** | **0.9980** | **0.5826** | **1.4214** | **0.0097** | **0.9982** |

**TABLE 4.** Prediction result of temperature and humidity data at Daman superstation.

| Prediction model | Temperature | | | | Humidity | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | PCCs | MAE | RMSE | MAPE | PCCs |
| MFSVR | 0.2980 | 0.3489 | 0.0173 | 0.9949 | 1.3262 | 1.5979 | 0.0218 | 0.9983 |
| MFCART | 0.1841 | 0.2191 | 0.0124 | 0.9986 | 0.6334 | 0.7752 | 0.0103 | 0.9986 |
| MFCEEMDCARTAR | 0.1743 | 0.2122 | 0.0113 | 0.9988 | 0.6480 | 0.7761 | 0.0104 | 0.9985 |
| MFEEMDCARTAR | 0.2377 | 0.3014 | 0.0149 | 0.9980 | 0.7044 | 0.8936 | 0.0117 | 0.9982 |
| **MFEMDCARTAR** | **0.0981** | **0.1355** | **0.0060** | **0.9995** | **0.2322** | **0.4729** | **0.0032** | **0.9996** |

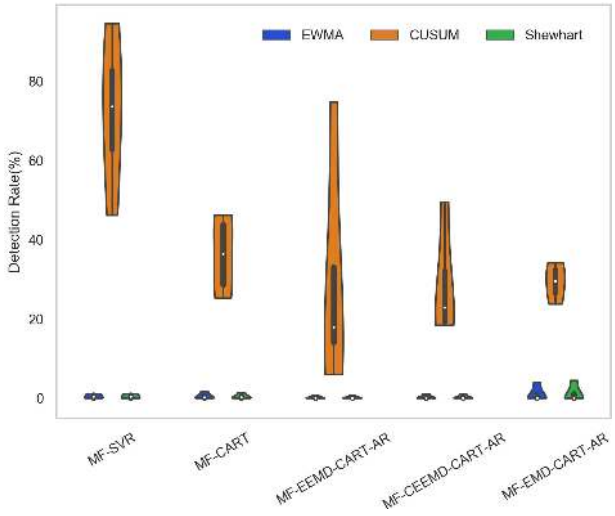


**FIGURE 20.** Comparison of the EWMA, CUSUM and Shewhart control chart integrated with the MF-SVR, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR, and MF-EMD-CART-AR models based on temperature and humidity data from Daman and Arou superstation.

41.5%, 39.1% and 34.7% improvements in the MAE, RMSE and MAPE, respectively, for the temperature data. Similarly, compared with MF-SVR, MF-CEEMD-CART-AR yielded improvements of 48.4%, 51.4% and 52.3% for the MAE, RMSE and MAPE, respectively, based on the humidity data. The MF-EMD-CART-AR model compared with MF-SVR yielded the maximum observed improvement for the temperature data, including approximately 67.1% for the MAE, 61.1% for the RMSE and 65.3% for the MAPE. Additionally, observed improvements were approximately 82.5%, 70.4%, and 85.3% for the MAE, RMSE and MAPE, respectively, based on the humidity data. Notably, the hybrid models include signal decomposition methods, which decompose non-stationary series into relatively stationary series with different characteristics to improve the accuracy of predictions. The results of this experiment demonstrate that the characteristics of stationary and non-stationary data have a considerable influence on the prediction accuracy.

Similarly, a performance comparison of the EMD-CART-AR model and the single models (e.g., MF-SVR and MF-CART) for both specific points and the entire dataset is given in Fig.16. According to the figure, the blue curve shows the MF-EMD-CART-AR predictions, and the black curve illustrates the original data. The findings presented in this figure indicate that the two curves largely coincide. In addition, to highlight the superiority of the EMD method, a hybrid model was constructed by combining the CEEMD and EEMD methods for comparison. As shown in Fig.16, the MF-EMD method provides better prediction performance than MF-EEMD and MF-CEEMD. The results confirm that to some extent, the EMD approach introduced in this paper performs better than EEMD and CEEMD in processing non-stationary data. Similarly, Table 4 shows that the MF-EMD-CART-AR model outperforms the MF-EEMD-CART-AR and MF-CEEMD-CART-AR models in terms of the prediction ability. For example, compared with MF-CEEMD-CART-AR, MF-EMD-CART-AR yields a 43.7% improvement in MAE, a 42.6% improvement in RMSE and a 46.9% improvement in MAPE for the temperature data. Similar, a comparison of MF-EMD-CART-AR and MF-CEEMD-CART-AR highlights increases of 64.2%, 39.1%, and 69.2% in the MAE, RMSE, and MAPE, respectively, for the humidity data. According to Table 2, the results confirm that the developed model performs better than MF-EEMD-CART-AR

**TABLE 5.** Prediction result of temperature and humidity data at Arou superstation.

| Prediction model | Temperature | | | | Humidity | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | PCCs | MAE | RMSE | MAPE | PCCs |
| MFSVR | 0.6146 | 0.8964 | 0.3118 | 0.9946 | 2.8883 | 3.3869 | 0.0409 | 0.9933 |
| MFKNN | 1.1422 | 1.3821 | 0.4598 | 0.9804 | 2.3217 | 2.9494 | 0.0329 | 0.9889 |
| MFCART | 0.4574 | 0.6061 | 0.3494 | 0.9931 | 1.3724 | 1.9487 | 0.0200 | 0.9932 |
| MFCEEMDCARTAR | 0.2061 | 0.2435 | 0.1420 | 0.9989 | 2.2640 | 3.1043 | 0.0350 | 0.9886 |
| MFEEMDCARTAR | 0.2448 | 0.3096 | 0.2399 | 0.9982 | 2.3055 | 3.1770 | 0.0357 | 0.9874 |
| **MFEMDCARTAR** | **0.1954** | **0.2212** | **0.0107** | **0.9991** | **0.9069** | **1.6192** | **0.0133** | **0.9950** |

**TABLE 6.** Control limits of EMWA, CUSUM and Shewhart control chart on temperature and humidity data from Daman superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | Temperature | Humidity | Temperature | Humidity | Temperature | Humidity |
| MFSVR | 1.0010 | 5.0801 | 0.7651 | 4.6794 | 1.0039 | 5.0431 |
| | -0.5323 | -4.3527 | -0.7651 | -4.6794 | -0.5262 | -4.3158 |
| MFCART | 0.6736 | 2.2167 | 0.6584 | 2.3079 | 0.6683 | 2.1985 |
| | -0.6536 | -2.4356 | -0.6584 | -2.3079 | -0.6484 | -4.2174 |
| MFCEEMDCARTAR | 0.6770 | 2.2972 | 0.6116 | 2.3266 | 0.6721 | 2.2788 |
| | -0.5560 | -2.3927 | -0.6116 | -2.3266 | -0.5511 | -2.3743 |
| MFEEMDCARTAR | 0.9592 | 2.3474 | 0.8361 | 2.5787 | 0.9526 | 2.3270 |
| | -0.7263 | -2.8501 | -0.8361 | -2.5787 | -0.7197 | -2.8305 |
| **MFEMDCARTAR** | **0.3973** | **1.5086** | **0.4060** | **1.3637** | **0.3940** | **1.4978** |
| | **-0.4212** | **-1.2405** | **-0.4060** | **-1.3637** | **-0.4179** | **-1.2296** |

**TABLE 7.** Control limits of EMWA, CUSUM and Shewhart control chart on temperature and humidity data from Arou superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | Temperature | Humidity | Temperature | Humidity | Temperature | Humidity |
| MFSVR | 2.1888 | 10.0766 | 2.5016 | 9.8984 | 2.1690 | 10.6874 |
| | -2.8539 | -9.1877 | -2.5016 | -9.8984 | -2.8341 | -9.1093 |
| MFCART | 1.9797 | 5.6851 | 1.8186 | 5.8285 | 1.7835 | 5.6490 |
| | -1.8679 | -6.0641 | -1.8186 | -5.8285 | -1.8535 | -6.0181 |
| MFKNN | 3.1885 | 8.0629 | 3.7518 | 8.6452 | 3.1589 | 7.9946 |
| | -3.3743 | -9.3640 | -3.7518 | -8.6452 | -4.3447 | -9.2957 |
| MFCEEMDCARTAR | 0.6983 | 5.9706 | 0.7252 | 7.6747 | 0.6926 | 5.9099 |
| | -0.7635 | -9.5002 | -0.7252 | -7.6747 | -0.7578 | -9.4396 |
| MFEEMDCARTAR | 0.8749 | 6.3509 | 0.9180 | 8.0142 | 0.8677 | 6.2876 |
| | -0.9757 | -9.8043 | -0.9180 | -8.0142 | -0.9684 | -9.7401 |
| **MFEMDCARTAR** | **0.6810** | **3.3669** | **0.6656** | **3.3681** | **0.6782** | **3.3403** |
| | **-0.6606** | **-3.4223** | **-0.6656** | **-3.3681** | **-0.6554** | **-3.3960** |

and MF-CEEMD-CART-AR. This result suggests that the EMD method displays better performance in processing the non-stationary data than do EEMD and CEEMD because of its consideration the dynamic behavior of sensor data, with obvious physical meaning.

For the Arou superstation test set, as shown in Fig.18, the applicability, generality and superiority of the MF-EMD-CART-AR model were further verified. Three evaluation criteria, the MAE, RMSE and MAPE, were used to compare the proposed model and other models. For the Arou superstation test set, the MF-KNN and MF-EMD-CART-AR models were compared, and the proposed model yielded a 68.2% improvement in MAE, an 83.9% average improvement in RMSE and a 97.6% improvement in MAPE for the temperature data, as well as 60.9%, 45.1% and 59.6% improvements in MAE, RMSE and MAPE, respectively, for the humidity data. The results using different prediction models are shown in 5. The findings show that the MF-EMD-CART-AR model proposed in this paper outperforms all others based on all three evaluation criteria.

The error measures for the MF-SVR, MF-KNN, MF-CART, MF-CEEMD-CART-AR, MF-EEMD-CART-AR and

**TABLE 8.** Performance of integrated detection model on temperature data from Daman superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | DR | FR | DR | FR | DR | FR |
| MFSVR | 0 | 32.67% | 94.55% | 4.46% | 0 | 32.67% |
| MFCART | 0 | 4.46% | 29.70% | 18.81% | 0 | 4.46% |
| MFCEEMDCARTAR | 0 | 5.94% | 49.50% | 13.37% | 0 | 5.94% |
| MFEEMDCARTAR | 0 | 17.33% | 74.75% | 18.31% | 0 | 17.33% |
| **MFEMDCARTAR** | **0** | **0** | **23.76%** | **23.76%** | **0** | **0** |

**TABLE 9.** Performance of integrated detection model on humidity data from Daman superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | DR | FR | DR | FR | DR | FR |
| MFSVR | 0 | 46.04% | 68.32% | 14.36% | 0 | 32.67% |
| MFCART | 0 | 8.91% | 25.25% | 14.37% | 0 | 8.91% |
| MFCEEMDCARTAR | 0 | 7.92% | 26.24% | 16.83% | 0 | 7.92% |
| MFEEMDCARTAR | 0 | 19.80% | 5.94% | 21.78% | 0 | 19.80% |
| **MFEMDCARTAR** | **3.96%** | **3.96 %** | **34.16%** | **34.16%** | **4.46%** | **4.46%** |

MF-EMD-CART-AR models based on data from the Daman and Arou superstations are shown in Fig.17 and Fig.18. Notably, the largest improvement was obtained by MF-EMD-CART-AR. The two figures also show that the errors associated with the MF-CEEMD-CART-AR and MF-EEMD-CART-AR results are lower than the errors for the single model results, such as MF-SVR, MF-KNN and MF-CART. The largest differences between the MF-CEEMD-CART-AR and MF-EMD-CART-AR MAE, RMSE and MAPE values were 43.7%, 36.1%, and 46.9%, respectively. Fig.18 and Fig.19 also show that the majority of the improvement in the overall error is due to the preprocessing and signal decomposition methods. The performance, which was evaluated based on the three criteria, confirms that the accuracy of the MF-EMD-CART-AR model is higher than that of the other models.

The prediction accuracy and statistical interpretation performance can be summarized as follows: a. the hybrid model can effectively provide predictions based on sensor data; b. the combination of the CART and AR models enhances the performance of the hybrid model; c. the comparison of the MF-EMD-CART-AR model and other models indicates that the proposed model displays superior performance; d. the comparison of the four sets of temperature and humidity experiments with different sampling times and sample numbers indicates the MF-EMD-CART-AR model has good generalization ability; and e. as shown in Table 4 and Table 5, the model is accurate, broadly applicable, robust and effective. In summary, the MF-EMD-CART-AR model provides an effective method for outlier detection based on predictions for sensor data.

### D. RESULTS AND DISCUSSION OF THE OUTLIER DETECTION MODEL

In this section, we use residual sequences of the real and predicted values of temperature and humidity taken from Daman superstation and Arou superstation to evaluate the ability of the detector. At the outlier detection stage, the UCL and LCL control limits of employed methods are computed based on the confidence level of 99.73% ($3\delta$). Tables 6-7 show the control limits of these methods. Fig.20 shows the performance of EMWA, CUSUM and Shewhart control charts based on prediction for detecting outliers using grouped violin plots in all experimental test sets. Each bar is a sideways plot of the distribution of each DR or FR across per group test sets.

From Table 6, it can be found that the UCL and LCL control limits of MF-EMD-CART-AR-EWMA are (-0.4212, 0.3973) for temperature data and (-1.2405, 1.5086) for humidity data. It shows that the MF-EMD-CART-AR-EWMA method has narrowest control limits compared to the others (e.g., MF-EMD-SVR-EWMA, MF-EEMD-CART-AR-EWMA, MF-CEEMD-CART-AR-EWMA, MF-SVR and MF-CART). Notice that the three control charts almost have the same control limits, while having different strategies for detecting outliers. Likewise, it can be seen in Table 6 that the MF-EMD-CART-AR-EWMA has the narrowest control limits of (-0.6606, 0.6810) and (-3.4223, 3.3669) for temperature data. An important problem in a detection model is the accuracy of the prediction method that leads to the change of control limits of the EMWA, CUSUM and Shewhart control charts, compromising the final detection results achieved by the detector operation.

For the Daman and Arou superstation test sets, MF-SVR, MF-CART, MF-EEMD-CART-AR, and MF-CEEMD-CART-AR are combined with the EWMA control chart to evaluate the performance of the proposed scheme, the MF-EMD-CART-AR-EWMA model. The results detected by the different detection schemes are presented in Tables 8-11. The findings presented tables are evaluated by detection ratio (DR) and fail-detection rate (FR). Therefore, DR is defined as the ratio of the amount of the points in which the outlier is detected to the total amount of test points. The FR is the ratio of the amount of the points in which the outlier failed to be detected. The results confirm that the detection scheme introduced in this work achieves comparable performance with MF-SVR-EMWA, MF-CART-EWMA, MF-EEMD-CART-AR-EWMA, and MF-CEEMD-CART-AR-EWMA across all dataset groups. This is because the proposed MF-EMD-CART-AR has superior performance compared to MF-SVR, MF-CART, MF-EEMD-CART-AR, and MF-CEEMD-CART-AR. As a result, MF-EMD-CART-AR-EWMA achieves good accuracy, thereby reducing a failure detection in the outlier detection model.

Similarly, to see the functionalities and performance of the proposed detection method, some contrast tests were performed which include the CUSUM control chart and Shewhart control chart. A performance comparison of DR and

**TABLE 10.** Performance of integrated detection model on temperature data from Arou superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | DR | FR | DR | FR | DR | FR |
| MFSVR | 0.69% | 30.90% | 46.18% | 5.90% | 0.69% | 11.11% |
| MFCART | 0.35% | 22.91% | 43.06% | 30.20% | 0.35% | 9.72% |
| MFKNN | 0.34% | 70.14% | 54.51% | 28.13% | 0.35% | 15.28% |
| MFCEEMDCARTAR | 0 | 0.35% | 19.44% | 1.74% | 0 | 0.69% |
| MFEEMDCARTAR | 0 | 3.47% | 16.67% | 3.13% | 0 | 3.45% |
| **MFEMDCARTAR** | **0** | **0** | **27.43%** | **27.43%** | **0** | **0** |

**TABLE 11.** Performance of integrated detection model on humidity data from Arou superstation.

| Detection Model | EMWA | | CUSUM | | Shewhart | |
|---|---|---|---|---|---|---|
| | DR | FR | DR | FR | DR | FR |
| MFSVR | 1.04% | 37.50% | 78.82% | 17.36% | 1.04% | 34.04% |
| MFCART | 1.74% | 5.21% | 46.18% | 8.68% | 1.39% | 3.13% |
| MFKNN | 1.04% | 20.14% | 65.97% | 20.83% | 1.39% | 7.29% |
| MFCEEMDCARTAR | 1.04% | 22.57% | 18.40% | 10.42% | 1.04% | 2.08% |
| MFEEMDCARTAR | 0.69% | 21.53% | 19.09% | 8.68% | 0.69% | 1.74% |
| **MFEMDCARTAR** | **0** | **0** | **31.60%** | **31.59%** | **0** | **0** |

FR is given in Tables 8-11. Notice that none of the methods can be said to be consistently superior in the four group test sets. For example, compared with the Shewhart control chart, EWMA has almost the same DR in all datasets. Meanwhile, MF-EMD-CART-AR-EWMA and MF-EMD-CART-AR-Shewhart show superior performance in general with comparably low DR and FR compared to MF-EMD-CART-AR- CUSUM. MF-EMD-CART-AR- CUSUM has high DR and FR due to its detection strategy. Therefore, the CUSUM control chart is not particularly suited for time-series sensor data.It is easy to find that our adaptive methods offer a great improvement in detection rate compared to MF-EMD-CART-AR-CUSUM. Additionally, EWMA control chart among Shewhart control chart are process control strategy for monitoring outliers, while Shewhart control chart assumes that observations obey a Gaussian distribution, EWMA control chart are robust against this assumption and particularly suited for time-series data. Thus, we employed EWMA control chart as the outlier detection method to achieve an adaptive outlier detection approach towards time-series sensor.

## V. CONCLUSION

In this paper, the proposed three-level hybrid model, which integrates preprocessing, prediction and outlier detection tasks, achieves excellent performance in outlier detection for non-stationary and nonlinear data collected by environment monitoring network networked sensors. To address the sensitivity of the prediction model with respect to outliers, preliminary screening based on the MF method, as the first level of the model, is conducted, and this approach significantly outperforms five other methods in preprocessing obvious outliers. EMD can decompose non-stationary data into stationary data series, and the prediction model simultaneously considers the accuracy and robustness of the prediction result. In this context, the EMD-CART-AR prediction model is proposed as the second level of the model, and it outperforms other models in predictions based on sensor data. For instance, compared with a single model, e.g., MF-SVR, the maximum observed improvements for temperature data from Daman superstation are approximately 67.1% for MAE, 61.1% for RMSE and 65.3% for MAPE by applying MF-EMD-CART-AR, and compared with hybrid models, e.g., MF-CEEMD-CART-AR, the improvements in the MAE, RMSE, and MAPE are 43.7%, 36.1%, and 46.9% for the humidity data, respectively. Then, an EWMA control chart, as the last level in the model, is formulated to detect minor deviations in the data. This approach is especially suitable for outlier detection in predicted values. A three-level hybrid model is constructed to identify and treat outliers in environmental monitoring data.

We evaluate the performance of the proposed approach with four data series from a real-world sensor data set of the hydrometeorological observation network in the Heihe River Basin. The experimental results suggest that the preprocessing and prediction methods proposed in this paper achieve a better generalization ability and higher accuracy levels than other models in dealing with non-stationary and nonlinear sensor data. Moreover, the detection method displays outstanding effectiveness in terms of minor outlier detection. This research provides a new perspective for outlier detection and improvements to environmental monitoring data. However, this research evaluates only temperature and humidity data, including humidity data with weak non-stationary characteristics. In future work, the proposed method will be further expanded and optimized to detect outliers in different sensor data.

## REFERENCES

[1] Y. K. Chen, "Challenges and opportunities of Internet of Things," in *Proc. Asia South Pacific Design Automat. Conf.*, 2012, pp. 383–388.
[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generat. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.
[3] X. Li, N. Zhao, R. Jin, S. Liu, X. Sun, X. Wen, D. Wu, Y. Zhoug, J. Guo, S. Chen, Z. Xu, M. Ma, T. Wang, Y. Qu, X. Wang, F. Wu, and Y. Zhou, "Internet of things to network smart devices for ecosystem monitoring," *Sci. Bull.*, vol. 64, no. 64, pp. 1234–1245, 2019.
[4] S. Sen and C. Jayawardena, "Analysis of network techniques and cybersecurity for improving performance of big data IoT and cyber-physical communication Internetwork," in *Proc. IEEE Int. Conf. Ind. Technol.*, Feb. 2019, pp. 780–787.
[5] P. Buonadonna, D. Gay, J. M. Hellerstein, W. Hong, and S. Madden, "Task: Sensor network in a box," in *Proc. 2nd Eur. Workshop Wireless Sensor Netw.*, 2005, pp. 133–144.
[6] T. H. D. Nguyen, B. Adams, M. J. Zhen, A. E. Hassan, M. Nasser, and P. Flora, "Automated detection of performance regressions using statistical process control techniques," in *Proc. 3rd ACM/SPEC Int. Conf. Perform. Eng.*, 2012, pp. 299–310.
[7] L. Xie, S. Bao, L. J. Pietrafesa, K. Foley, and M. Fuentes, "A real-time hurricane surface wind forecasting model: Formulation and verification," *Monthly Weather Rev.*, vol. 134, no. 5, pp. 1355–1370, 2006.
[8] V. Chandola and V. Kumar, "Outlier detection: A survey," *ACM Comput. Surv.*, vol. 14, no. 15, pp. 1–83, 2007.

[9] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.

[10] S. K. Asl, "Outlier detection in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1302–1325, 2011.

[11] J. Mori and J. Yu, "Quality relevant nonlinear batch process performance monitoring using a kernel based multiway non-Gaussian latent subspace projection approach," *J. Process Control*, vol. 24, no. 1, pp. 57–71, 2014.

[12] M. A. Engle, M. Gallo, K. T. Schroeder, N. J. Geboy, and J. W. Zupancic, "Three-way compositional analysis of water quality monitoring data," *Environ. Ecol. Statist.*, vol. 21, no. 3, pp. 565–581, 2014.

[13] H. Fanaee-T and J. Gama, "Event detection from traffic tensors: A hybrid model," *Neurocomputing*, vol. 203, pp. 22–33, Aug. 2016.

[14] A. Arning, R. Agrawal, and P. Raghavan, "A linear method for deviation detection in large databases," *Proc. KDD*, vol. 1141, no. 50, pp. 164–169, 1996.

[15] E. M. Knorr and R. T. Ng, "A unified notion of outliers: Properties and computation," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, vol. 97, 1997, pp. 219–222.

[16] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, vol. 29, no. 2, pp. 427–438.

[17] A. H. Yaacob, I. K. T. Tan, F. C. Su, and H. K. Tan, "Arima based network anomaly detection," in *Proc. Int. Conf. Commun. Softw. Netw.*, 2010, pp. 205–209.

[18] R. S. Tsay, "Nonlinear time series models: Testing and applications," in *A Course in Time Series Analysis*. New York, NY, USA: Wiley, 2000, pp. 267–285.

[19] P. Galeano, D. Peña, and R. S. Tsay, "Outlier detection in multivariate time series by projection pursuit," *J. Amer. Stat. Assoc.*, vol. 101, no. 474, pp. 654–669, 2006.

[20] A. Alfons, C. Croux, and S. Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *Ann. Appl. Statist.*, vol. 7, no. 1, pp. 226–248, 2013.

[21] E. V. Burnaev, M. E. Panov, and A. A. Zaytsev, "Regression on the basis of nonstationary Gaussian processes with Bayesian regularization," *J. Commun. Technol. Electron.*, vol. 61, no. 6, pp. 661–671, 2016.

[22] S. Mahadevan and S. L. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," *J. Process Control*, vol. 19, no. 10, pp. 1627–1639, 2009.

[23] E. George and D. P. Foster, "Outliers in multivariate time series," *Biometrika*, vol. 87, no. 4, pp. 789–804, 2000.

[24] X. Cheng, K. Xie, and W. Dong, "Network traffic anomaly detection based on self-similarity using HHT and wavelet transform," in *Proc. Int. Conf. Inf. Assurance Secur.*, vol. 1, 2009, pp. 710–713.

[25] Z. Wu, C. Wang, and R. Aijun, "Optimal selection of wavelet base functions for eliminating signal trend based on wavelet analysis," *Trans. Beijing Inst. Technol.*, vol. 33, no. 8, pp. 811–814, 2013.

[26] L. Wang, X. Li, and Y. Bai, "Short-term wind speed prediction using an extreme learning machine model with error correction," *Energy Convers. Manage.*, vol. 162, pp. 239–250, Apr. 2018.

[27] B. Huang and A. Kunoth, "An optimization based empirical mode decomposition scheme," *J. Comput. Appl. Math.*, vol. 240, no. 240, pp. 174–183, 2013.

[28] M. Li, X. Wu, and X. Liu, "An improved emd, method for time–frequency feature extraction of telemetry vibration signal based on multi-scale median filtering," *Circuits Syst. Signal Process.*, vol. 34, no. 3, pp. 815–830, 2015.

[29] L. Angrisani, P. Daponte, M. D. D'Apuzzo, and A. Testa, "Measurement method based on the wavelet transform for power quality analysis," *IEEE Trans. Power Del.*, vol. 13, no. 4, pp. 990–998, 1998.

[30] D. Brownrigg, "The weighted median filter," *Commun. ACM*, vol. 27, no. 8, pp. 807–818, Aug. 1984.

[31] N. E. Huang, S. Zheng, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. T. Chi, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[32] N. E. Huang, M. L. Wu, W. Qu, S. R. Long, and S. S. P. Shen, "Application of Hilbert-Huang transform to non-stationary financial time series analysis," *Appl. Stochastic Models Bus. Ind.*, vol. 19, no. 3, pp. 245–268, 2010.

[33] I. Daubechies, J. Lu, and H. T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 243–261, 2011.

[34] L. Breiman, *Classification Regression Trees*. London, U.K.: Routledge, 2017.

[35] H. Liu, R. Bu, J. Liu, W. Leng, Y. Hu, L. Yang, and H. Liu, "Predicting the wetland distributions under climate warming in the great Xing'an mountains, northeastern China," *Ecol. Res.*, vol. 26, no. 3, pp. 605–613, 2011.

[36] C.-W. Lu and M. Reynolds, Jr., "EWMA control charts for monitoring the mean of autocorrelated processes," *J. Qual. Technol.*, vol. 31, no. 2, pp. 166–188, 2018.

[37] D. A. T. Tran, Y. Chen, H. L. Ao, and H. N. T. Cam, "An enhanced chiller FDD strategy based on the combination of the LSSVR-DE model and EWMA control charts," *Int. J. Refrig.*, vol. 72, pp. 81–96, 2016.

[38] A. S. Neubauer, "The EWMA control chart: Properties and comparison with other quality-control procedures by computer simulation," *Clinical Chem.*, vol. 43, no. 4, pp. 594–601, 1997.

[39] O. Ibidunmoye, A. R. Rezaie, and E. Elmroth, "Adaptive anomaly detection in performance metric streams," *IEEE Trans. Netw. Service Manage.*, to be published.

[40] X. Li, G. Cheng, S. Liu, Q. Xiao, M. Ma, R. Jin, T. Che, Q. Liu, W. Wang, and Y. Qi, "Heihe watershed allied telemetry experimental research (HiWATER): Scientific objectives and experimental design," *Bull. Amer. Meteorol. Soc.*, vol. 94, no. 8, pp. 1145–1160, 2013.

[41] S. Liu *et al.*, "The heihe integrated observatory network: A basin-scale land surface processes observatory in China," *Vadose Zone J.*, vol. 17, no. 1, pp. 1–21, 2018.

[42] G. Cheng, X. Li, W. Zhao, Z. Xu, F. Qi, S. Xiao, and H. Xiao, "Integrated study of the water–ecosystem–economy in the Heihe river basin," *Nat. Sci. Rev.*, vol. 1, no. 3, pp. 413–428, 2014.

[43] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.*, vol. 4, no. 9, pp. 973–977, Nov. 1987.

[44] Y. Liu, C. Liu, and D. Wang, "A 1D time-varying median filter for seismic random, spike-like noise elimination," *Geophysics*, vol. 74, no. 1, pp. 17–24, 2009.

[45] G. Rilling, P. Flandrin, and P. G. Calvès, "On empirical mode decomposition and its algorithms," in *Proc. IEEE-EURASIP Workshop Nonlinear Signal Image Process.*, Jun. 2003, vol. 3, no. 3, pp. 8–11.

[46] S. Sarkar, A. Patel, S. Madaan, and J. Maiti, "Prediction of occupational accidents using decision tree approach," in *Proc. IEEE Annu. India Conf.*, 2016, pp. 1–6.

[47] J. Wang, W. Zhang, Y. Li, J. Wang, and Z. Dang, "Forecasting wind speed using empirical mode decomposition and Elman neural network," *Appl. Soft Comput.*, vol. 23, no. 5, pp. 452–459, 2014.

[48] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, "Gis-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran," *Environ. Monitor. Assessment*, vol. 188, no. 1, p. 44, 2016.

[49] M. Gan, H. Peng, X. Peng, X. Chen, and G. Inoussa, "A locally linear RBF network-based state-dependent ar model for nonlinear time series modeling," *Inf. Sci.*, vol. 180, no. 22, pp. 4370–4383, 2010.

[50] W. Ugaz, I. Sánchez, and A. M. Alonso, "Adaptive EWMA control charts with time-varying smoothing parameter," *Int. J. Adv. Manuf. Technol.*, vol. 93, nos. 9–12, pp. 3847–3858, 2017.

[51] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho, and S. W. Baik, "Improving electric energy consumption prediction using cnn and bi-lstm," *Appl. Sci.*, vol. 9, no. 20, p. 4237, 2019.

[52] P. Protopapas, J. M. Giammarco, L. Faccioli, M. F. Struble, R. Dave, and C. Alcock, "Finding outlier light curves in catalogues of periodic variable stars," *Monthly Notices Roy. Astronom. Soc.*, vol. 369, no. 2, pp. 677–696, 2010.

**MINGHU ZHANG** is currently pursuing the Ph.D. degree with the Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. His research interests include the Internet of Things, unmanned aerial vehicle communications, machine learning, and big data analysis.

**XIN LI** (SM'14) received the B.Sc. degree from Nanjing University, Nanjing, China, in 1992, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 1998. He has been a Professor with the Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences (CAS), Lanzhou, China, since 1999. He is currently the Director and a Professor with the National Tibetan Plateau Data Center, Institute of Tibetan Plateau Research, CAS. He has led the Watershed Allied Telemetry Experimental Research and the Heihe Watershed Allied Telemetry Experimental Research, which are comprehensive remote sensing experiments conducted sequentially in recent years with over 350 participants in China. His current research interests include land data assimilation, the application of remote sensing and geography information systems in hydrology and cryosphere science, and integrated watershed modeling. He is a Senior Member of the IEEE Geoscience and Remote Sensing Society, a member of the Global Energy and Water Exchanges (GEWEX) Scientific Steering Committee, the International Science Advisory Panel of Global Water Futures, the American Geophysical Union, and the American Meteorological Society, and the Vice President of GEWEX China. He was a recipient of the several honors by CAS and the Chinese government for his outstanding contribution.

**LILI WANG** received the Ph.D. degree from the Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China, in 2019. She is currently a Lecturer with the College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou, China. Her research interests include machine learning, data mining, and big data analysis.