Front. Earth Sci. 2015, 9(3): 381–393
DOI 10.1007/s11707-014-0483-2

RESEARCH ARTICLE

Quality control of specific humidity from surface stations based on EOF and FFT—Case study

Hong ZHAO[1], Xiaolei ZOU (✉)[2], Zhengkun QIN[1]

1 Center of Data Assimilation for Research and Application, Nanjing University of Information Science & Technology, Nanjing 210044, China
2 Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20740-3823, USA

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2014

Abstract Comparisons between observations and background fields indicate that amplitude and phase differences in oscillations result in a non-Gaussian distribution in observation minus background vectors (OMB). Empirical Orthogonal Function (EOF) quality control (QC) and Fast Fourier Transform (FFT) quality control are proposed from the perspective of data assimilation and are applied to the surface specific humidity from ground-based stations. The QC results indicate that the standard deviation between observations and background is reduced effectively, and the frequency distribution for the observation increment is closer to a normal distribution. The specific humidity outliers occur primarily in mountainous and coastal regions. Comparing the two QC methods, it is found that the EOF QC performs better than the FFT QC as it can keep large scale of fluctuation information from the original field, preventing these waves from entering into the residual field and being removed by the QC process.

Keywords specific humidity, quality control, EOF, FFT

1 Introduction

Surface observations from manual and automatic surface stations are important data sources for weather and climate research as they have high space-time resolution and can record distinctive meso-scale weather phenomena in real-time (Lanzante, 1996). Estimating surface information correctly is especially important for the initialization of short-medium range numerical weather prediction (NWP) with high resolution (Mohanty et al., 1986; Hong and Pan, 1996; Ha et al., 2007). However, the quality of the observation data has been greatly reduced because the surface observations are easily affected by the station positions, observing instruments, human factors and so on (Hubbard and You, 2005). Therefore, it is important to carry out quality control (QC) before data application.

Errors in surface station data can be broadly divided into three categories: gross error, systematic error and random error (Gandin, 1988). The most widespread QC methods include limit checks (Wade, 1987), internal consistency checks (Reek et al., 1992), spatial consistency checks (Eischeid et al., 1995) and temporal consistency checks (Shafer et al., 2000). These methods can effectively identify gross errors in surface stations (Baker, 1992; Feng et al., 2004). For data assimilation, the generalized QC method is background consistency checks, using this method, observations will be assimilated only when the differences between background field and observations (O-B, OMB) do not exceed a given threshold value (Ghil and Malanotte-Rizzoli, 1991). This approach can effectively avoid a lack of coordination between observations and background field in NWP; however, data reflecting the abnormal state of the atmosphere is also removed (Wang et al., 2013). Moreover, errors are always assumed to be of Gaussian type in most variational data assimilation systems (Lorenc and Hammon, 1988; Ingleby and Lorenc, 1993; Anderson and Järvinen, 1999), and data after OMB QC may not line up with the Gaussian assumption (Qin et al., 2010). On the other hand, water vapor plays an important role in the radiation and water cycle. Surface observations can provide a reasonable and accurate description of large-scale water vapor analysis. The thermal effects of water vapor can affect local convections and turbulences (Ramanathan et al., 1989), and the quality of the water vapor will impact the assimilation results in NWP (Mohanty et al., 1986). Specific humidity at surface sites is selected as the study object in this paper, and the paper is arranged as follows. Section 2 introduces the data and methods. Section 3 compares observations and background fields. Then, a bi-weighted standard deviation QC method is applied to residual fields on the basis of

Received March 25, 2014; accepted August 4, 2014

E-mail: xzou1@umd.edu

Empirical Orthogonal Function (EOF) and Fast Fourier Transform (FFT). Section 4 analyzes QC results and comparisons between them. Section 5 summarlizes and presents discussion and directions for future work.

2 Data and method

2.1 Data

An extraordinarily brutal snowstorm struck China in January 2008, causing abnormal low temperatures, excessive precipitation, and severe snow and ice disasters over most of China. The specific humidity data of January 2008 over China were chosen for this study. There are 3,340 station observations in the area between 0 and 70°N, 55°E−145°E from SYNOP (surface synoptic observations) after conventional QC with three-hour temporal resolution. The observed surface specific humidity can be expressed as follows:

$$q_{k,n}^{OBS},\ k = 1,2,...,3340,\ n = 1,2,...,239,$$

where k is the station index and n indicates the observing time.

The National Centers for Environmental Prediction (NCEP) Final (FNL) Operational Global Analysis data (Kalnay et al., 1996) were chosen as the background field, with a 6-h temporal resolution and a $1°\times 1°$ spatial resolution. Because the spatial and temporal resolutions of the data from the background field are different from the station observations, the LaGrange polynomial interpolation method was employed to interpolate the grid data into the stations spatially. As Zou and Qin (2010) have noted the data after time interpolation would not affect its cycle characteristics, so the data at 0300 UTC, 0900 UTC, 1500UTC and 2100 UTC can be obtained by the cubic spline interpolation method. Then, the background specific humidity can be represented as follows:

$$q_{k,n}^{BF},\ k = 1,2,...,3340,\ n = 1,2,...,239.$$

2.2 Methods used to analyze specific humidity

Power spectral analysis can decompose the total power into different frequencies (Wei, 2007). Then, the major cycle can be determined by those fluctuations. The major oscillation should pass the standard spectrum test, which would mean the period is significant. Figure 1 shows the power spectrum density of specific humidity from observations (q^{OBS}) and background fields (q^{BF}). Data of the first eight days in January 2008 were used in the research. It can be observed that there are three distinct periodic oscillations: diurnal, semidiurnal and 8-h oscillations, all of which have passed the 95% confidence level. The diurnal cycle is the strongest, and 8-h oscillation is the weakest. In addition, the intensity of the diurnal oscillation is slightly weaker in the observations than in the background field. The data for other days in January 2008 yield the similar results.

The Fast Fourier Transform is another method to determine the major oscillation of meteorological variables (Cooley and Tukey, 1965). To meet the requirements of operational application, the data used in each process should not be too large. Thus every 64 times data (eight days) were chosen as a sample. In total, there are 176 samples (l_n = 1, 2,…, 176) for 239 observing times. All the

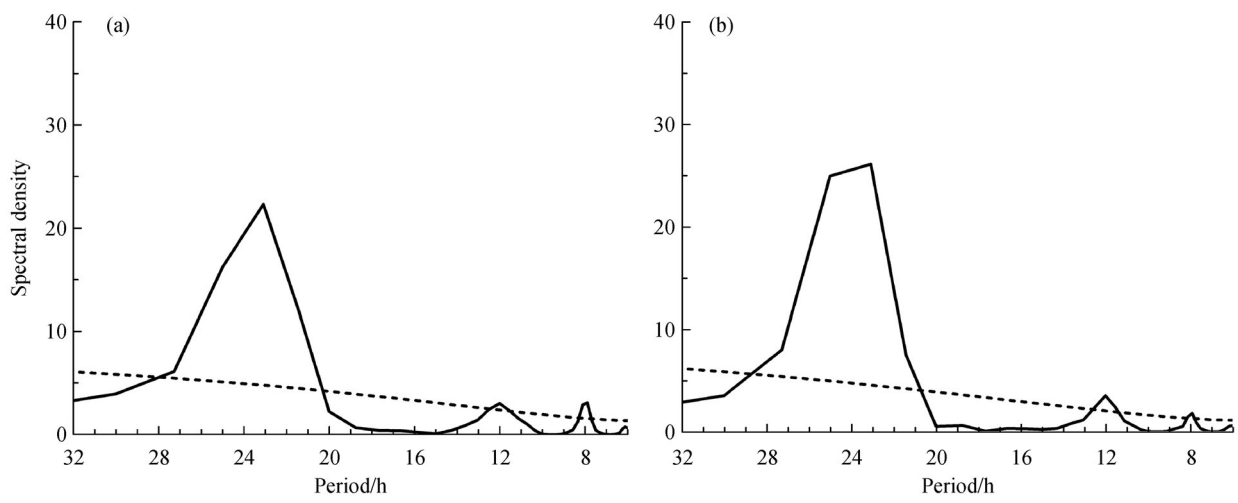

Fig. 1 Power spectrum density of surface specific humidity from observations (a) and background field (b), in 55°E−145°E, 0−70°N. The 95% confidence level is shown (dashed line). Data are from the first eight days of January 2008.

analyses and computations are based on these samples. After FFT, the amplitude of q can be expressed by the following:

$$a_{k,m} = \sum_{j=0}^{N-1} q_{k,j} \mathrm{e}^{-i2\pi m \frac{j}{N}},\ m = 0,1,...,N-1, \tag{1}$$

where k is the station index, m is the wavenumber, j is the observing time, and $N = 64$, namely eight days.

The original field can be calculated by inverse FFT:

$$q_{k,j} = \frac{1}{N} \sum_{m=0}^{N} a_{k,m} \mathrm{e}^{i2\pi j \frac{m}{N}},\ j = 0,1,...,N-1, \tag{2}$$

Thus, the specific humidity q can be divided into two parts, i.e., $q_{FFT_{reb}}$ and $q_{FFT_{res}}$. $q_{\mathrm{FFT}_{reb}}$ includes the average state and three main oscillations, m in Eq. (2) is the wavenumber, as surface specific humidity has been decomposed by FFT every 64 times, namely, eight days. $m = 0$ indicates the average state in specific humidity. $m = 8$ means that there are eight waves in eight days, which corresponds to the diurnal oscillations. By parity of reasoning, m = 16 corresponds to the semidiurnal oscillations and m = 24 represents the 8-h oscillations. Therefore, m in Eq. (2) can be taken as 0, 8, 16, 24, which can include the main oscillations in specific humidity, while $q_{FFT_{res}}$ represents other unapparent cycles.

EOF can also separate specific humidity into two parts: $q_{EOF_{reb}}$ and $q_{EOF_{res}}$. Different from FFT, which extract the period characteristics from the original field, the EOF analysis can separate the main space distribution structures from the original field and obtain a few unrelated typical modes, which can replace the original variable field. Each mode contains sufficient information of the original field (Lorenz, 1956). In recent years, EOF has been considered for quality control for varieties of data (Li et al., 2004; Zou et al., 2012; Zhao et al., 2013).

EOF is expanded for the space and time domain, which can be expressed as the following matrix:

$$\boldsymbol{X} = \begin{bmatrix} q_{1,l_n} & q_{2,l_n} & \cdots & q_{k,l_n} \\ q_{1,l_n+1} & q_{2,l_n+1} & \cdots & q_{k,l_n+1} \\ \vdots & \vdots & \ddots & \vdots \\ q_{1,l_n+63} & q_{2,l_n+63} & \cdots & q_{k,l_n+63} \end{bmatrix}, \tag{3}$$

where q_{k,l_n+n-1} indicates the specific humidity at the kth surface station and n^{th} time of the $l_n{}^{\mathrm{th}}$ sample.

Then, the original data can be written as follows:

$$q_{k,l_n+n-1} = \sum_{m=1}^{64} e_{n,m}^{l_n} c_{m,k}^{l_n},\ n = 1,2,...,64, \tag{4}$$

where $e_{n,m}^{l_n}$ is the eigenvector reflecting the time information; $e_m^{l_n} = (e_{1,m}^{l_n}, e_{2,m}^{l_n}, \cdots, e_{64,m}^{l_n})^{\mathrm{T}}$ represents the principal components (PCs); and $c_{m,k}^{l_n}$ represents the EOF modes which show the space layouts. The extent of the ith EOF mode describing the original variable field can be explained by the ratio of the ith variance to the sum of all variance ($\lambda_i^{l_n} / \sum_{i=1}^{64} \lambda_i^{l_n}$), as the sum of all eigenvalues is equal to the total variance of the data.

2.3 QC methods

Most of the gross errors can be eliminated by conventional QC methods, but the random error, system errors, and other small errors are difficult to address by these QC methods. The distributions of most meteorological variables are usually assumed to be Gaussian in normal conditions, but a Gaussian distribution is not suitable for surface variables such as surface air temperature and specific humidity because of the dominating diurnal and semidiurnal oscillations. Errors induced by inconsistent descriptions of the dominant oscillations between observations and background fields can be removed if the oscillation descriptions from the background are improved. Therefore, $q^{OBS} - q^{BF}$ and $q_{EOF_{reb}}^{OBS} - q_{EOF_{reb}}^{BF} (q_{FFT_{reb}}^{OBS} - q_{FFT_{reb}}^{BF})$ cannot represent the quality of observations. By using EOF and FFT, the diurnal and semidiurnal oscillations from observations, background fields, and large-scale weather characteristics are removed, and the small-scale features caused by observation errors will be left in the residual parts. Therefore, it is assumed that the residual parts should be a Gaussian-distribution, and such a bi-weighted QC is applied to the residual parts to identify instrument noise or environmental errors in the observations.

Taking EOF QC as an example, the QC process can be described in the following steps:

1) Use the EOF analysis to obtain the residual parts of the observations $q_{EOF_{res}}^{OBS}$ and background field $q_{EOF_{res}}^{BF}$.

2) Calculate the weight function in the kth station at the nth time $w_{k,n}$,

$$w_{k,n} = \frac{(q_{EOF_{res}}^{OBS} - q_{EOF_{res}}^{BF})_{k,n} - M}{c \times \mathrm{MAD}}, \tag{5}$$

where M is the median, $M = \mathrm{median}(q_{EOF_{res}}^{OBS} - q_{EOF_{res}}^{BF})$ and MAD is the median absolute deviation, $\mathrm{MAD} = \mathrm{median}\left[\left|(q_{EOF_{res}}^{OBS} - q_{EOF_{res}}^{BF}) - \mathrm{median}(q_{EOF_{res}}^{OBS} - q_{EOF_{res}}^{BF})\right|\right]$; $c = 7.5$; If $|w_{k,n}| > 1.0$, then $w_{k,n} = 1.0$.

The weight for each point is intended to keep data with large deviations from the average value from affecting the calculation of the integral mean and standard deviation.

3) Write the bi-weighted mean ($\overline{q_{EOF_{res}}^{OBS} - q_{EOF_{res}}^{BF}}^{\mathrm{bw}}$) and the bi-weighted standard deviation (STD_{bw}) as follows:

$$\overline{q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}}^{\mathrm{bw}} = M + \frac{\sum_{k,n}\left((q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}})_{k,n} - M\right)(1-w^2_{k,n})^2}{\sum_{k,n}(1-w^2_{k,n})^2}, \quad (6)$$

$$STD_{bw} = \frac{\left[N\sum_{i=1}\left((q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}})_{k,n} - M^2(1-w^2_{k,n})^4\right)\right]^{0.5}}{\left|\sum_{k,n=1}(1-w^2_{k,n})(1-5w^2_{k,n})\right|}, \quad (7)$$

4) Write the definition of outliers as follows:

$$Z_{k,n} = \frac{\left|(q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}})_{k,n} - \left(\overline{q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}}^{\mathrm{bw}}\right)\right|}{STD_{bw}} > Z, \quad (8)$$

here $Z_{k,n}$ is called Z-score of the specific humidity difference at the data point (k, n). Z is a specified threshold.

The flow of FFT QC is the same as EOF QC, only $q^{OBS}_{EOF_{res}}, q^{BF}_{EOF_{res}}$ are replaced by $q^{OBS}_{FFT_{res}}, q^{BF}_{FFT_{res}}$ in Eqs. (5)–(8).

3 Characteristics of specific humidity from observations and background fields

Figure 2 shows cumulative variances explained by the first to fifteenth EOF modes from observations (shaded areas) and background fields (dashed line) within an 8-day period ending at different observing times indicated by the numbers on the x-axis. It can be observed that the first EOF mode has explained 96.0% of the total variance, while the cumulative variance of the first ten EOF modes has reached 99.5% throughout January 2008. Therefore, the first ten modes ($c^{l_n}_{m,k}, m = 1,2,\cdots,10$) are chosen as the rebuild fields of EOF for the following analysis. In addition to the EOF modes, PCs ($e^{l_n}_{n,m} = (e^{l_n}_{1,m}, e^{l_n}_{2,m}, \ldots, e^{l_n}_{64,m})^{\mathrm{T}}$, $m = 1,2,\ldots,64$) show the time variation characteristics of those EOF modes. The PCs of the first to tenth EOF modes from the observations and background fields are presented in Fig. 3. The diurnal cycle can clearly be observed obviously in the 1st, 4th, 5th and 6th modes in both the observations and background fields. The semidiurnal oscillations can be found from the 8th to 10th EOF modes in the observations instead of the background fields, where the low frequency waves are dominant in those modes. Moreover, there exists an obvious phase and amplitude difference between the observations and background fields for the first ten PCs.

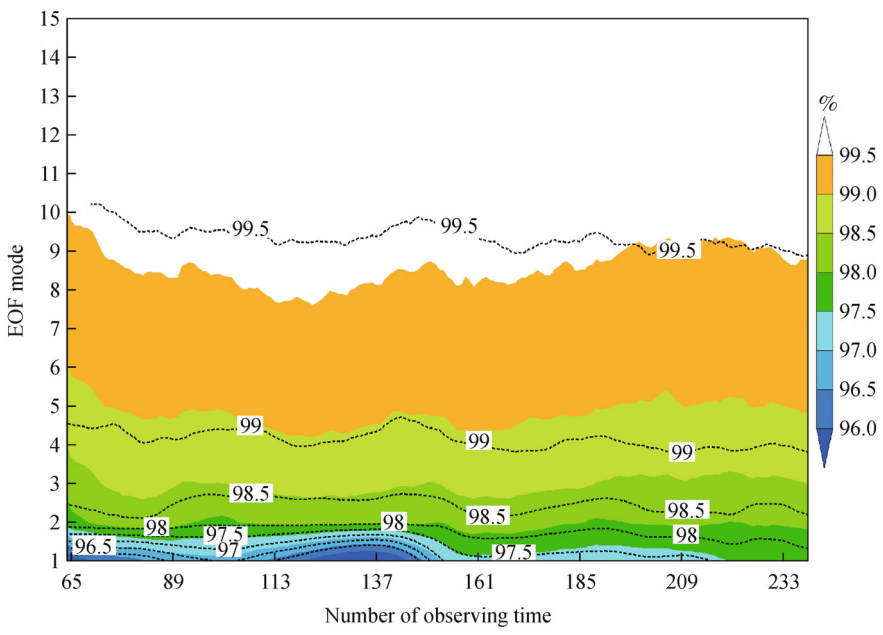

Fig. 2 Cumulative variances explained by the first 15 EOF modes for specific humidity from observations (shaded) and background fields (dashed line) within an 8-day period ending at different observing times indicated by the numbers on the x-axis. The time of 0000 UTC 9 January 2008 is the 65th observing time in January 2008.

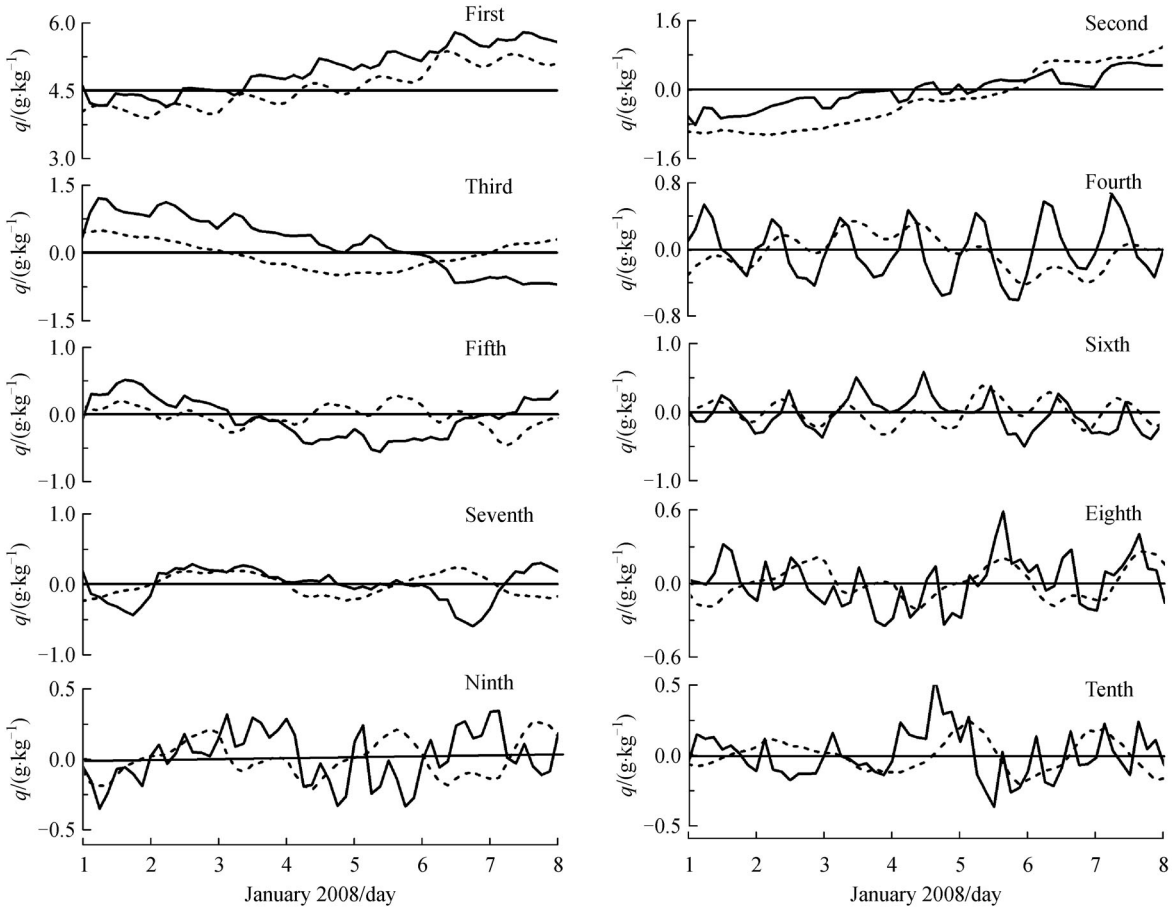

Fig. 3 PCs of the first ten EOF modes of specific humidity from observations (solid line) and background fields (dashed line) from 0000 UTC 1 January to 0000 UTC 8 January 2008. The number indicates the EOF mode number.

To obtain a clear understanding of the main oscillation characteristics of specific humidity in EOF modes, wavelet analysis (Torrence and Compo, 1998) and Fourier analysis are employed to analyze the top ten EOF modes. Figure 4 presents the wavelet modulus varying with the periods of oscillations. For the sake of finding the characteristics of diurnal and semidiurnal oscillations in the EOF modes, the period is divided into two phases: 6−18 h and 18−72 h. It is found that the diurnal cycle characteristics of the surface observations are more significant in the 4th, 6th and 8th EOF modes. However, for the background fields, in addition to the modes mentioned above, the 5th mode also shows significant diurnal oscillations. The long wave oscillations are also remarkable but only exist in the last few modes. The conditions of the semidiurnal cycles are more complicated than those of the diurnal cycle. First, semidiurnal oscillation is the most important feature for the 2nd, 6th and 8th−10th modes in the observations and the 5th and 8th modes in the background fields. Second, it should be noted that the modulus of wavelet analyses from the background fields is smaller than the observations, which means that the semidiurnal oscillation in the background fields is weaker than that in the observations. Different from spectral analysis, the 8-h oscillation is not significant in the EOF modes. Furthermore, the first ten EOF modes are decomposed by FFT, and the results distinctively show that the more significant the cycles of the modes, the larger their amplitudes (figure omitted).

Skewness (Mardia, 1970) is a measure of the asymmetry of the probability distribution of a real-valued random variable. A skewness of zero indicates that the variable values are relatively evenly distributed on both sides of the mean, typically implying a symmetric distribution. Kurtosis (Mardia, 1970) is a measure of the "peakedness" of the probability distribution of a real-valued random variable. Distributions with less than 3 or greater than 3 are called platykurtic or leptokurtic respectively. As the truth is unknown, the observation error and background error cannot be determined. However, the OMB vector can replace them to perform analysis, because the difference of two variables with Gaussian distribution is Gaussian. As both the EOF and FFT rebuild fields can reflect the main

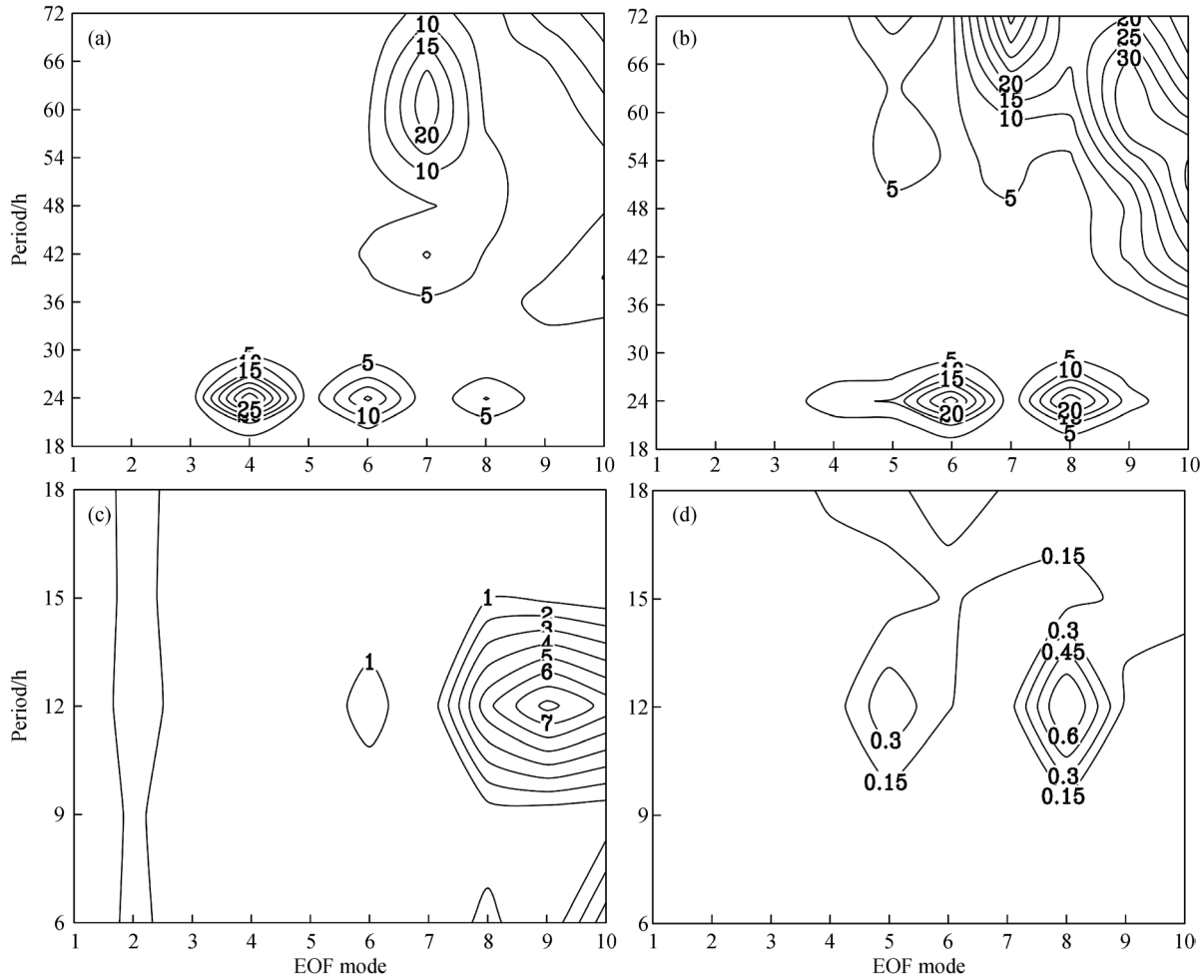

Fig. 4 Modulus of wavelet analysis for the PCs of the first ten EOF modes of specific humidity from observations (a, c) and background fields (b, d) for oscillation periods with 18−72 h (top panels a and b) and 6−18 h (bottom panels c and d).

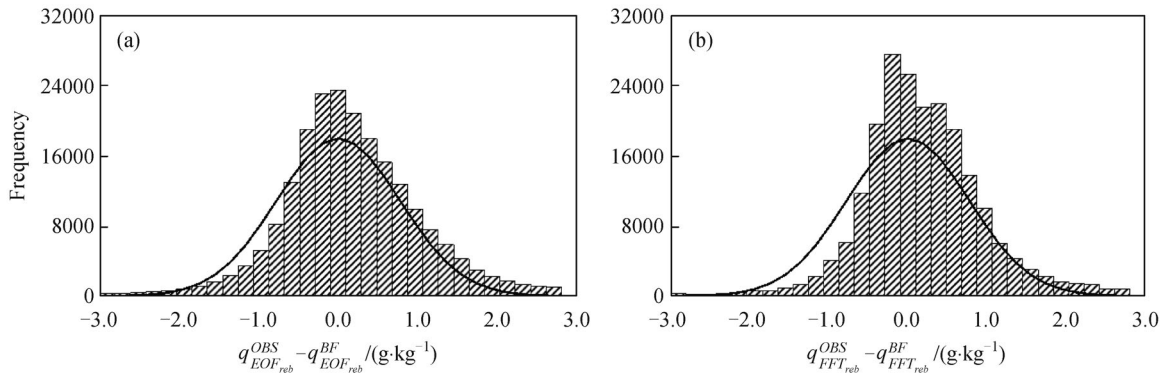

Fig. 5 Frequency distributions of the differences between the observations and background fields for the EOF rebuild terms $q^{OBS}_{EOF_{reb}} - q^{BF}_{EOF_{reb}}$ (a), and the composite fields extracted from FFT $q^{OBS}_{FFT_{reb}} - q^{BF}_{FFT_{reb}}$ (b). The solid curve is the Gaussian fitting.

oscillations, Fig. 5 presents the OMB distributions of the rebuild fields from EOF and FFT. It shows that the two distributions are different from a Gaussian distribution with the characteristics of nonzero mean, right-skewed and leptokurtic distribution. Similar to temperature, the inaccurate descriptions of the intensity and phase differ-

ences in diurnal, semidiurnal and 8-h oscillations contribute to the non-Gaussian distributions of OMB. These differences mainly exist in the first ten EOF modes.

4 QC Results

The standard deviation spatial distributions of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ are shown in Fig. 6. The data calculated from all available stations are selected from the first eight days in January 2008. The conditions of the other days are similar. We can clearly find that the *STD* (standard deviation) of $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ is relatively larger than that of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$, and the *STD* of the difference between observations and background fields in the India Peninsula, Indo-China Peninsula and the islands in south-east is much larger than that in the north. Therefore, the whole domain can be separated into two parts: a south region (S) and a north region (N). The solid line in Fig. 6 indicates the boundary between these two regions. There are 836 stations in the south region and 2504 stations in the north region. The STD_{bw} (bi-weighted standard deviation) time series of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ in the two regions can be observed in Fig. 7. It is shown that the STD_{bw} remains nearly constant in January 2008, so it is reasonable to choose a constant Z-score to perform QC during the whole month. Moreover, through comparing the STD_{bw} determined by EOF and FFT in the same region, the STD_{bw} of EOF is approximately 0.15 g/kg smaller than FFT in the north region and 0.2 g/kg smaller in the south region.

The STD_{bw} in the south region is greater than that in the north, due to the nature of the variability of the data. Observations will be identified as the outliers when $|q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}|$ exceed the given value-Z. By testing many different thresholds of *Z*, the thresholds of Z are empirically specified, for EOF QC, 3.0 and 3.5 are used for the north region and the south region, respectively for FFT, they are 2.8 and 3.3, respectively.

Because the impact of QC on the frequency distributions of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ occurred mostly on the tails of the distributions, Figure 8 gives the frequency distributions before and after QC on the left and right tails. It seems that the stations with large OMB can be eliminated effectively after QC, and the OMB error ranges are consequently reduced when the stations with large OMB are eliminated by QC. Additionally, the EOF QC method can remove more outliers. The distributions of frequencies after QC are much more similar to a Gaussian distribution.

To illustrate more objectively the characteristics of the frequency statistics before and after QC, Table 1 gives the mean, standard deviation, skewness and kurtosis of the OMB of variables q, $q_{EOF_{1-10}}$, $q_{FFT_{0,8,16}}$, $q_{EOF_{res}}$ and $q_{FFT_{res}}$ before and after QC. It is noted that after EOF or FFT QC, the mean, standard deviation, skewness and kurtosis of the distribution of $q^{OBS} - q^{BF}$ shows slight improvement. This is due to the non-Gaussian characteristics still being present in the $q^{OBS} - q^{BF}$ after QC, and the non-Gaussian characteristics are mainly from the rebuild terms of EOF ($q^{OBS}_{EOF_{reb}} - q^{BF}_{EOF_{reb}}$) or FFT ($q^{OBS}_{FFT_{reb}} - q^{BF}_{FFT_{reb}}$), which are contained in inaccurate descriptions of the diurnal, semidiurnal and 8-h oscillations. In addition, when the Gaussian and non-Gaussian distributions are both superposed together in the observation, the OMB distribution is dependent on the relative strength of two signals (Zou, 2009). The QC process for residual parts can greatly

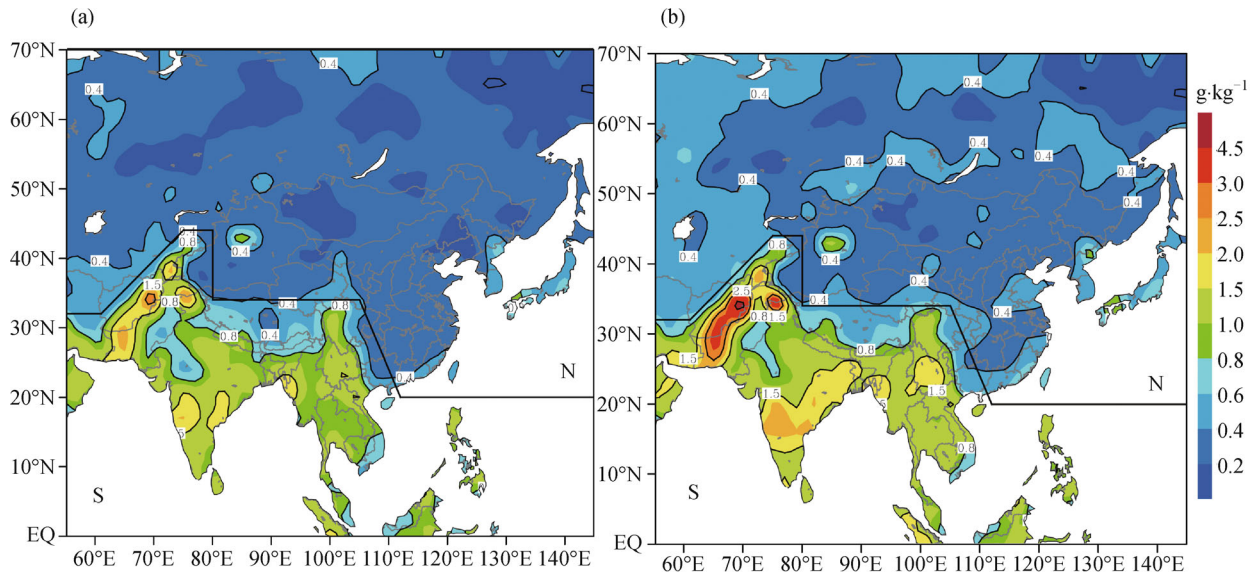

Fig. 6 The spatial distributions of standard deviations for the EOF residual terms (a) and FFT residual terms (b) of specific humidity.

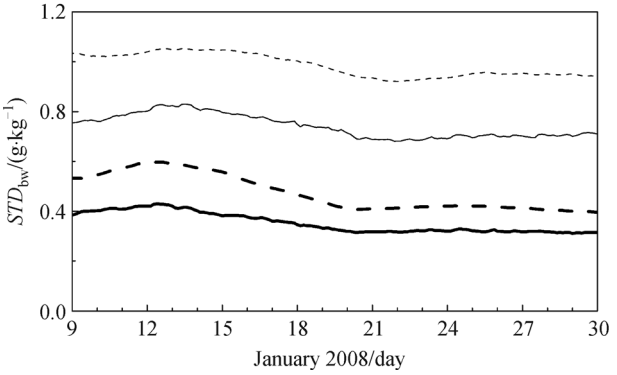

Fig. 7 The bi-weighted standard deviations of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ (solid lines) and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ (dashed lines) in the north region (thick lines) and the south region (thin lines) (See Fig. 6 for the two regions).

improve the frequency distributions of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$. The mean and the standard deviation are close to 0, and the skewness and kurtosis are close to 0 and 3, respectively. This means the frequency distributions after QC are more Gaussian.

Figures 9(a) and 9(b) present the spatial distribution of the total number of station data points removed by the EOF QC and FFT QC for specific humidity. The specific humidity outliers which occur more than 20 times are mainly present at the periphery of the Qinghai-Tibet plateau, around Gansu and Sichuan provinces and the Yunnan-Guizhou Plateau. These areas are mountainous and sparsely populated with fewer data sources, causing a large deviation between the observations and the background field. Other regions with more outliers are found in the coastal areas such as southern China, the Korean Peninsula and Japan. The data in those regions contain random and environmental errors due to the influence of the water vapor from the ocean. The results of the EOF QC and FFT QC appear similar, but the FFT QC identifies more outliers in southern China. Figure 9(c) shows the outliers identified by both QCs. Outliers with more than 50 occurrences are mainly present in Gansu and the coastal areas in southern China and Japan. In addition, the EOF QC distinguishes 4.29% and 3.13% of the outliers in the

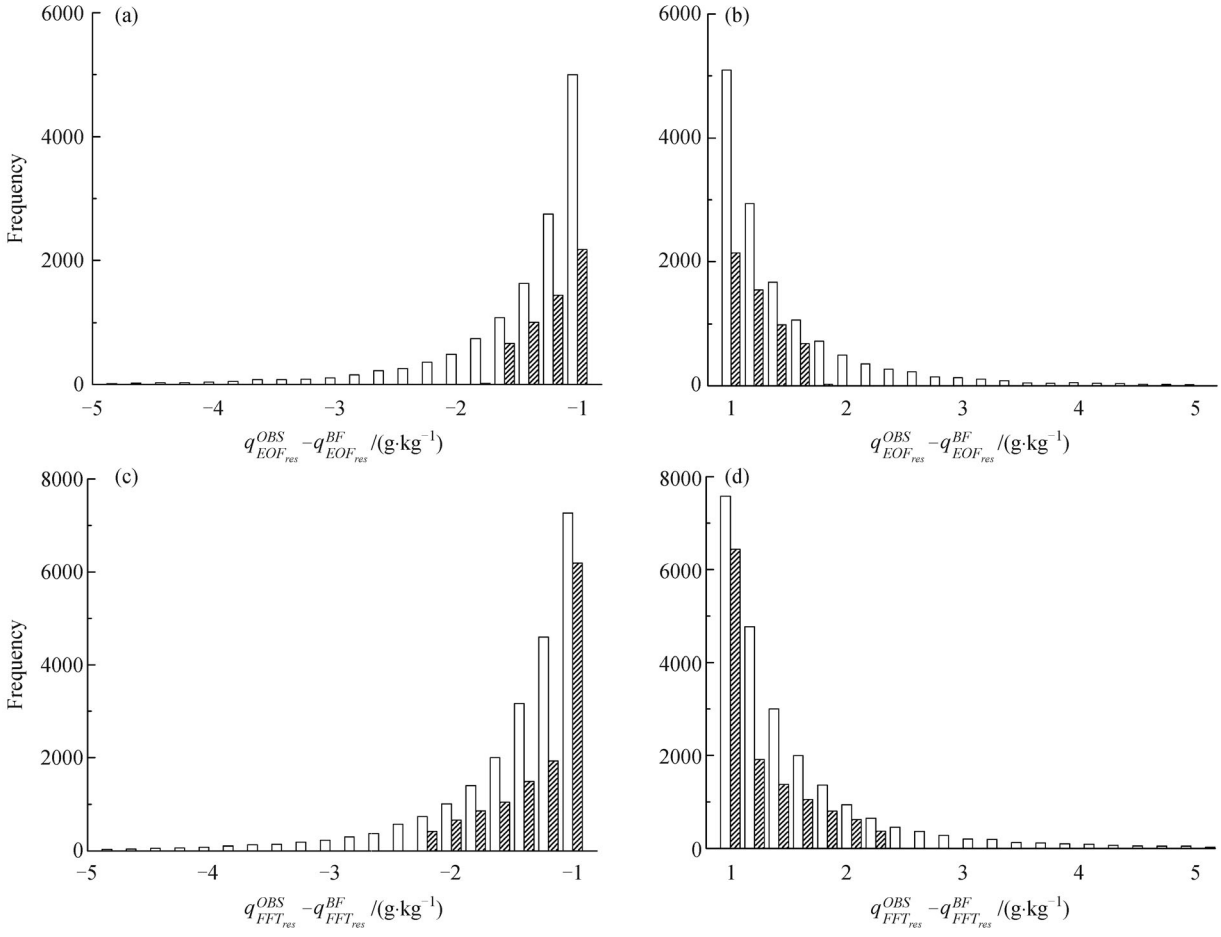

Fig. 8 Frequency distributions of the differences of $q^{OBS}_{EOF_{res}} - q^{BF}_{EOF_{res}}$ (a, b) and $q^{OBS}_{FFT_{res}} - q^{BF}_{FFT_{res}}$ (c, d) before (white bar) and after (shaded bar) QC. For clarity, only the tails of the frequency distributions are shown.

Table 1 Mean, Standard Deviation, Skewness, and Kurtosis of OMB of variables $q^{OBS}-q^{BF}$, $q^{OBS}_{EOF_{reb}}-q^{BF}_{EOF_{reb}}$, $q^{OBS}_{FFT_{reb}}-q^{BF}_{FFT_{reb}}$, $q^{OBS}_{EOF_{res}}-q^{BF}_{EOF_{res}}$, $q^{OBS}_{FFT_{res}}-q^{BF}_{FFT_{res}}$, before and after QC.

	Mean/(g·kg^{-1})	STD/(g·kg^{-1})	Skewness	Kurtosis
$q^{OBS}-q^{BF}$	0.1086	0.8375	0.1834	4.3971
$q^{OBS}-q^{BF}$ (after EOF QC)	0.1197	0.7968	0.2633	4.2744
$q^{OBS}-q^{BF}$ (after FFT QC)	0.1273	0.7750	0.2414	4.6269
$q^{OBS}_{EOF_{reb}}-q^{BF}_{EOF_{reb}}$ (before/after EOF QC)	0.1214 0.1274	0.7968 0.7637	0.2183 0.1959	4.3292 4.1707
$q^{OBS}_{FFT_{reb}}-q^{BF}_{FFT_{reb}}$ (before/after FFT QC)	0.1661 0.1660	0.7151 0.6808	0.4204 0.3180	5.6551 5.2848
$q^{OBS}_{EOF_{res}}-q^{BF}_{EOF_{res}}$ (before/after EOF QC)	−0.0065 −0.0043	0.5674 0.4104	−0.0178 −0.0049	7.0047 3.7451
$q^{OBS}_{FFT_{res}}-q^{BF}_{FFT_{res}}$ (before/after FFT QC)	−0.0457 −0.0294	0.6858 0.5366	−0.1161 −0.0208	5.6329 3.6687

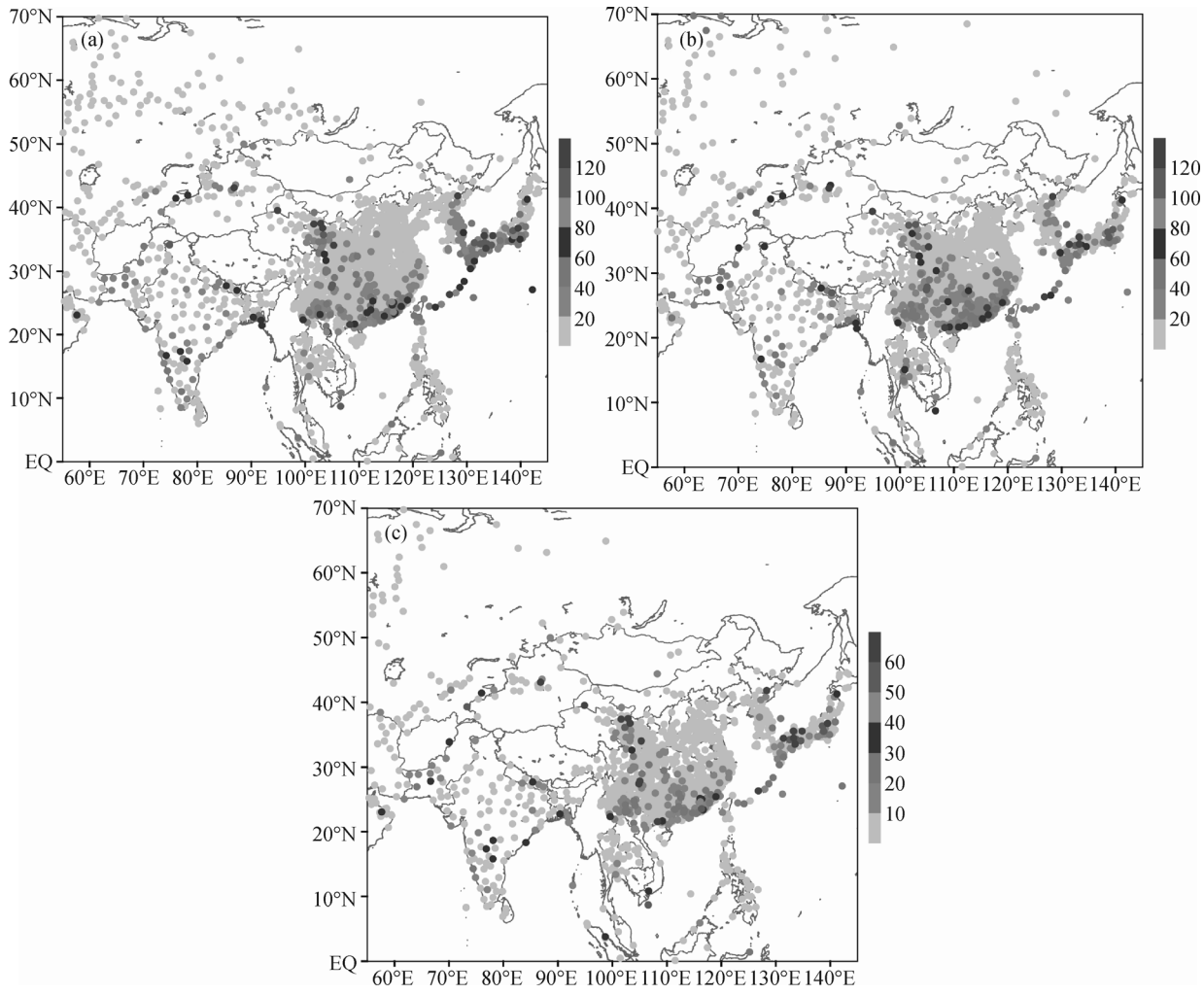

Fig. 9 Total number of specific humidity data removed at each surface station by (a) EOF QC, (b) FFT QC and (c) both QCs.

south and north regions, respectively while the percentages of data removed by FFT QC are 4.51% and 3.08%, respectively. Furthermore, outliers identified by both QC methods hold up to 40.71% and 43.66% for the total outliers of EOF QC and FFT QC, respectively.

Figure 10 shows that the percentage of data removed

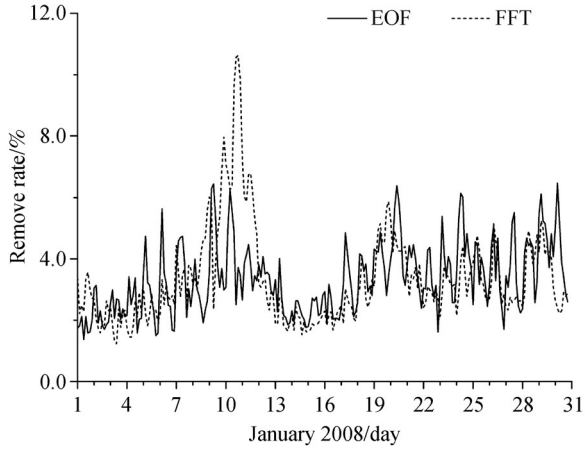

Fig. 10 The percentage of surface specific humidity removed at each observing time by QC methods using the EOF and FFT methods.

varied with time for specific humidity during January 2008. The outliers' percentage of EOF QC and FFT QC is nearly the same during the whole month, except for the period from 10th January to 13th January of 2008. We have found that a large number of FFT outliers are clustered into one area at 1500 UTC on 10 January 2008 over northern and central China, as shown in the subdomains in Fig. 11.

To see the example clearly, the subdomains in Fig. 11 are amplified, and Fig. 12 presents the spatial distributions of q, $q_{FFT_{reb}}$,$q_{EOF_{reb}}$,$q_{FFT_{res}}$ and $q_{EOF_{res}}$ from the observations, the background and the differences between them. It is noted that the specific humidity is presented in the form of two ridges with a deep trough, but the locations of the two ridges and the deep trough in the background fields are

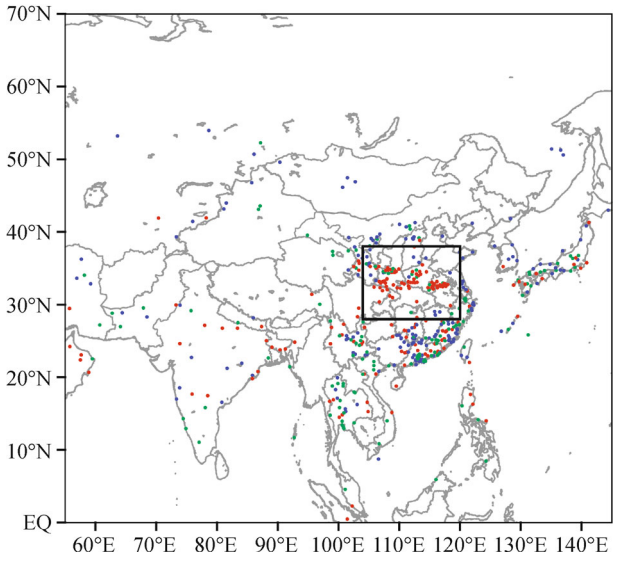

Fig. 11 Spatial distributions of the outliers identified by only the EOF QC (blue points), only the FFT QC (red points), and by both methods (green points) at 1500 UTC 10 January 2008.

ahead of that in the observations. Thus, the difference between observations and background fields shows a structure similar to a saddle-type field (Fig. 12(a)). The FFT rebuild fields can depict the distributions of water vapor troughs and ridges but fail to capture their intensity. The differences between observations and background fields show a large positive error over the research area (Fig. 12(b)) because the rebuild field only catches the mean and main oscillation components from the original fields. Furthermore, the FFT QC method is carried out at individual stations, so the wavelet information with other periods will be kept in the residual parts and removed as outliers during the QC process (Fig. 12(d)). Compared with FFT, the first few EOF modes can capture the spatial structures for both observations and background fields (Fig. 12(c)); the residual fields looks more homogeneous, with a small OMB value (Fig. 12(e)).

5 Summary and discussion

A local surface station observation network has been set up in many areas in China, and considerable attention has been paid to the application of surface data (Baker, 1992; Guo et al., 2002). Specifically, data quality is one of the roadblocks to surface data application, especially data assimilation (Fan and Zhang, 2006). The poor quality of the data results in the low utilization rate in NWP. Systematic quality control procedures should be applied to the station data to ensure they are representative and accurate to capture weather phenomena in the boundary layer so that effective information can be extracted from the "good" observations and a better initial background field can be discerned for numerical simulation.

Comparisons between observations and background fields show that amplitude and phase differences between observations and background fields result in a non-Gaussian distribution in OMB, which violates the assumption of normalized distribution observation error in data assimilation systems (Sasaki, 1970). For this reason, a bi-weighted standard deviation QC method (Zou and Zeng, 2006) based on EOF and FFT is proposed. The chief features of the data, such as their distinctive oscillations and information related to abnormal weather events, are retained in the rebuild terms. The QC for residual terms can avoid the influence of systematic errors in the background field. The standard deviation between observations and the background fields is reduced after QC, and the frequency distribution for OMB is closer to the normal distribution. The specific outliers occur mainly in mountainous and coastal regions, which is mainly due to the location of the observation sites. The data are sparse in the mountainous areas, causing large deviations between observations and the background fields. For the coastal areas, the specific humidity is vulnerable to the water vapor from the ocean, which may produce environmental errors.

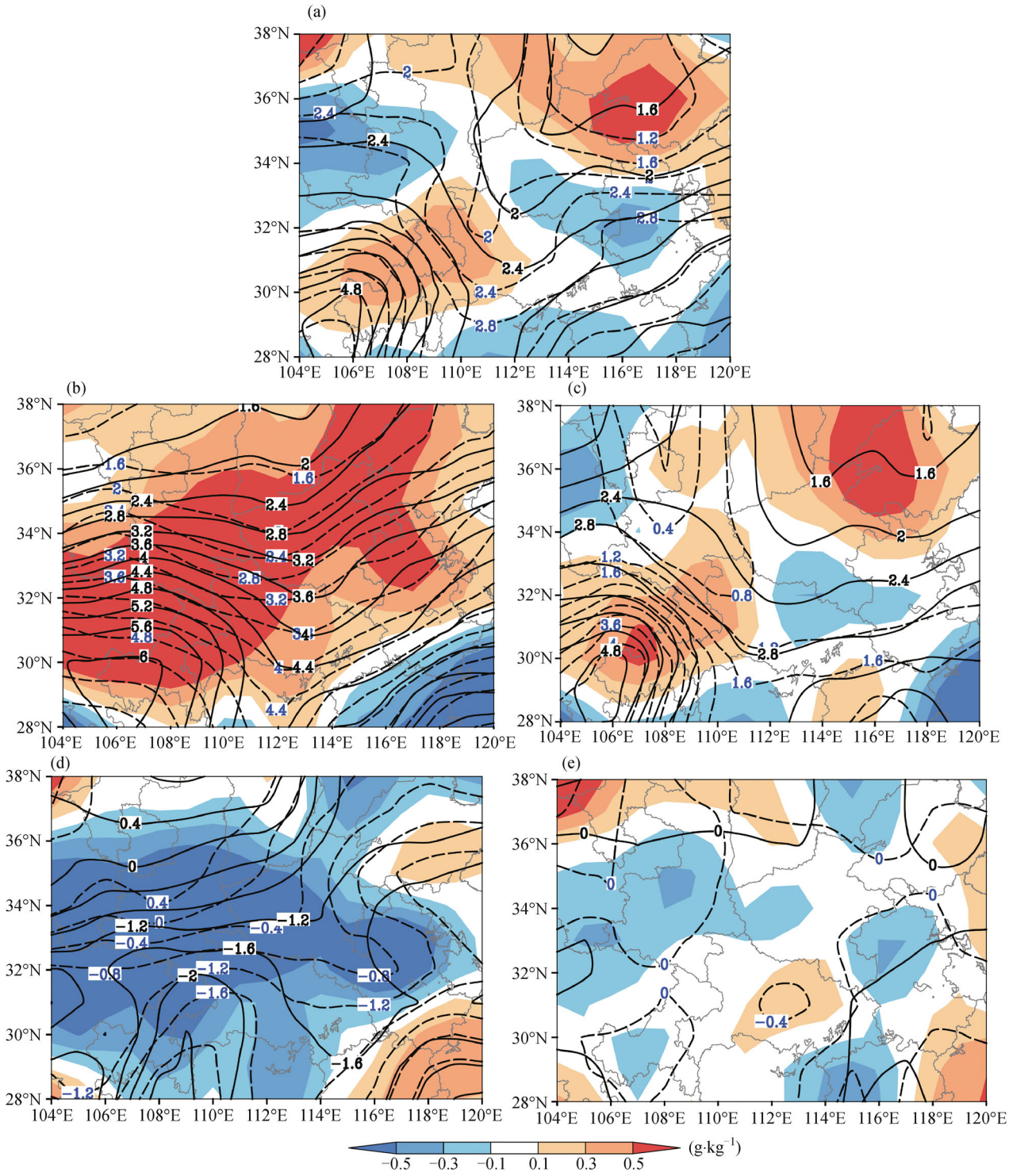

Fig. 12 Spatial distributions of (a) q, (b) $q_{FFT_{reb}}$, (c) $q_{EOF_{reb}}$, (d) $q_{FFT_{res}}$ and (e) $q_{EOF_{res}}$ from observations (solid lines), FNL data (dashed lines), and the differences between observations and FNL data (shaded areas: $q^{OBS}-q^{BF}$ in (a), $q^{OBS}_{FFT_{reb}}-q^{BF}_{FFT_{reb}}$ in (b), $q^{OBS}_{EOF_{reb}}-q^{BF}_{EOF_{reb}}$ in (c), $q^{OBS}_{FFT_{res}}-q^{BF}_{FFT_{res}}$ in (d), and $q^{OBS}_{EOF_{res}}-q^{BF}_{EOF_{res}}$ in (e)) at 1500 UTC 10 January 2008 in the subdomain indicated by the small box in Fig. 11.

The results of the two QC methods show a slight difference. EOF QC can keep large-scale fluctuation information from the original field, avoiding these waves entering into the residual field and being removed by the QC process, so it performs better than FFT QC.

This paper presents an effective QC method for data assimilation regarding surface specific humidity. The next work will focus on the flowing aspects: (1) comparing

other surface variables (e.g., surface pressure and horizontal winds) with other reanalysis data such as ERA-Interim (Simmons et al., 2007), JRA-25 (Onogi et al., 2007) and T639 (Chen et al., 2007); (2) applying this QC method to other surface variables and verifying the QC results; and (3) evaluating the impact of surface data passing QC on meso- and micro-scale weather system prediction.

Acknowledgements This work was jointly supported by the National Basic Research Program of China (No. 2010CB951600), China Special Fund for Meteorological Research in the Public Interest (No. GYHY201406008), the Research Innovation Program for college graduates of Jiangsu Province (CXZZ13_0503) and the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions. Thanks for NCEP for providing FNL data on http://rda.ucar.edu/.

References

Anderson E, Järvinen H (1999). Variational quality control. Q J R Meteorol Soc, 125(554): 697–722

Baker N L (1992). Quality control for the Navy operational atmospheric database. Weather Forecast, 7(2): 250–261

Chen Q Y, Guan C G, Yao M M, Tong H, Mariano H, Li Z C (2007). Development of key techniques and experiments in global model upgrading. Acta Meteorologica Sinica, 65(4): 478–492 (in Chinese)

Cooley J W, Tukey O W (1965). An algorithm for the machine calculation of complex Fourier series. Math Comput, 19(90): 297–301

Eischeid J K, Bruce Baker C, Karl T R, Diaz H F (1995). The quality control of long-term climatological data using objective data analysis. J Appl Meteorol, 34(12): 2787–2795

Fan S Y, Zhang C L (2006). Numerical assessing experiments on individual component impact of the meteorological observation network on the "July 2000" torrential rain in Beijing. Acta Meteorologica Sinica, 20(4): 389–401 (in Chinese)

Feng S, Hu Q, Qian W (2004). Quality control of daily meteorological data in China, 1951–2000: a new dataset. Int J Climatol, 24(7): 853–870

Gandin L S (1988). Complex quality control of meteorological observations. Mon Weather Rev, 116(5): 1137–1156

Ghil M, Malanotte-Rizzoli P (1991). Data assimilation in meteorology and oceanography. Adv Geophys, 33: 151−156

Guo Y R, Shin D H, Lee J H, Xiao Q N, Barker D M, Kuo Y H (2002). Application of the MM5 3DVAR system for a heavy rain case over the Korean Peninsula. In: Proccedings: the Twelfth PSU/NCAR Mesoscale Model Users' Workshop NCAR, Boulder, 24–25

Ha K J, Jeon E H, Oh H M (2007). Spatial and temporal characteristics of precipitation using an extensive network of ground gauge in the Korean Peninsula. Atmos Res, 86(3–4): 330–339

Hong S Y, Pan H L (1996). Nonlocal boundary layer vertical diffusion in a medium-range forecast model. Mon Weather Rev, 124(10): 2322–2339

Hubbard K G, You J (2005). Sensitivity analysis of quality assurance using the spatial regression approach—A case study of the maximum/minimum air temperature. J Atmos Ocean Technol, 22 (10): 1520–1530

Ingleby N B, Lorenc A C (1993). Bayesian quality control using multivariate normal distributions. Q J R Meteorol Soc, 119(513): 1195–1225

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woolen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo K C, Ropelewski C, Wang J, Jenne R, Joseph D (1996). The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc, 77(3): 437–471

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. Int J Climatol, 16 (11): 1197–1226

Li L, Njoku E G, Im E, Chang P S, Germain K S (2004). A preliminary survey of radio-frequency interference over the US in Aqua AMSR-E data. Geoscience and Remote Sensing. IEEE Transactions on, 42(2): 380–390

Lorenc A C, Hammon O (1988). Objective quality control of observations using Bayesian methods. Theory, and a practical implementation. Q J R Meteorol Soc, 114(480): 515–543

Lorenz E N (1956). Empirical orthogonal functions and statistical weather prediction, scientific Report No. 1, Statistical Forecasting Project. Massachusetts Institute of Technology department of meteorology

Mardia K V (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3): 519–530

Mohanty U C, Kasahara A, Errico R (1986). The impact of diabatic heating on the initialization of a global forecast model. J Meteorol Soc Jpn, 64(6): 805–817

Onogi K, Tsutsui J, Koide H, Sakamoto M, Kobayashi S, Hatsushika H, Matsumoto T, Yamazaki N, Kamahori H, Takahashi K, Kadokura S, Wada K, Kato K, Oyama R, Ose T, Mannoji N, Taira R (2007). The JRA-25 reanalysis. J Meteorol Soc Jpn, 85(3): 369–432

Qin Z K, Zou X L, Li G, Ma X L (2010). Quality control of surface station temperature data with non-Gaussian observation-minus-background distributions. J Geophys Res, 115(D16): D16312

Ramanathan V, Barkstrom B R, Harrison E F (1989). Climate and the earth's radiation budget. Phys Today, 42(5): 22–32

Reek T, Doty S R, Owen T W (1992). A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. Bull Am Meteorol Soc, 73(6): 753–762

Sasaki Y (1970). Some basic formalisms in numerical variational analysis. Mon Weather Rev, 98(12): 875–883

Shafer M A, Fiebrich C A, Arndt D S, , Fredrickson S E, Hughes T W (2000). Quality assurance procedures in the Oklahoma Mesonetwork. J Atmos Ocean Technol, 17(4): 474–494

Simmons A, Uppala S, Dee D, Kobayashi S (2007). ERA-Interim: new ECMWF reanalysis products from 1989 onwards. ECMWF newsletter, 110(110): 25–35

Torrence C, Compo G P (1998). A practical guide to wavelet analysis. Bull Am Meteorol Soc, 79(1): 61–78

Wade C G (1987). A quality control program for surface mesometeorological data. J Atmos Ocean Technol, 4(3): 435–453

Wang Y, Xu Z F, Fan G Z (2013). Research of quality control based on EOF for 2m temperature. Plateau Meteorology, 32(2): 564–574 (in Chinese)

Wei F Y (2007). Statisticial Diagnosis and Prediction for Modern Climate. Beijing: China Meteorological Press, 71–76 (in Chinese)

Zhao J, Zou X L, Weng F Z (2013). WindSat radio-frequency interference signature and its identification over Greenland and Antarctic. IEEE Trans Geosci Rem Sens, 51(9): 4830–4839

Zou X L (2009). Theories and Applications for Data Assimilation (I). Beijing: China Meteorological Press (in Chinese)

Zou X L, Ma Y, Qin Z K (2012). Fengyun-3B microwave humidity sounder (MWHS) data noise characterization and filtering using principle component analysis. IEEE Trans Geosci Rem Sens, 50(12): 4892–4902

Zou X L, Qin Z K (2010). Time zone dependence of diurnal cycle errors in surface temperature analyses. Mon Weather Rev, 138(6): 2469–2475

Zou X L, Zeng Z (2006). A quality control procedure for GPS radio occultation data. J Geophys Res, 111: D02112