ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 35, DECEMBER 2018, 1522–1532

• **Original Paper** •

Probabilistic Automatic Outlier Detection for Surface Air Quality Measurements from the China National Environmental Monitoring Network

Huangjian WU[1,3], Xiao TANG*[1], Zifa WANG[1,3], Lin WU[1], Miaomiao LU[1], Lianfang WEI[1], and Jiang ZHU[2]

[1]*State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029*, China*

[2]*International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029*, China*

[3]*University of Chinese Academy of Science, Beijing* 100049*, China*

(Received 20 March 2018; revised 22 May 2018; accepted 15 June 2018)

ABSTRACT

Although quality assurance and quality control procedures are routinely applied in most air quality networks, outliers can still occur due to instrument malfunctions, the influence of harsh environments and the limitation of measuring methods. Such outliers pose challenges for data-powered applications such as data assimilation, statistical analysis of pollution characteristics and ensemble forecasting. Here, a fully automatic outlier detection method was developed based on the probability of residuals, which are the discrepancies between the observed and the estimated concentration values. The estimation can be conducted using filtering—or regressions when appropriate—to discriminate four types of outliers characterized by temporal and spatial inconsistency, instrument-induced low variances, periodic calibration exceptions, and less PM_{10} than $PM_{2.5}$ in concentration observations, respectively. This probabilistic method was applied to detect all four types of outliers in hourly surface measurements of six pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) from 1436 stations of the China National Environmental Monitoring Network during 2014–16. Among the measurements, 0.65%–5.68% are marked as outliers, with PM_{10} and CO more prone to outliers. Our method successfully identifies a trend of decreasing outliers from 2014 to 2016, which corresponds to known improvements in the quality assurance and quality control procedures of the China National Environmental Monitoring Network. The outliers can have a significant impact on the annual mean concentrations of $PM_{2.5}$, with differences exceeding 10 μg m^{-3} at 66 sites.

Key words: probabilistic automatic outlier detection, air quality observation, low pass filter, spatial regression, bivariate normal distribution

Citation: Wu, H. J., X. Tang, Z. F. Wang, L. Wu, M. M. Lu, L. F. Wei, and J. Zhu, 2018: Probabilistic automatic outlier detection for surface air quality measurements from the China National Environmental Monitoring Network. *Adv. Atmos. Sci.*, **35**(12), 1522–1532, https://doi.org/10.1007/s00376-018-8067-9.

1. Introduction

Surface pollutant measurements are fundamental in air quality as they provide crucial information for validating theoretical concepts, testing numerical models, and enabling applications such as data assimilation and ensemble forecasting. However, outliers can occur despite the quality assurance and quality control procedures applied by most air quality networks, consequently posing a considerable challenge for the various uses of surface pollutant measurements. Manual inspection (Fiebrich et al., 2010) is an effective choice to identify these outliers. However, this becomes cumbersome when facing large amounts of data, which makes it inapplicable to near real-time applications. Another limitation lies in the fact that data quality after manual inspection varies from operator to operator. Here, a probabilistic automatic method is proposed to detect outliers from China National Environmental Monitoring Center (CNEMC) surface pollutant measurements.

The CNEMC started to monitor the concentrations of six air pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) from 2012. By March 2017, the China National Environmental Monitoring Network (CNEMN) had included 1436 monitoring sites from 369 cities. Real-time hourly observations of the six pollutants at every monitoring site are uploaded to the CNEMC and released to the public (http://www.cnemc.cn/). These real-time observations are being assimilated into the chemical transport model at the CNEMC to improve air quality forecasts. They are also an important data source for the Ministry of Environmental Protection of the People's Republic of China to assess urban air quality and establish air

* Corresponding author: Xiao TANG
Email: tangxiao@mail.iap.ac.cn

quality control strategies. The observations are widely used in many research fields, such as data assimilation (Tang et al., 2016), model verification (Wang et al., 2016), health impact (Li et al., 2017), satellite product verification (Gu et al., 2017), and analyzing the formation processes of pollution episodes (Shan et al., 2009; Zheng et al., 2015).

The quality of ambient air quality datasets is vital for interpreting results in related research fields. However, instrument malfunctions, the influence of harsh environments and the limitation of measuring methods can cause outliers in the observed datasets. For example, wearing, insufficient air tightness and excess axial resistance of the pump of the monitoring instrument will interfere with the measured values (Luo, 2016). A blockage of the filter core or a breakage or running out of filter tapes can lead to significant errors in the monitoring of particulate matter. After rainstorms, observed concentrations of particulate matter might be negative if the mass of particulate matter accumulated is less than the water evaporated. Malfunction of the cooling system or the photomultiplier tube will cause NO_2 outliers, while instable infrared sources will cause CO outliers (Guan, 2016).

Outlier detection involves separating inconsistent observations from regular ones, based on statistical or physical criteria that characterize the regular or inconsistent observations (Aggarwal, 2016). Outlier detection has been widely applied to a diverse range of environmental data. In meteorology, outliers can be detected for those observations that are inconsistent in space and time (You et al., 2008; Steinacker et al., 2011; Liao et al., 2014), or against the characteristics of observed variables such as surface wind, precipitation and snowfall (Golz et al., 2005; Durre et al., 2010; Jiménez et al., 2010). There are also some detailed studies about outlier detection for variables in oceanography (e.g., temperature and salinity profiles) and soil science (e.g., temperature and moisture) (Ingleby and Huddleston, 2007; Fiebrich et al., 2010; Dorigo et al., 2013).

Outliers in air quality are difficult to detect, because pollutants show particularly rich patterns of variations in space and time on multiple scales. These variations are driven by complex processes of chemical reactions, atmospheric transport, emissions and depositions, and thus cannot be easily represented by either statistical models or chemistry transport models. Recent outlier detection studies in air quality mainly probe the inconsistency in space and time in pollutant concentration observations from clusters of stations. For instance, Kracht et al. (2014) identified outliers as daily PM_{10} measurements that manifest extreme values compared with the smoothed values—the weighted averages from neighboring background stations in the European air quality database AirBase; whereas, Bobbia et al. (2015) compared hourly PM_{10} measurements with the weighted median from nearest-neighbor stations or with geostatistical spatial predictions using classical kriging techniques. Araki et al. (2017) performed leave-one-out predictions of daily mean $PM_{2.5}$ and NO_2 in Japan using ordinary kriging of residuals from a land use regression model, and identified observations with large kriging errors as outliers. Interestingly, Čampulová et al. (2015) proposed an automatic procedure to mark out potential outliers from two urban stations measuring hourly PM_{10} concentrations by checking whether observations lie in a plausible interval obtained by analyzing residuals of data smoothing.

The main challenge for outlier detection in air quality is that the true spatiotemporal variations of pollutants can only be estimated to serve as detection criteria to separate outliers from signals, while all estimations have limitations. For instance, in kriging estimation, a more detailed diurnal covariance structure can be introduced (Wu et al., 2010), but covariance modelling in high dimensional space is a very difficult issue (Bickel and Levina, 2008). In practice, manual inspection (Fiebrich et al., 2010) is still an effective choice to identify outliers. A majority (80%) of outliers were identified by manual inspection at four sites in Shanghai in China from 2014 to 2016 (Guan, 2016). In this paper, outliers are classified into four types based on their characteristics. Then, an automatic outlier detection framework is proposed based on the probability of residuals between the observations and their estimation discriminating the known characteristics of different types of outliers. Our hope is that this probabilistic outlier detection method will help in routinely identifying outliers in pollutant observations from the CNMEC in an automatic manner such that cumbersome real-time manual inspections can be avoided. Subsequently, better datasets could be constructed to support various research aims related to the severe air pollution problem that is currently a national concern in China.

2. Methods

2.1. *Outlier classification*

Complex sources of errors in observational datasets make it difficult to identify the cause of errors based on the observed data itself (Sciuto et al., 2013). Therefore, outliers are classified into four types based on their characteristics:

(1) Spatially and temporally inconsistent outliers (ST-outliers). Similar to meteorological parameters (Steinacker et al., 2011), the scale of pollution phenomena exceeds that resolved by the observational network, and one can expect the measured concentrations to be smooth both in time and space. Observations that differ greatly from values observed at the adjacent time or in neighboring areas are defined as ST-outliers.

(2) Low variance outliers (LV-outliers). This type of outlier has a very low variance in time series compared to neighboring sites. Some LV-outliers do not change over time and can be observed when the pump of the instruments is stuck or the filter tape is depleted. Most outliers that change very slowly come from CO observations measured by the pressure difference between two chambers. These outliers can be observed when the aging of the light sources in the two chambers is not synchronized (Luo, 2016).

(3) Periodic outliers (P-outliers). This type of outlier usually appears every 24 h, as shown in Fig. 1e. For some in-

struments, the accuracy may decrease with time due to the ageing of light sources and the changing of the ambient environment. Regular calibration is required for such instruments. However, the calibration processes may interfere with the observations, thus inserting abnormal values into the online measurement datasets.

(4) Lower PM_{10} than $PM_{2.5}$ outliers (LP-outliers). This type of outlier involves $PM_{2.5}$ concentrations being higher than PM_{10} concentrations observed at the same hour and same site. Most $PM_{2.5}$ monitors in the CNEMN network were installed around a decade after the PM_{10} monitors, with more advanced instruments used for $PM_{2.5}$ monitoring, and the loss of semi-volatile components of particulate matter is better handled (see section 2.7 for details). Therefore, $PM_{2.5}$ data are more reliable than PM_{10} data, and PM_{10} data are marked as LP-outliers if the observed concentration of $PM_{2.5}$ is higher than that of PM_{10}.

Figure 1 shows examples of classified outliers. It is important to note that the four types of outliers are not exclusive. Some P-outliers are also ST-outliers, and some LP-outliers are also LV-outliers.

2.2. *Probabilistic automatic outlier detection*

The most common technique for identifying outliers in meteorological data is the z-score method (Lanzante, 1996; Feng et al., 2004; Durre et al., 2010). This method normalizes the observed values using their mean and standard deviation, then removes those values whose z-score exceeds a specified threshold. The z-score is calculated as follows:

$$z(i) = \frac{|f(i) - \bar{f}|}{\sigma}, \tag{1}$$

where $f(i)$ and $z(i)$ are the observed value and z-score at time i; $\bar{f}$ and σ are the mean and standard deviation of the observed values.

The z-score method can identify outliers that are considerably deviated from most observations. The limitation of this method is the normal distribution assumption made for observations (Durre et al., 2010), when in fact pollutant concentrations are always positive and their distributions are known to be closer to lognormal (Leiva et al., 2008). To deal with this limitation and combine detection methods by probability theory, three modifications are made to the z-score method:

(1) Instead of directly assessing the pollutant concentration observations, the residuals between the observed and the estimated concentration values are evaluated:

$$R(i) = f(i) - F(i). \tag{2}$$

Here, $F(i)$ is the estimated concentration at time i. Such an estimation can be conducted using either filters or regression

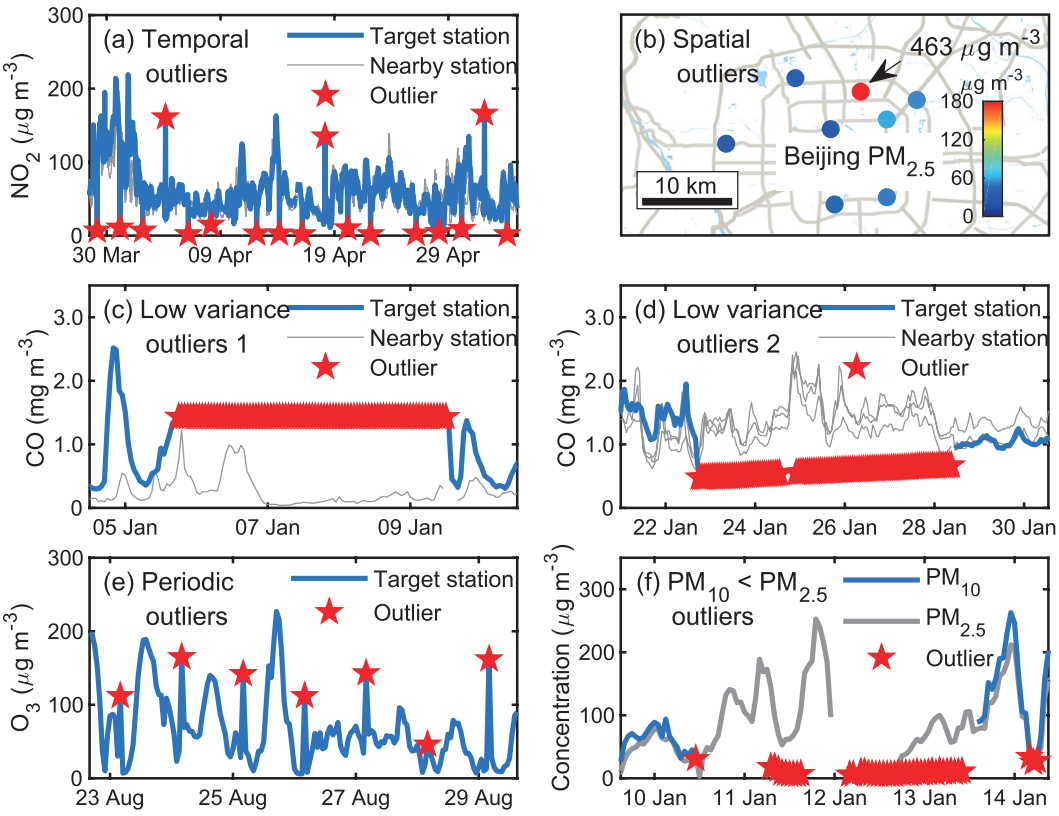

Fig. 1. Samples of outliers. (a, b) Spatiotemporal outliers have large differences with neighboring observations in time and space. (c, d) Low variance outliers either stay the same or change abnormally slowly in time and differ significantly with observations from nearby sites. (e) Periodic outliers appear periodically, usually every 24 h. (f) $PM_{10} < PM_{2.5}$ outliers are the PM_{10} observations that are lower than the $PM_{2.5}$ observations at the same time and site.

models. Observations with large residuals are more prone to be marked as outlier candidates.

(2) The standard deviation of the residuals is computed within a sliding window and updated constantly so that the outlier detection method will be more sensitive to local outliers. The standard deviation S of the residuals is calculated by

$$S(i) = \sqrt{\frac{\sum_{k=-n}^{n} R(i+k)^2}{2n}}\,, \tag{3}$$

where $i-n$ and $i+n$ are the start and end of the sliding window.

Substituting the observed values f and the standard deviation σ in Eq. (1) by the residual R and its standard deviations S, it becomes:

$$Z(i) = \frac{|R(i)-\bar{R}|}{S(i)}\,. \tag{4}$$

Generally, the mean of residuals $\bar{R}$ should be zero, and the calculation of probability is not affected by the sign of Z. The numerator $|R(i)-\bar{R}|$ can then be simplified to $R(i)$; therefore,

$$Z(i) = \frac{R(i)}{S(i)}\,. \tag{5}$$

(3) Instead of the z-score values, the probability of Z is introduced as a criterion to determine whether or not an observation is abnormal. The z-values in Eq. (5) are set to be normally distributed, and its probability is calculated by

$$P(i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z(i)^2}\,. \tag{6}$$

The normal distribution is chosen because it is a central distribution appropriate for residuals; plus, among all distributions with a given mean and a given variance, the normal distribution maximizes the entropy and is thus least informative. The introduction of the probability provides a framework based on which multiple rules for identifying abnormal observations can be combined (see sections 2.4.3 and 2.5). Next, we provide further details on how to use this probabilistic automatic outlier detection (PAOD) method to identify different types of outliers in surface observations of air pollutants.

2.3. *Detection of outliers with large errors*

The first detection involves identifying and removing outliers with large observational errors. These outliers might increase the residuals of normal observations and decrease the residuals of outliers. They make it more difficult to identify outliers with small observational errors, and should be removed before other detections. The detection and removal of outliers with large errors consists of the following two steps:

(1) Firstly, outliers exceeding the measurable range of the instrument are removed. The measurable range of the instrument for the six pollutants are specified by China National Standards (HJ653-2013, HJ654-2013), as listed in Table 1.

(2) In the second step, the PAOD method described in section 2.2 is applied to further identify outliers with large observational errors. Here, the estimated values are calculated by a median filter [Eq. (7)], which is less likely to be affected by the outliers:

Table 1. Measurement ranges of the monitors used in the CNEMC's air quality monitoring network.

Pollutant	Measurement Range
$PM_{2.5}$	0–1000 μg m^{-3} or 0–10000 μg m^{-3}
PM_{10}	0–1000 μg m^{-3} or 0–10000 μg m^{-3}
SO_2	0–1428 μg m^{-3}
NO_2	0–1026 μg m^{-3}
CO	0–62.5 mg m^{-3}
O_3	0–1071 μg m^{-3}

Note: The ranges are specified by environmental protection standards (HJ653-2013, HJ654-2013). There are two ranges for the measurement of particulate matter, because both the BAM and TEOM measurement methods can be applied.

$$F_{\rm m}(i) = \mathcal{M}(f(i+k)), \quad k \in [-n, n] \tag{7}$$

where $F_{\rm m}$ is the value estimated by the median filter at time i; $\mathcal{M}$ is the median function; $i-n$ and $i+n$ represent the start and end of the sliding window. The length of the sliding window, $2n+1$, is set to one month.

The residual $R_{\rm m}$ is obtained by substituting $F_{\rm m}$ into Eq. (2). The standard deviation of the residual $S_{\rm m}$ is calculated using the median absolute deviation (MAD), as follows:

$$S_{\rm m}(i) = 1.4826\mathcal{M}(|R_{\rm m}(i+k)|), \quad k \in [-n, n] \tag{8}$$

Compared with the conventional method described by Eq. (3), obtaining the standard deviation by MAD is more robust to outliers (Dunn et al., 2012). The probability $P_{\rm m}$ can be calculated through using the regression residual $R_{\rm m}$, its standard deviation $S_{\rm m}$ and Eqs. (5) and (6). The probability threshold is set to 10^{-15} after several sensitivity tests, and the data with probability $P_{\rm m}$ less than the threshold value are marked as outliers and removed from the datasets.

2.4. *Detection of ST-outliers*

After removing the outliers with large observational errors, spatiotemporal outlier detection is implemented to remove the ST-outliers described in section 2.1. To better identify these outliers, both the observed data of the target site at adjacent times and the data at neighboring sites are simultaneously used. The spatial and temporal residuals are assumed to follow the bivariate normal distribution, which makes it convenient to combine the estimations from both temporal and spatial estimation models.

2.4.1. *Temporal consistency estimation*

The estimation of pollutant series of temporal consistency is carried out using a low-pass filter:

$$F_{\rm t}(i) = \sum_{k=-15}^{15} f(i-k)h(k)\,, \tag{9}$$

Table 2. Filter coefficients and the time shift for the low-pass filter used in the temporal consistency estimation.

Time shift (k)	Filter coefficient (h)	Time shift (k)	Filter coefficient (h)	Time shift (k)	Filter coefficient (h)	Time shift (k)	Filter coefficient (h)
−15	−0.00053	−7	0.015445	1	0.130196	9	−0.00270
−14	−0.00139	−6	0.031918	2	0.117345	10	−0.00546
−13	−0.00275	−5	0.052580	3	0.098184	11	−0.00560
−12	−0.00436	−4	0.075568	4	0.075568	12	−0.00436
−11	−0.00560	−3	0.098184	5	0.052580	13	−0.00275
−10	−0.00546	−2	0.117345	6	0.031918	14	−0.00139
−9	−0.00270	−1	0.130196	7	0.015445	15	−0.00053
−8	0.003967	0	0.134722	8	0.003967		

where $F_t(i)$ is the estimated value at time i, f is the observed value at the target site, and $h(k)$ represents the low-pass filter coefficient (listed in Table 2). The filter coefficients are calculated following the algorithm designed by Karam and McClellan (1995), and the passband and stopband frequency is set to 1/8 and 1/24 h, respectively. The low-pass filter tends to preserve the low-frequency signals of normal variations from atmospheric chemistry and restrain the abnormal high-frequency signals accompanied by outliers. Compared with a moving average, it makes the residuals of normal observations smaller through giving higher weights to the data closer to the checkpoint.

2.4.2. *Spatial consistency estimation*

The pollutant series of spatial consistency is estimated by spatial regression:

$$F_s(i) = \sum_r \frac{f_r(i)a_r(i)c_r}{\sum_r a_r(i)c_r}, \tag{10}$$

where $F_s(i)$ is the estimated value at the checkpoint i; f_r is the observed value from neighboring site r; a_r and c_r are the index of agreement and localization coefficient between the target and a neighboring site.

Following the method of Legates and McCabe (1999), the index of agreement is defined as

$$a_r(i) = 1 - \frac{\sum_{k=-n}^{n} |f_r(i+k) - f(i+k)|}{\sum_{k=-n}^{n} [|f(i+k) - \bar{f}_r| + |f_r(i+k) - \bar{f}_r|]}, \tag{11}$$

where f_r is the observed value at neighboring sites, and $\bar{f}_r$ is the time-average of f_r within the sliding window. The length of the sliding window, $2n+1$, is also set to one month.

The index of agreement is often employed to evaluate the simulated results of models, and has been used in quality assurance as well (Durre et al., 2010). It provides a measure for both the covariation and the absolute difference between the data of two time-series. Compared with the correlation coefficient, it is less affected by outliers, but still affected by sampling errors. To deal with this problem, this paper borrows the idea of localization from data assimilation and adopts a localization method introduced by Gaspari and Cohn (1999). The method reduces the influence from remote sites, and the localization coefficient is calculated as follows:

$$c_r = \begin{cases} -\frac{1}{4}\left(\frac{|d|}{d_c}\right)^5 + \frac{1}{2}\left(\frac{|d|}{d_c}\right)^4 + \frac{5}{8}\left(\frac{|d|}{d_c}\right)^3 - \frac{5}{3}\left(\frac{|d|}{d_c}\right)^2 + 1, & 0 \leqslant |d| \leqslant d_c \\ \frac{1}{12}\left(\frac{|d|}{d_c}\right)^5 - \frac{1}{2}\left(\frac{|d|}{d_c}\right)^4 + \frac{5}{8}\left(\frac{|d|}{d_c}\right)^3 + \frac{5}{3}\left(\frac{|d|}{d_c}\right)^2 - 5\left(\frac{|d|}{d_c}\right) + 4 - \frac{2}{3}\left(\frac{d_c}{|d|}\right), & d_c \leqslant |d| \leqslant 2d_c \\ 0, & 2d_c \leqslant |d| \end{cases}, \tag{12}$$

where d is the distance between the target and a neighboring site, and d_c is the characteristic length of localization.

2.4.3. *Combining temporal and spatial consistency estimations*

After obtaining the estimated values F_t in Eq. (9) and F_s in Eq. (10) using the observed data at adjacent times and the data at neighboring sites respectively, the corresponding residuals at the checkpoints can be calculated using Eq. (2). Then, the residuals are normalized to Z_t and Z_s using Eqs. (3) and (5). The spatial and temporal consistency are evaluated simultaneously under our PAOD framework, and compute the probability of Z_t and Z_s by a bivariate normal distribution:

$$P_{st}(i) = \frac{1}{2\pi\sqrt{1-\rho(i)^2}} e^{\left(-\frac{1}{2(1-\rho(i)^2)}[Z_t(i)^2 + Z_s(i)^2 - 2Z_t(i)Z_s(i)]\right)}, \tag{13}$$

where $P_{st}(i)$ is the joint probability at time i, and ρ is the correlation coefficient between Z_t and Z_s:

$$\rho(i) = \frac{\sum_{k=-n}^{n} [Z_t(i+k) - \overline{Z_t}][Z_s(i+k) - \overline{Z_s}]}{\sqrt{\sum_{k=-n}^{n} [Z_t(i+k) - \overline{Z_t}]^2 \sum_{k=-n}^{n} [Z_s(i+k) - \overline{Z_s}]^2}}. \tag{14}$$

Here, $\overline{Z_t}$ and $\overline{Z_s}$ are the means of Z_t and Z_s, respectively, within the sliding window; $i-n$ and $i+n$ are the start and end of the sliding window. Observations with low probability at the checkpoint are identified as outliers.

Figure 2 displays the outliers detected in two ways. One uses the joint probability of Z_t and Z_s, and the other only employs the probability of Z_t (only considering the temporal consistency). When only the temporal consistency is considered, the method identifies the outliers at checkpoint Nos. 1 and 2, but misses the outlier at No. 3 and misrecognizes the normal observations at Nos. 4 and 5 as outliers. After using the joint probability, the method deals with the above defects properly. This suggests that the detection method using the joint probability of the residuals improves the accuracy of the detection method. The two detection methods are applied to the raw data described in section 3. For $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3 observations, 37%–83% of the data marked with spatiotemporal inconsistency are also marked with temporal inconsistency. For some outliers with moderate temporal inconsistency but strong spatial inconsistency (such as observation 3 in Fig. 2), it can only be identified by the detection using spatiotemporal consistency. Among the data marked with spatiotemporal inconsistency, 46% of them are not marked by the detection using temporal inconsistency. For normal observations with strong spatial consistency but week temporal consistency (such as observations 1 and 2 in Fig. 2), it might be misidentified by the detection using temporal inconsistency. Among the data marked with temporal inconsistency, 54% of the them are not marked by the detection using spatiotemporal consistency.

2.5. *Detection of LV-outliers*

To deal with the LV-outliers that stay the same or change slowly, the low variance periods are first detected by checking the first and second derivatives of the observed values over time. As some low variance observations will be normal observations when the ambient air is clean or stable, the spatial consistency is combined to decide whether the data in those periods should be mark as outliers. The residual of a low variance period is an average of the residuals of the spatial consistency estimation in this period:

$$R_v = \frac{\sum_{i=b}^{e} R_s(i)}{(e-b+1)}, \tag{15}$$

where $R_s(i)$ is the residual of the spatial consistency estimation, as in section 2.4.2, and b and e are the beginning and end of this period. The standard deviation of R_v is calculated according to the standard error of the mean for samples that are normally distributed:

$$S_v = \frac{\sum_{i=b}^{e} S_s(i)}{(e-b+1)\sqrt{(e-b+1)}}, \tag{16}$$

where $S_s(i)$ is the standard deviation of the residuals of the spatial consistency estimation in section 2.4.2. Using Eqs. (5) and (6), the normalized residual Z_v and its probability P_v for the whole period can be obtained. If P_v is smaller than a predefined value (10^{-6} in this study), the data within the whole period are identified as LV-outliers.

2.6. *Detection of P-outliers*

P-outliers are mainly caused by the daily self-calibration of the instrument and appear every 24 h. According to this characteristic, the time series of the observed data within 11 days are firstly processed into diurnal-variation data:

$$f_p(i) = \frac{\sum_{k=-5}^{5} f(i+24k)}{11}, \tag{17}$$

where f is the hourly observed concentrations and f_p is the data after processing. Then, a median filter is applied:

$$F_p(i) = \mathcal{M}(f_p(i+k)), \quad k \in [-1,1] \tag{18}$$

where $F_p(i)$ is the estimated value and $\mathcal{M}$ is the median function.

Instead of Eq. (3), the standard deviation $S_p(i)$ of the residuals is calculated using the following method:

$$S_p(i) = g(R_p(i+k)), \quad k \in [-72,72] \tag{19}$$

where R_p is the residual and g is a function that finds the 93.75th percentiles. The advantage of using the 93.75th percentiles of the residuals is that the obtained standard deviation is the second largest residual in one day, and only the observation with the largest residual in a day might be identified as the P-outliers. Using Eqs. (5) and (6), the probability

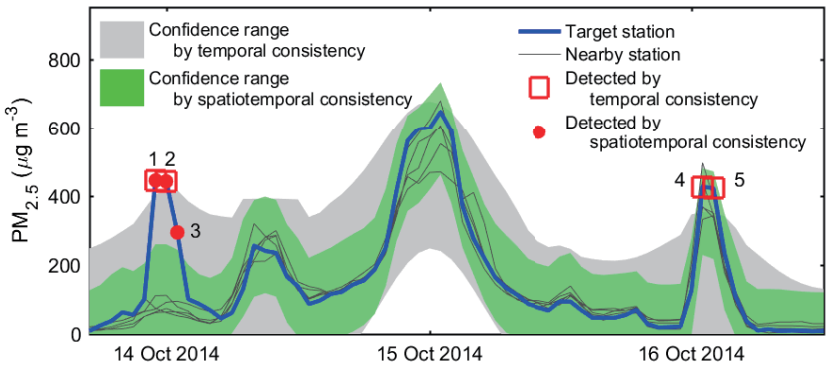

Fig. 2. Comparison of outlier detections based on temporal consistency and spatiotemporal consistency. The detection based on spatiotemporal consistency can detect outliers 1, 2 and 3 while preserving valid observations 4 and 5.

P_p of the residual can be obtained. If P_p is smaller than a predefined value (10^{-4} in our check), the corresponding observation will be identified as the P-outliers. Figure 3 gives an example of the P-outliers that shows a peak concentration of ozone at 0400 LST (LST=UTC+8) every day. It should be noted that most of the P-outliers present large differences from their neighboring observations and can be identified by the detection of ST-outliers in section 2.4. However, there are also some P-outliers that have a relatively small difference from the neighboring observations and need to be identified by the P-outlier detection.

2.7. *Detection of LP-outliers*

The last detection is to mark the outliers with an observed concentration of $PM_{2.5}$ higher than that of PM_{10} at the same hour and same site. This step is very simple but very important for the observed datasets from the China Nationwide Air Quality Monitoring Network. The $PM_{2.5}$ and PM_{10} are mainly measured by the beta attenuation monitoring method (BAM) or tapered element oscillating microbalance method (TEOM). Both methods use heaters to reduce the humidity of the sampled air to prevent fogging and inhibit particle growth under high humidity. However, the heating process may lead to the volatilization of semi-volatile organic compounds of particulate matter. To deal with this problem, most new monitors adopt a filter dynamic measurement system to measure the volatile portion of the sample air when the concentrations are measured by TEOM. Also, a smart heater is implemented when the concentrations are measured by BAM, to keep the relative humidity at around 35% instead of keeping the temperature at around 50°C. The $PM_{2.5}$ monitoring started from 2013, about a decade later than the PM_{10} monitoring. A filter dynamic measurement system or a smart heater is mandatory for the $PM_{2.5}$ measuring instruments, while they are optional for the PM_{10} measuring instruments. Also, neither of them is implemented for most PM_{10} measuring instruments that were established before 2013 (Pan et al., 2014). As a result, the $PM_{2.5}$ data are more reliable than the PM_{10} data, and the LP-outliers are more likely to occur at nighttime and on foggy days when the relative humidity is higher (Niu, 2017).

Most outliers or "bad data" in the observed PM_{10} data are LP-outliers before 2016. The percentage of "bad data" can be higher than 7% (Fig. 4). However, as measuring instruments and management have been upgraded, the percentage of "bad data" had reduced to about 1% in 2016, and will hopefully continue to decrease in the future.

3. Results

The PAOD method was applied to detect outliers in the raw data of the hourly observations of six pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) during 2014–16. The raw data were monitored by the CNEMN network. The network contains 1436 monitoring stations across China, and the monitored data are directly transmitted to the data center at the CNEMC. The raw data in this paper are hourly observations directly acquired from the data center at the CNEMC. The number of outliers identified by the method and their proportions are shown in Table 3. Among the raw data, 0.65%–5.68% are identified as outliers. There are more outliers in the PM_{10} and CO observations than other pollutants, accounting for 5.68% and 1.03% respectively. The NO_2 and SO_2 observations have fewer outliers than the other pollutants, accounting for only 0.65% and 0.73% respectively. For $PM_{2.5}$, SO_2, NO_2, CO and O_3 observations, the ST-outliers are the most frequent among the four types of outliers, accounting for 0.46%–0.77% of the raw data. The LV-outliers rank second, accounting for 0.09%–0.19%. The P-outliers are an important type of outlier for gaseous pollutants, especially CO and SO_2. For PM_{10} observations, the LP-outliers account for 4.73% of the raw data and are the main source of abnormal observations (see section 2.7 for why the $PM_{2.5}$ data are more reliable than the PM_{10} data).

Figure 4 shows the removal ratios of the six pollutants during the three years. The removal ratio is the fraction of raw data being labeled as outliers. During 2014–15, the PM_{10}

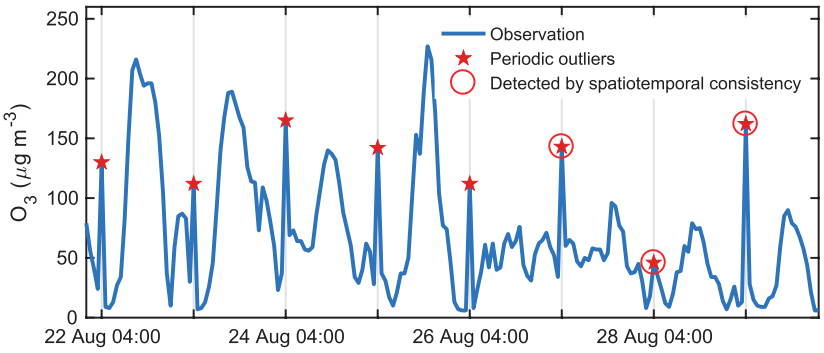

Fig. 3. Periodic outliers of O_3 observations detected for a site in Wuhan, China, between 22 and 30 August 2014. Some periodic outliers can be detected by spatiotemporal consistency, while others have relatively small measurement error compared to the variation of neighboring observations and can only be identified in the detection of periodic outliers.

Table 3. Number of hourly observations in China from 2014 to 2016, as well as number and ratio of outliers detected.

Pollutant	Number of raw records	Number of outliers	Ratio of outliers	Ratio of ST-outliers	Ratio of LV-outliers	Ratio of P-outliers	Ratio of LP-outliers
PM_{10}	33 124 620	1 879 899	5.68%	0.82%	0.14%	0.01%	4.73%
$PM_{2.5}$	34 105 146	312 121	0.92%	0.74%	0.19%	0.01%	0%
SO_2	34 166 889	248 844	0.73%	0.58%	0.09%	0.08%	0%
NO_2	34 150 069	222 934	0.65%	0.46%	0.19%	0.03%	0%
CO	32 915 111	337 996	1.03%	0.77%	0.19%	0.11%	0%
O_3	33 988 052	292 208	0.86%	0.68%	0.18%	0.03%	0%

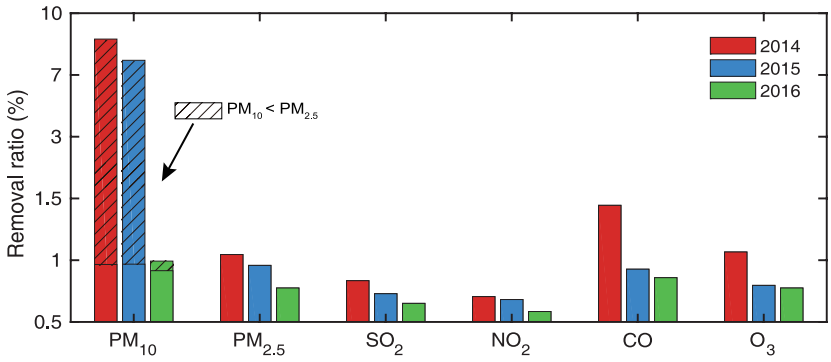

Fig. 4. Outlier ratio of all sites in China From 2014 to 2016. Outlier ratios decrease more or less for all the six pollutants from 2014 to 2016. Most of the reduction in the ratio of PM_{10} outliers comes from the decrease in $PM_{10} < PM_{2.5}$ outliers.

observations have a high removal ratio (more than 7%). Most of the removed data are LP-outliers. However, in 2016, the removed ratio decreases sharply to about 1%. This might be related to the implementation of a compensation algorithm for the loss of semi-volatile materials in the PM_{10} measurements. Interestingly, the removal ratios of the $PM_{2.5}$, SO_2, NO_2, CO and O_3 observations also decrease. This indicates that the data quality of the CNEMN has been improved from 2014 to 2016.

To evaluate the impact of the outlier detections on the observed concentrations, Fig. 5a compares the observed annual $PM_{2.5}$ concentrations before and after quality assurance in Shijiazhuang City during 2014–16. The results show a big difference of more than 150 μg m^{-3} for estimating the annual mean concentration in 2014, which is mainly caused by some outliers that exceed the measurement range. Figure 5b presents the diurnal variations of the observed O_3 concentrations before and after quality assurance at a site in Wuhan in 2015. Due to the P-outliers that occurred at 0400 LST, the raw data display a false peak at that time. The quality assurance reduces this false peak, and the data after quality assurance show more reasonable daily variations of O_3 concentrations.

Figure 6 displays the differences in annual concentrations caused by outliers in 2015. For $PM_{2.5}$, the differences are lower than 1 μg m^{-3} at most sites. However, there are 66 sites and 17 sites whose differences are greater than 10 μg m^{-3} and 50 μg m^{-3} respectively. For PM_{10}, the differences at most sites are within 1–10 μg m^{-3}, while there are 92 sites and 23 sites with differences greater than 10 μg m^{-3} and 50 μg m^{-3} respectively. For CO, big differences of more than 1 mg m^{-3} can be observed at 80 sites. For O_3 and NO_2, the differences are relatively small, and only a few stations (<20) have differences of more than 10 μg m^{-3}. For SO_2, there are 38 sites with differences of more than 10 μg m^{-3}. The above results suggest that outliers might lead to significant biases in the estimation of annual mean concentrations of these six pollutants. Identifying and removing outliers is a necessary step before using the online dataset.

4. Conclusions and future work

A POAD method is proposed to detect outliers for hourly surface concentration observations of six pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) from the 1436 stations of the CNEMN during 2014–16. This outlier detection method takes advantage of the known characteristics of outliers [temporal and spatial inconsistency (ST-outliers), instrument-induced low variances (LV-outliers), periodic calibration exceptions (P-outliers), and less PM_{10} than $PM_{2.5}$ in concentration observations (LP-outliers)] by computing the probability of residuals between the observations and the estimations that discriminate these known characteristics of outliers. The outlier detection process is fully automatic; hence, it will help in avoiding the cumbersome manual inspection of outliers when seeking to achieve reliable air quality data in the CNEMN network.

The outliers detected account for 0.65% to 5.68% of the observations for the six pollutants. PM_{10} observations have the most outliers, among which LP-outliers contribute the

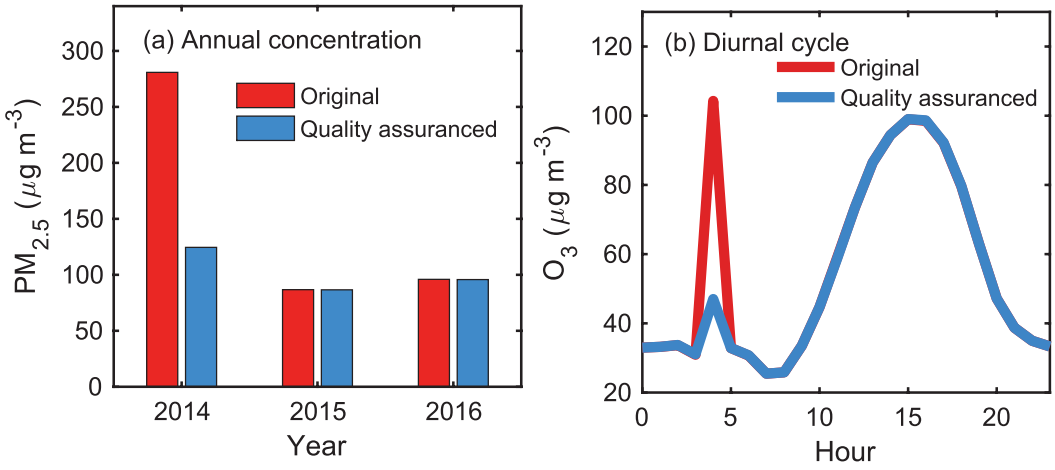

Fig. 5. (a) Annual $PM_{2.5}$ concentration in Shijiazhuang city before and after quality assurance. (b) Diurnal cycle of O_3 before and after quality assurance for a site in Wuhan, China, 2015.

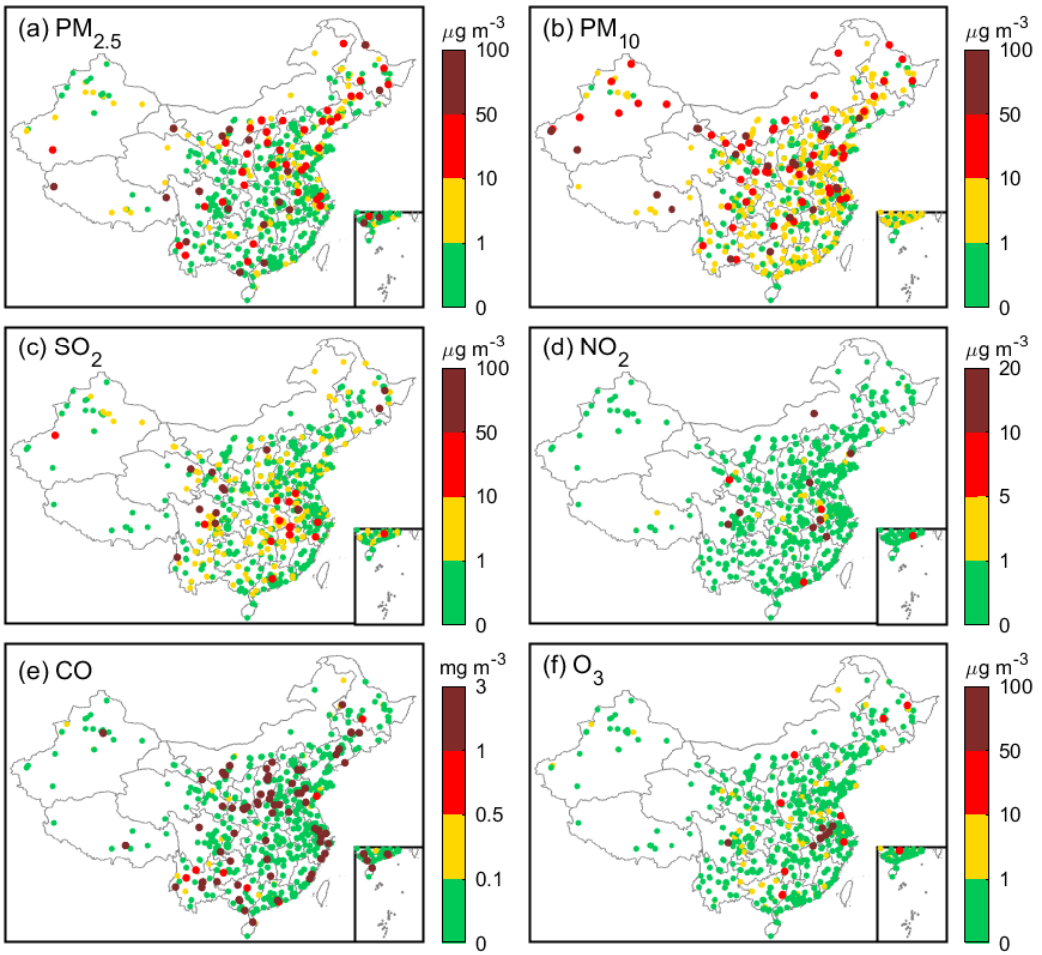

Fig. 6. Absolute difference in annual concentrations before and after quality assurance for six pollutants and all sites in 2015. Quality assurance has a significant effect on the annual concentration for some sites.

most (see section 2.7 for why the $PM_{2.5}$ data are more reliable than the PM_{10} data). The proportions of outliers in the six pollutants all decrease from 2014 to 2016, which suggests an improvement in the data quality of the CNEMN network. The impact of outliers is estimated by the difference in annual mean concentrations of $PM_{2.5}$ between the raw data and the data after outlier detection. The differences are less than 1 μg m^{-3} at most sites, but there are 66 sites whose differences

are greater than 10 μg m^{-3}. This suggests that outlier detection is essential before using the monitoring datasets, even for evaluation of the annual mean concentrations.

The outlier detection method was developed with the help of the CNEMC, which is also the institute responsible for releasing the real-time air quality data. The method has been used in the assimilation system of the CNEMC, and is going to be integrated into the data management system. Hopefully, outliers in the real-time air quality data will be removed by our method in the near future.

Although the application of the PAOD method has brought some positive and interesting results, improvements can still be made for better performance in further in-depth studies. First, the outlier detection method is performed separately for each species; however, these different pollutants are closely linked to one another by atmospheric chemistry. Further developments of the outlier detection method could take into consideration information from multiple pollutants at the same time to account for chemical transformations. Second, the method does not take into account the specific locations of the stations in the urban environment. Stations in the proximity of heavy traffic highways will have much stronger variations in observations than those in urban green spaces. Developing different parameters sets for different types of stations might improve the performance. Third, the outlier detection method assumes that outliers account for a small proportion of the raw data. However, at some stations, most of the raw data on a weekly basis can be bad data, providing little or no information on reals concentrations. Examination of the distributions within the raw data might help in such cases. Finally, this method uses the normal distribution to compute the probability of residuals. However, when more information is available, new distribution types should be tested to better adapt to the CNEMN dataset.

Acknowledgements. The authors express their sincere gratitude to the CNEMC for providing the air quality observations for the period 2014–16. This study was supported by the National Natural Science Foundation (Grant Nos. 91644216 and 41575128), the CAS Information Technology Program (Grant No. XXH13506-302) and Guangdong Provincial Science and Technology Development Special Fund (No. 2017B020216007).

REFERENCES

Aggarwal, C. C., 2016: *Outlier Analysis*. 2nd ed., Springer, Cham, 263 pp.

Araki, S., H. Shimadera, K. Yamamoto, and A. Kondo, 2017: Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan. *Atmos. Environ.*, **153**, 83–93, https://doi.org/10.1016/j.atmosenv.2016.12.057.

Bickel, P. J., and E. Levina, 2008: Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199–227, https://doi.org/10.1214/009053607000000758.

Bobbia, M., M. Misiti, Y. Misiti, J.-M. Poggi, and B. Portier, 2015: Spatial outlier detection in the PM_{10} monitoring network of Normandy (France). *Atmospheric Pollution Research*, **6**, 476–483, https://doi.org/10.5094/APR.2015.053.

Čampulová, M., P. Veselík, and J. Michálek, 2017: Control chart and Six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}. *Atmospheric Pollution Research*, **8**, 700–708, https://doi.org/10.1016/j.apr.2017.01.004.

Dorigo, W. A., and Coauthors, 2013: Global automated quality control of in situ soil moisture data from the international soil moisture network. *Vadose Zone Journal*, **12**, https://doi.org/10.2136/vzj2012.0097.

Dunn, R. J. H., K. M. Willett, P. W. Thorne, E. V. Woolley, I. Durre, A. Dai, D. E. Parker, and R. S. Vose, 2012: HadISD: A quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, **8**, 1649–1679, https://doi.org/10.5194/cp-8-1649-2012.

Durre, I., M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, **49**, 1615–1633, https://doi.org/10.1175/2010JAMC2375.1.

Feng, S., Q. Hu, and W. H. Qian, 2004: Quality control of daily meteorological data in China, 1951–2000: A new dataset. *International Journal of Climatology*, **24**, 853–870, https://doi.org/10.1002/joc.1047.

Fiebrich, C. A., C. R. Morgan, A. G. McCombs, P. K. Hall, and R. A. McPherson, 2010: Quality assurance procedures for mesoscale meteorological data. *J. Atmos. Oceanic Technol.*, **27**, 1565–1582, https://doi.org/10.1175/2010JTECHA1433.1.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Golz, C., T. Einfalt, M. Gabella, and U. Germann, 2005: Quality control algorithms for rainfall measurements. *Atmospheric Research*, **77**, 247–255, https://doi.org/10.1016/j.atmosres.2004.10.027.

Gu, J. B., and Coauthors, 2017: Ground-level NO_2 concentrations over China inferred from the satellite OMI and CMAQ model simulations. *Remote Sensing*, **9**, 519, https://doi.org/10.3390/rs9060519.

Guan, Q. Y., 2016: Judgment and handling of abnormal data during ambient air automatic monitoring data audit. *Environmental Monitoring and Forewarning*, **8**, 59–63, https://doi.org/10.3969/j.issn.1674-6732.2016.05.015 (in Chinese).

Ingleby, B., and M. Huddleston, 2007: Quality control of ocean temperature and salinity profiles—Historical and real-time data. *J. Mar. Syst.*, **65**, 158–175, https://doi.org/10.1016/j.jmarsys.2005.11.019.

Jiménez, P. A., J. F. González-Rouco, J. Navarro, J. P. Montávez, and E. Garcia-Bustamante, 2010: Quality assurance of surface wind observations from automated weather stations. *J. Atmos. Oceanic Technol.*, **27**, 1101–1122, https://doi.org/10.1175/2010JTECHA1404.1.

Karam, L. J., and J. H. McClellan, 1995: Complex Chebyshev approximation for FIR filter design. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, **42**, 207–216, https://doi.org/10.1109/82.372870.

Kracht, O., M. Gerboles, and H. I. Reuter, 2014: First evaluation of a novel screening tool for outlier detection in large scale ambient air quality datasets. *International Journal of Environment and Pollution*, **55**, 120–128, https://doi.org/10.1504/IJEP.2014.065912.

Lanzante, J. R., 1996: Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology*, **16**, 1197–1226, https://doi.org/10.1002/(SICI)1097-0088(199611)16:11<1197::AID-JOC89>3.0.CO;2-L.

Legates, D. R., and G. J. McCabe, 1999: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.*, **35**, 233–241, https://doi.org/10.1029/1998WR900018.

Leiva, V., M. Barros, G. A. Paula, and A. Sanhueza, 2008: Generalized Birnbaum-Saunders distributions applied to air pollutant concentration. *Environmetrics*, **19**, 235–249, https://doi.org/10.1002/env.861.

Li, H. M., and Coauthors, 2017: Chemical partitioning of fine particle-bound metals on haze–fog and non-haze–fog days in Nanjing, China and its contribution to human health risks. *Atmospheric Research*, **183**, 142–150, https://doi.org/10.1016/j.atmosres.2016.07.016.

Liao, J., B. Wang, and Q. X. Li, 2014: A new method for quality control of Chinese rawinsonde wind observations. *Adv Atmos Sci*, **31**, 1293–1304, https://doi.org/10.1007/s00376-014-4030-6.

Luo, M., 2016: Quality control research of air pollutant hourly monitoring data. M.S thesis , Dept. of School of Geographic Sciences, East China Normal University (in Chinese).

Niu, G., 2017: Features and cause analysis of abnormal data of particulate matter in ambient air monitoring. *Anhui Chemical Industry*, **43**, 103–105, https://doi.org/10.3969/j.issn.1008-553X.2017.02.033 (in Chinese).

Pan, B.F., H. H. Zheng, L. N. Li, and W. Wang, 2014: The characteristic and reason about the reversal between $PM_{2.5}$ and PM_{10} in ambient air quality automatic monitoring. *Environmental Monitoring in China*, **30**, 90–95 (in Chinese).

Sciuto, G., B. Bonaccorso, A. Cancelliere, and G. Rossi, 2013: Probabilistic quality control of daily temperature data. *International Journal of Climatology*, **33**, 1211–1227, https://doi.org/10.1002/joc.3506.

Shan, W. P., Y. Q. Yin, H. X. Lu, and S. X. Liang, 2009: A meteorological analysis of ozone episodes using HYSPLIT model and surface data. *Atmospheric Research*, **93**, 767–776, https://doi.org/10.1016/j.atmosres.2009.03.007.

Steinacker, R., D. Mayer, and A. Steiner, 2011: Data quality control based on self-consistency. *Mon. Wea. Rev.*, **139**, 3974–3991, https://doi.org/10.1175/MWR-D-10-05024.1.

Tang, X., J. Zhu, Z. F. Wang, A. Gbaguidi, C. Y. Lin, J. Y. Xin, T. Song, and B. Hu, 2016: Limitations of ozone data assimilation with adjustment of NO_X emissions: Mixed effects on NO_2 forecasts over Beijing and surrounding areas. *Atmospheric Chemistry and Physics*, **16**, 6395–6405, https://doi.org/10.5194/acp-16-6395-2016.

Wang, L. T., Y. Zhang, K. Wang, B. Zheng, Q. Zhang, and W. Wei, 2016: Application of Weather Research and Forecasting Model with Chemistry (WRF/Chem) over northern China: Sensitivity study, comparative evaluation, and policy implications. *Atmos. Environ.*, **124**, 337–350, https://doi.org/10.1016/j.atmosenv.2014.12.052.

Wu, L., M. Bocquet, and M. Chevallier, 2010: Optimal reduction of the ozone monitoring network over France. *Atmos. Environ.*, **44**, 3071–3083, https://doi.org/10.1016/j.atmosenv.2010.04.012.

You, J. S., K. G. Hubbard, and S. Goddard, 2008: Comparison of methods for spatially estimating station temperatures in a quality control system. *International Journal of Climatology*, **28**, 777–787, https://doi.org/10.1002/joc.1571.

Zheng, B., and Coauthors, 2015: Heterogeneous chemistry: A mechanism missing in current models to explain secondary inorganic aerosol formation during the January 2013 haze episode in North China. *Atmospheric Chemistry and Physics*, **15**, 2031–2049, https://doi.org/10.5194/acp-15-2031-2015.