Environmental Pollution 289 (2021) 117846

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Source apportionment of atmospheric particle number concentrations with wide size range by nonnegative matrix factorization (NMF)☆

Chun-Sheng Liang [a,c,1], Dingli Yue [b,1], Hao Wu [f], Jin-Sen Shi [a,c], Ke-Bin He [d,e,*]

[a] Collaborative Innovation Center for West Ecological Safety, Lanzhou University, Lanzhou, 730000, China
[b] Guangdong Environmental Monitoring Center, State Environmental Protection Key Laboratory of Regional Air Quality Monitoring, Guangzhou, 510308, China
[c] Key Laboratory for Semi-Arid Climate Change of the Ministry of Education, College of Atmospheric Sciences, Lanzhou University, Lanzhou, 730000, China
[d] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, 100084, China
[e] State Environmental Protection Key Laboratory of Sources and Control of Air Pollution Complex, Beijing, 100084, China
[f] Key Laboratory of China Meteorological Administration Atmospheric Sounding, School of Electrical Engineering, Chengdu University of Information Technology, Chengdu, 610225, China

ARTICLE INFO

Keywords:
Source apportionment
Particle number concentrations
Nonnegative matrix factorization
Particle number size distributions
Receptor model
Particle sizer

ABSTRACT

Quantifying the sources of atmospheric particles is essential to air quality control but remains challenging, especially for the source apportionment of particles based on number concentration with wide size range. Here, particle number concentrations (PNC) with size range 19–20,000 nm involving four modes Nucleation, Aitken, Accumulation, and Coarse are used to do source apportionment of PNC at the Guangdong Atmospheric Supersite (Heshan) during July–October 2015 by nonnegative matrix factorization (NMF) with 6 factors. For July 2015, separated source apportionments for three different size ranges from collocated instruments nano scanning mobility particle sizer (NSMPS), SMPS, and aerodynamic particle sizer (APS) and for two different size ranges (below and above 100 nm) show similar quantitative source information with that for the one whole size range. The mean absolute difference of contribution percentages of total particle number concentrations (TPNC) based on 5 unique apportioned sources is 5.6 % (4.3–7.6 %) for the instrument segregated apportionment and 4.2 % (0–5.3 %) for the size range segregated apportionment respectively, relative to the one whole apportionment. Moreover, the contribution percentages of TPNC are close to the weighted sum of contribution percentages of all size bins, with a mean absolute difference of 1.1 % (0–3.4 %). In both these two aspects, the consistency among different technical paths proves the matrix factorization by NMF is practically desirable and the simplicity of reducing some steps or calculations saves time. Besides, dust can be identified with the wide size range including larger than 3000 nm. Six apportioned sources in the 4 months are Accumulation (32.4 %), Nucleation (20.0 %), Aitken (15.2 %), traffic (14.6 %), dust (10.6 %), and Coarse (7.1 %). Therefore, NMF would serve as a promising tool for PNC source apportionment with wide size range and conducting the apportionment with the whole size range in one matrix factorization procedure and using the single TPNC contribution percentage are feasible.

1. Introduction

Source apportionment of atmospheric particles is very important to improve air quality for public health (Almeida et al., 2020; Hopke et al., 2020; Zhang et al., 2020), but particle source apportionment based on number concentration is immature. Although particle source apportionment based on number concentration is complementary to and verifiable with that based on mass concentration (Cai et al., 2020; Liang et al., 2020; Rodins et al., 2020), it has been conducted much less than that based on mass concentration. For example, there are 53 results for titles 'particle number' and 'source' and 1039 results for titles '$PM_{2.5}$' and 'source' from Web of Science Core Collection on March 12, 2021.

Two kinds of combinations can be found in the studies on particle number concentrations (PNC) with wide size range and PNC source

☆ This paper has been recommended for acceptance by Pavlos Kassomenos.

* Corresponding author. State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, 100084, China.
E-mail address: hekb@tsinghua.edu.cn (K.-B. He).

[1] These authors contributed equally.

https://doi.org/10.1016/j.envpol.2021.117846
Received 20 March 2021; Received in revised form 5 July 2021; Accepted 24 July 2021
Available online 24 July 2021

Table 1
R packages.

Roles	R packages (references)
Data transformation and statistics and character processing	magrittr (Bache and Wickham, 2014), readxl (Wickham and Bryan, 2019), dataprep (Liang et al., 2020; Liang et al., 2021), dplyr (Wickham et al., 2019b), reshape2 (Wickham, 2007; Wickham, 2017), openair (Carslaw and Ropkins, 2019; Carslaw and Ropkins, 2012), lubridate (Grolemond and Wickman, 2011; Spinu et al., 2018), forcats (Wickham, 2019a), scales (Wickham and Seidel, 2019), zoo (Zeileis and Grothendieck, 2005; Zeileis et al., 2019), RcppRoll (Ushey, 2018), tidyr (Wickham and Henry, 2019), expss (Demin, 2019), tibble (Müller and Wickham, 2019), broom (Robinson and Hayes, 2019), stringr (Wickham, 2019b)
Data collection	splitr (Iannone, 2021)
Models (NMF, CWT)	NMF (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2020), splitr (Iannone, 2021), openair (Carslaw and Ropkins, 2019; Carslaw and Ropkins, 2012)
Plotting	ggplot2 (Wickham, 2016; Wickham et al., 2019a), scales (Wickham and Seidel, 2019), cowplot (Wilke, 2019), lattice (Sarkar, 2008, 2018), openair (Carslaw and Ropkins, 2019; Carslaw and Ropkins, 2012), gtable (Wickham and Pedersen, 2019), ggrepel (Slowikowski, 2019), OpenStreetMap (Fellows and Stotz, 2019), ggplotify (Yu, 2019)

apportionment. One is about hardware (observations) and the other is about software (modelling). First, wider size ranges contain more abundant fingerprints about sources (Charron et al., 2008; Hussein et al., 2014), which generally need to be measured by combing multiple instruments. It is hard to measure wide size ranges such as 2–20,000 nm by using one single aerosol instrument (Kulkarni and Baron, 2011). Consequently, two or more instruments such as scanning mobility particle sizer (SMPS), aerodynamic particle sizer (APS), and optical particle spectrometer/sizer (OPS) have been jointly used in studies on particle number size distributions (PNSD) containing PNC with wide size ranges (Harrison et al., 2011; Masiol et al., 2016; Spielvogel et al., 2010; Wu and Boor, 2020; Xia et al., 2020). Second, the practical PNC source apportionment operations often involve combing receptor models and supplementary information such as related pollutants, meteorology, and dispersion models (Beddows et al., 2015; Harrison et al., 2011; Masiol et al., 2017; Vu et al., 2015). This combing has been summarized and proposed to adopt for PNC source apportionment (Liang et al., 2020).

In recent years, both instruments and receptor models have been developed and compared for observations and source apportionment of PNC. Portable and miniature ultrafine particle sizers (Liu et al., 2020; Sun et al., 2019; Yang et al., 2021), new nanoparticle sizer (Lee et al., 2020), and compact low-cost optical particle sizer (Njalsson and Novosselov, 2018) have been developed. The state-of-the-art aerosol technologies have been used and verified for high flow differential mobility particle sizer (Kangasluoma et al., 2018) and compared for field observations (Vo et al., 2018). Methods to assess the performance and uncertainties of SMPS instruments have also been explored (Coquelin et al., 2018; Stolzenburg and McMurry, 2018). With respect to receptor models, after comparing with *k*-means clustering (Beddows et al., 2009), principal components analysis (PCA) (Khan et al., 2015; Liang et al., 2013; Pey et al., 2009), factor analysis (FA) (Wåhlin et al., 2001), and positive matrix factorization (PMF) (Ogulei et al., 2006; Zhou et al., 2004), the nonnegative matrix factorization (NMF) (Lee and Seung, 1999) with synergetic advantages in factor distinction, nonnegative constraints, C++ optimized algorithms (Gaujoux and Seoighe, 2020), and parallel computing has been introduced and proposed to use for PNC source apportionment (Liang et al., 2020). Besides being aimed at PNC source apportionment, NMF has also been used to provide $PM_{2.5}$ and PM_{10} source profiles (Delmaire et al., 2010; Kfoury et al., 2016; Liu et al., 2019; Scerri et al., 2019; Shang et al., 2018; Suzuki et al., 2021; Zhang et al., 2019), deconvolve low-cost sensors data into CO-dominated and particle factors (Hagan et al., 2019), and identify sources of methylsiloxanes (Horii et al., 2021). Coupled with autocorrelation function (ACF), the diurnal, stochastic, and persistent sources can be distinguished by NMF (Ouaret et al., 2017). Informed weighted NMF methods using αβ-divergence would reduce the influence of outliers in practice (Delmaire et al., 2019).

Despite the above-mentioned advances, the performance of NMF in PNC source apportionment with wide size range is unknown. Some other unknown aspects include whether the source apportionment of PNC with wide size range observed by combing multiple instruments can be conducted in one matrix factorization procedure and whether using the single TPNC contribution percentage is feasible for reporting TPNC sources.

To make these unknowns known, PNC with wide size range 2–20,000 nm involving four modes Nucleation, Aitken, Accumulation, and Coarse are observed at the Guangdong Atmospheric Supersite (Heshan) during July–October 2015 by using instruments nano scanning mobility particle sizer (NSMPS), SMPS, and aerodynamic particle sizer (APS). The data are used in PNC source apportionment by NMF. Separated source apportionment (matrix factorization) procedures for three different size ranges from collocated instruments NSMPS, SMPS, and APS and for two different size ranges (below and above 100 nm) are compared with that for the whole size range in one matrix factorization procedure, using the data of July 2015. And, the contribution percentage of TPNC is compared with the weighted sum of contribution percentages of all size bins. In addition, PNC source apportionments in all 4 months are conducted with the selected technical paths based on these comparisons.

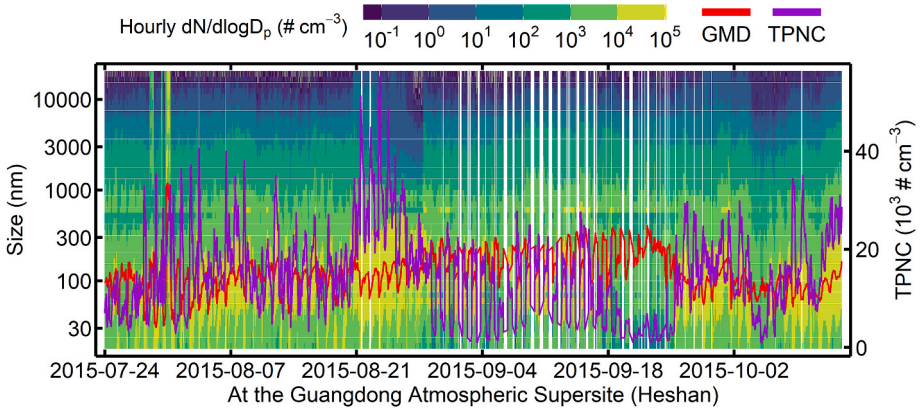

Fig. 1. Time series of PNSD, GMD, and TPNC at the Guangdong Atmospheric Supersite (Heshan) in 2015.

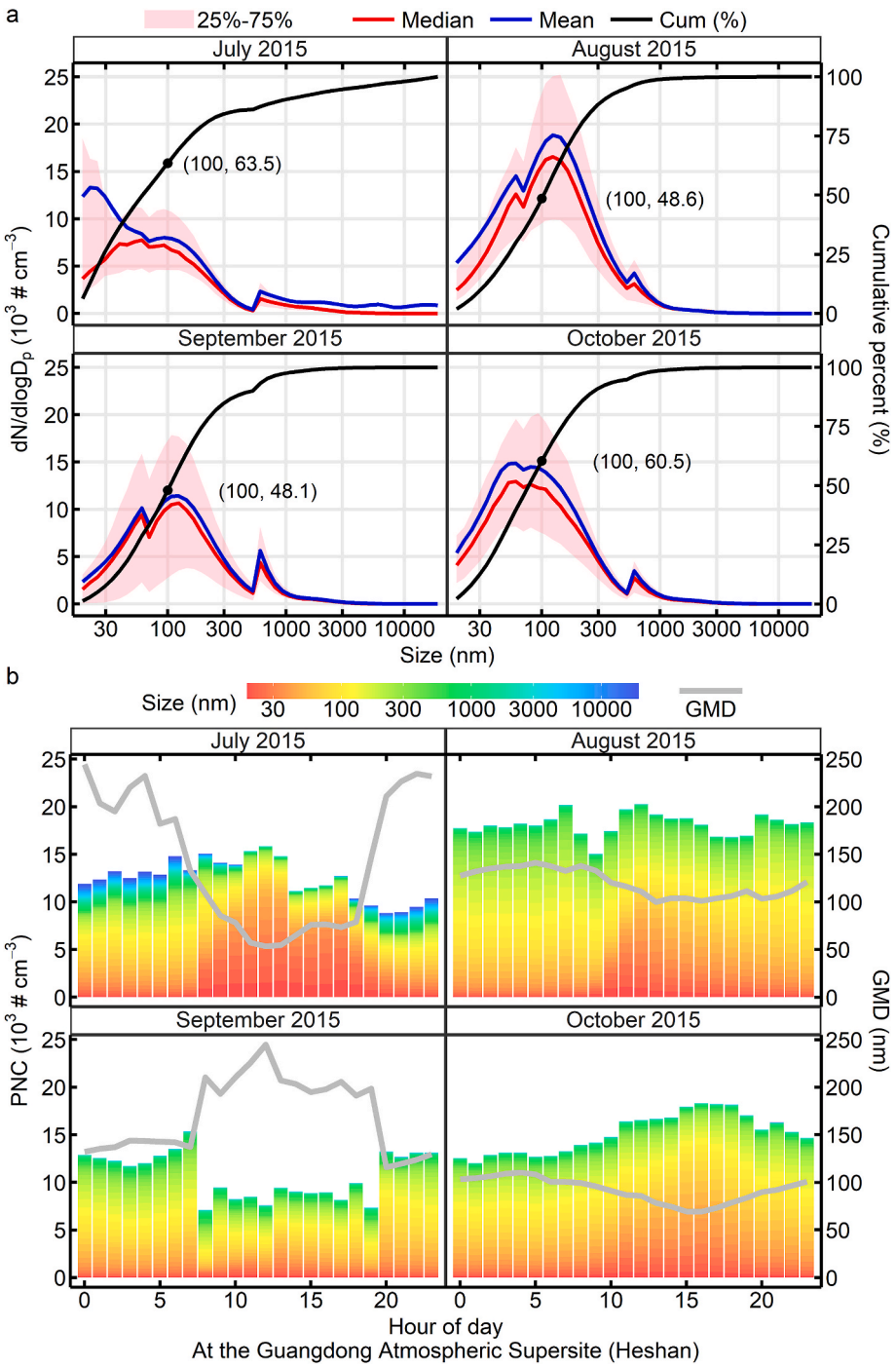

Fig. 2. Hourly averaged and cumulative PNSD and hour of day variations of size-divided PNC and GMD.

2. Materials and methodology

2.1. Location and observation

The site is at the Guangdong Atmospheric Supersite (Heshan, lat = 22.728, lon = 112.929) (Fig. S1). It is on a hill with trees, plants, and roads around (Chang et al., 2019) and influenced by regional transport of air pollutants from Guangzhou and Foshan (Yue et al., 2015).

Three particle sizers (measurement range) scanning mobility particle sizer (NSMPS, differential mobility analyzer (DMA) 3085, condensational particle counter (CPC) 3776, 2.21–60.4 nm, 24 bins), SMPS (DMA 3081, CPC 3775, 19.1–930.6 nm, 28 bins), and aerodynamic particle sizer (APS, 3321, one bin for < 0.523 μm and 51 bins for 0.542–19.81 μm) (TSI) were used for the PNSD observation from July to October 2015. The time resolution of all the measurements is 5-min. The channel resolution of these measurements is integrated into 32 per decade of particle size.

2.2. Accessory data

The accessory data of the PNSD spectra mainly include criteria (regular) pollutants (including ground-level ozone O_3, particulate matter PM, carbon monoxide CO, lead Pb, sulfur dioxide SO_2, and nitrogen dioxide NO_2, see https://www.epa.gov/criteria-air-pollutants, here only O_3, SO_2, NO_2, CO, and $PM_{2.5}$ are used) and meteorological parameters (wind speed (WS), wind direction (WD), atmospheric pressure (P),

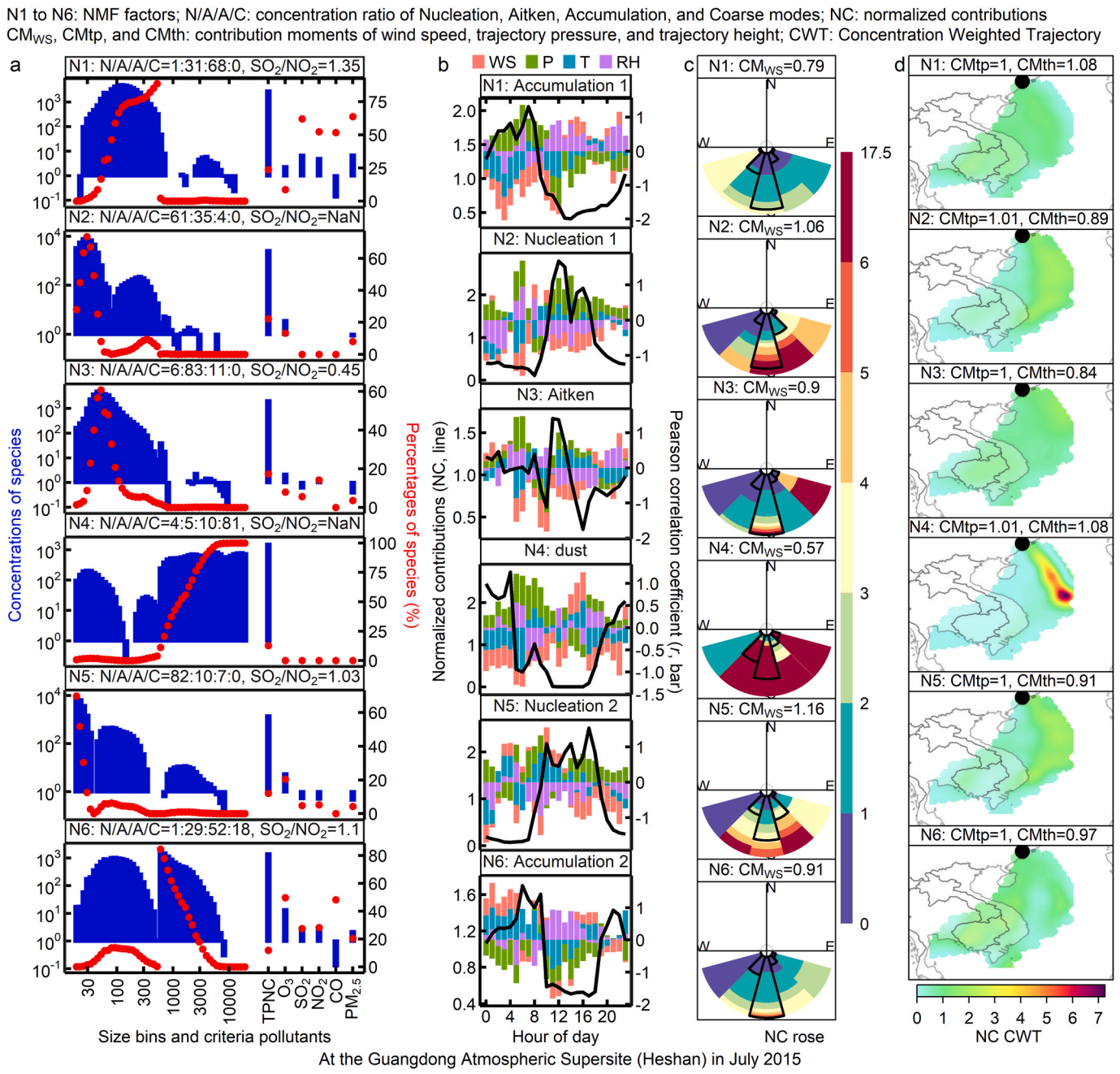

Fig. 3. NMF results of the whole apportionment in July 2015.

temperature (T), and relative humidity (RH)) monitored in the site and backward trajectories from National Oceanic and Atmospheric Administration (NOAA) accessed by using R package splitr (Iannone, 2021).

2.3. Data analysis tools

The R packages in system library (The R Core Team, 2021) and some other packages in user library were used to deal with the data. The selected user library R packages are listed in Table 1.

2.4. Nonnegative matrix factorization and contribution moment

The formula of NMF is:

$$X \approx WH \quad (1)$$

where X is a matrix of M variables and N observations (Lee and Seung, 1999). NMF can find the nonnegative base matrix W (M × L) and the coefficient matrix H (L × N) to meet X ≈ WH (Gaujoux and Seoighe, 2010; Gaujoux and Seoighe, 2020).

Contribution moment (CM) was put forward to compare the influence of parameters (variables) such as wind and trajectory (Liang et al., 2020). CM equals a normalized parameter (average is 1) times a normalized contribution (average is 1).

$$CM_{Parm} = \mathrm{Parm}_1^{Mean} \times C_1^{Mean} \quad (2)$$

where Parm means parameter, Parm_1^{Mean} means the normalized parameter (average is 1), and C_1^{Mean} means the normalized contribution (average is 1) (Liang et al., 2020). The specific expressions were clarified in the reference (Liang et al., 2020).

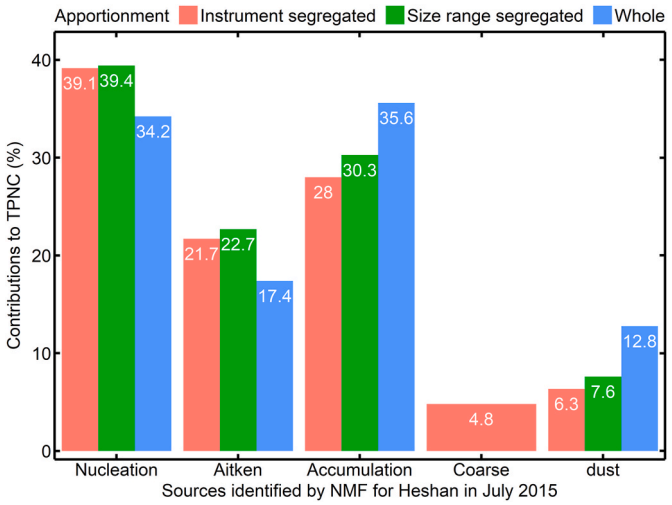

Fig. 4. Source apportionment (percentages using TPNC) by NMF in three different methods in July 2015.

3. Results and discussion

3.1. Preprocessing of data

The new R package dataprep (Liang et al., 2020, 2021) was used to preprocess the raw PNC data (Fig. S2). The values in data from NSMPS and SMPS equal to 0 were removed (replaced by missing values). The arguments in function were set according to the principles of accuracy and completeness to keep more samples and remove obvious outliers. The 'accuracy' principle means deleting variables, observations, and outliers neither excessively nor insufficiently within strict constraints. The 'completeness' principle means fully respecting the original data by avoiding changes or preprocessing whenever feasible and rigorously interpolating the missing values by considering the spatiotemporal gaps in the horizontal and vertical directions to keep as many samples (observations and variables) and values as possible. The outliers are defined as follows:

$$x = \text{Max and } x\ (\text{Max}) > Q_T \times (1 + E_T) + M_{TQ} \times C_{TM} \quad (3)$$

where Max is the maximum particle number concentration of a size bin, Q_T is the top quantile (percentile) with a default of 0.995 (99.5 %), E_T is the top allowable error coefficient with a default of 0.1, M_{TQ} is the order of magnitude of Q_T, and C_{TM} is the coefficient (default is 0.2) of M_{TQ}.

$$x = \text{Min and } x\ (\text{Min}) < Q_B \times (1 - E_B) - M_{BQ} \times C_{BM} \quad (4)$$

where Min is the minimum particle number concentration of a size bin, Q_B is the bottom quantile (percentile) with a default of 0.0025 (0.25 %), E_B is the bottom allowable error coefficient with a default of 0.2, M_{BQ} is the order of magnitude of Q_B, and C_{BM} is the coefficient (default is 0.4) of M_{BQ}. The different parameter values are chosen based on the improvement (Liang et al., 2020) from previous work (Masiol et al., 2017; Masiol et al., 2016). Allowable error coefficient and order of magnitude of quantiles and its coefficient were added as constraints to improve the traditional percentile-based outlier removal method. The new method deletes less values, deletes more completely, and generates no new outliers, compared with previous methods. Results (e.g., calculations and line plots by methods in R package dataprep (Liang et al., 2021)) of top and bottom percentiles, number of retained samples, amount of removed outliers, and descriptive statistics (e.g., standard deviation) should be considered in setting these values. The top and bottom parameters play symmetrical roles but their values are not because the bottom is less dispersed and influential than the top (Liang et al., 2020).

The outliers are removed in a one-by-one way by considering every extremum. The quantiles and constraints are re-estimated after each removal. Both the top and bottom outliers are removed simultaneously (order-irrelevant) after they are calculated together in one step. Usually, this procedure will be repeated after a previous removal operation for many times. Besides the default values for most arguments, the value of fraction in the function dataprep was set as 0.18 for the raw PNC data here.

After preprocessing, 20645 observations were retained from the original 23,616 observations, outliers were removed, and missing values were linearly interpolated based on the incomplete and dispersed degrees of the raw data (Fig. S3). A total of 15 variables (size bins) were deleted: 2.21, 2.55, 2.94, 3.40, 3.92, 4.53, 5.23, 6.04, 6.98, 8.06, 9.31, 10.7, 12.4, 14.3, and 16.5 nm. It is because of the difficulty in measurement. For example, the standard deviations and coefficients of variation of the size bins below 17 nm are very high (Fig. S4). The preprocessing here keeps more samples than ordinary methods and adds no new outliers after interpolation.

3.2. Pollution characteristics

The 5-min data (Fig. S3) were averaged into hourly data (Fig. 1) for analysis of characteristics and sources. The trends of geometric mean diameter (GMD) and TPNC are opposite. In July and August (hotter months) the peaks of TPNC are higher than those in September and October. The TPNC are highly dependent on PNC of size range 30–300 nm.

The concentration percentages of ultrafine particles (UFP) are about half or even more of the TPNC, ranging from 48.1 % to 63.5 % (Fig. 2a). However, these ratios are not so high than usually observed, because the 15 smallest size bins were deleted due to excessive missing values, resulting in an important part of UFP is not taken into account. Particles larger than 1000 nm account for very few (1.5–7.1 %) in TPNC. The TPNC are generally higher in daytime than in nighttime, as shown in July, August, and October (Fig. 2b). Compared with these three months, the case of September is not like this and less representative because of containing relatively more missing observations.

3.3. Source apportionment

The preprocessed PNC data are from three different particle sizers NSMPS, SMPS, and APS, still involving the four modes Nucleation, Aitken, Accumulation, and Coarse, but 15 smallest size bins were deleted from the raw data. To test the performance of NMF, three kinds of apportionments based on size ranges were compared. First, PNC data from different instruments were separately used in source apportionments by NMF. Second, similarly, PNC data with sizes below and above 100 nm were separately input into NMF for source apportionments. Third, the whole PNC data were used in one source apportionment procedure by NMF.

3.3.1. Instrument segregated apportionment

According to the peaks of PNSD and normalized contributions, the correlations between normalized contributions and meteorological parameters and their diurnal patterns, contributions to criteria pollutants (O_3, SO_2, NO_2, CO, and $PM_{2.5}$), and contribution moments (wind rose and backward trajectories) (Liang et al., 2020), the factors N1 to N9 (Fig. S5) in the instrument segregated apportionment are named as: N1

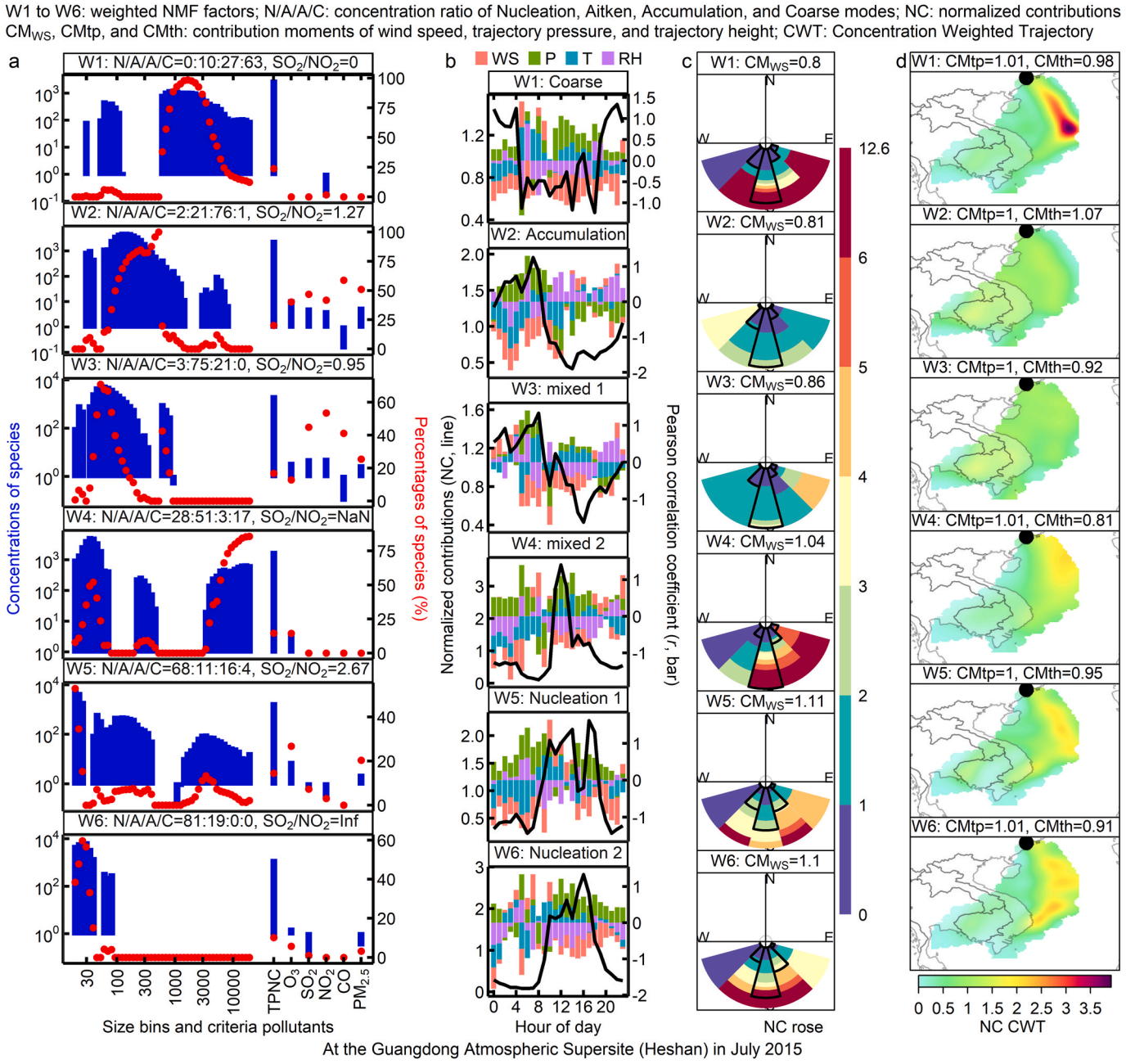

Fig. 5. 0–1 normalization weighted NMF results of the whole apportionment in July 2015.

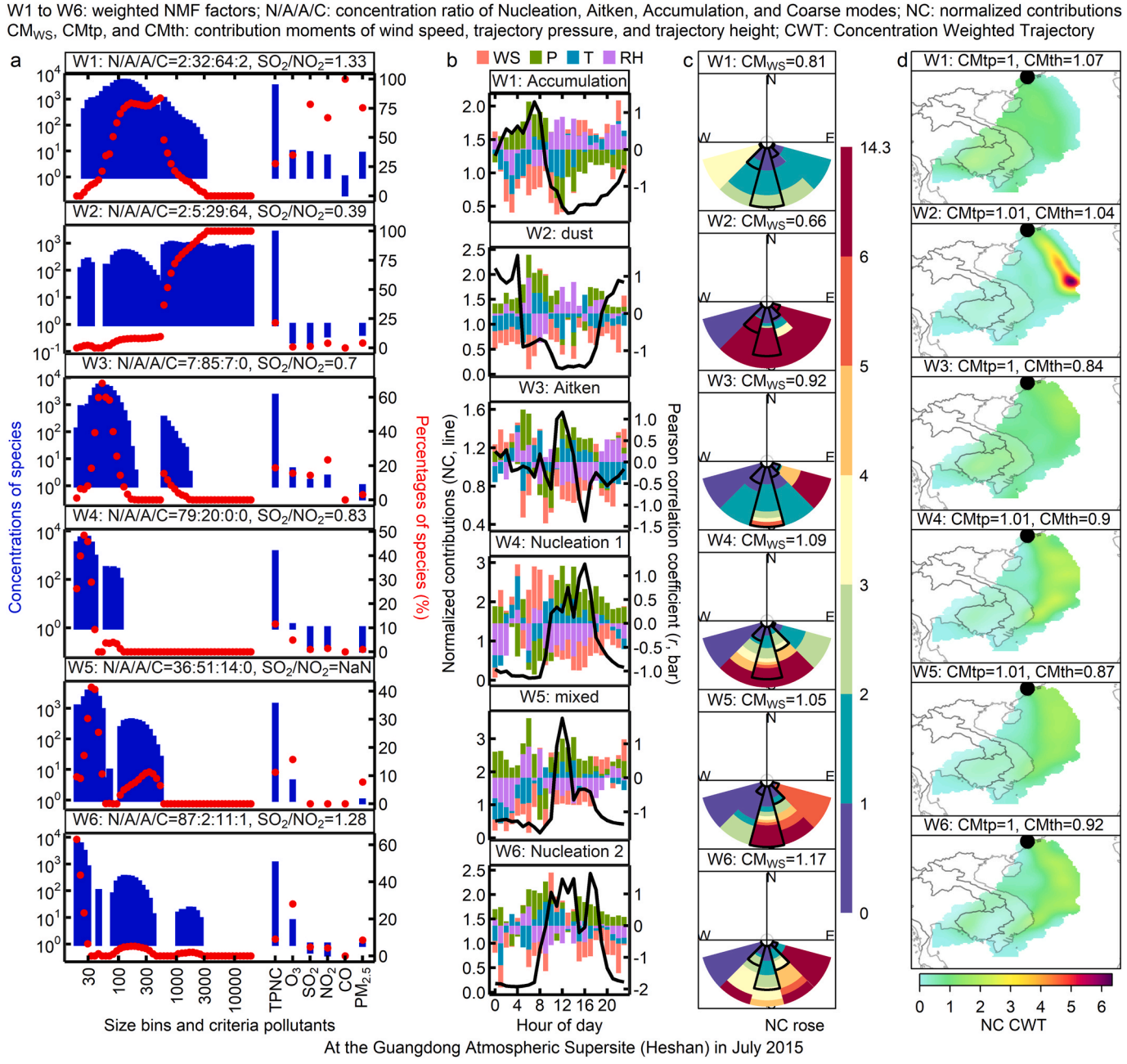

Fig. 6. Binary weighted NMF results of the whole apportionment in July 2015.

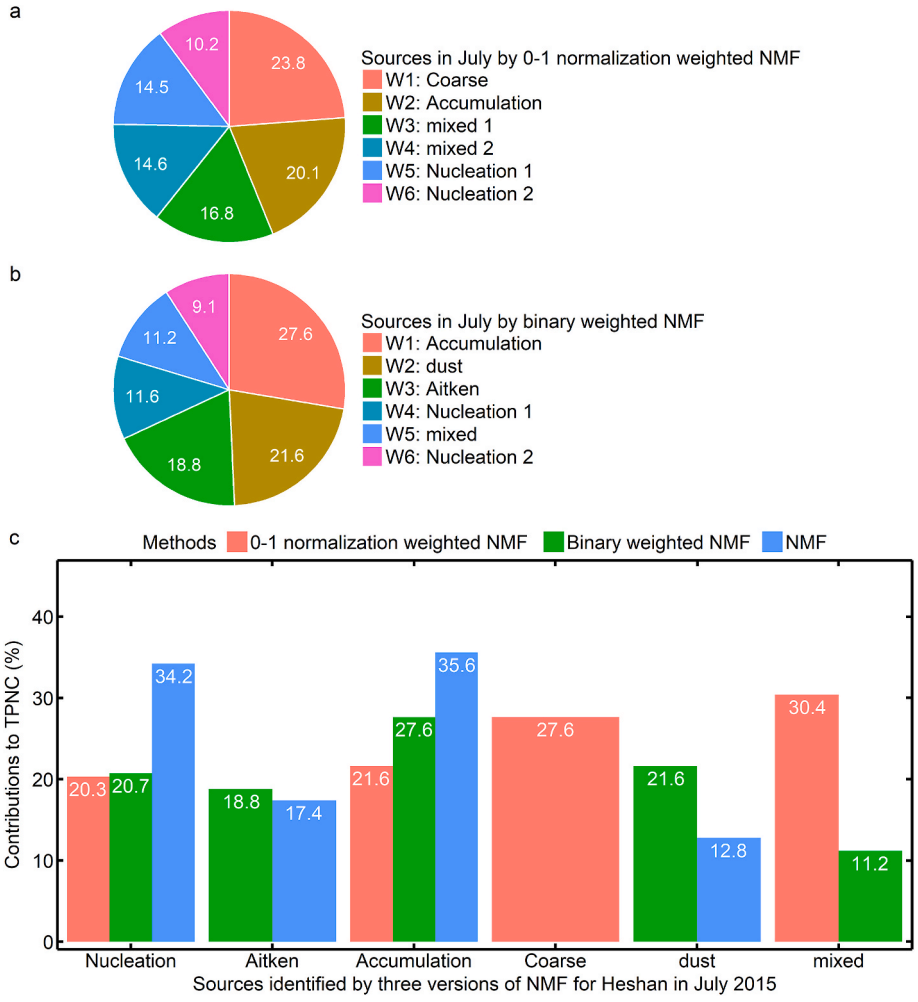

Fig. 7. Sources and their percentages using three versions of NMF on a whole apportionment basis.

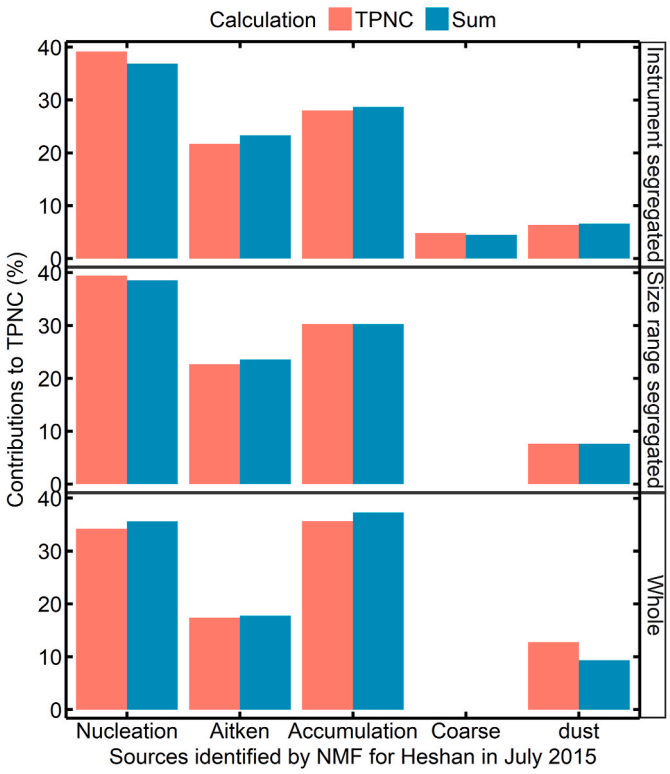

Fig. 8. Source apportionment by NMF for TPNC with two different calculations on a whole apportionment basis in July 2015.

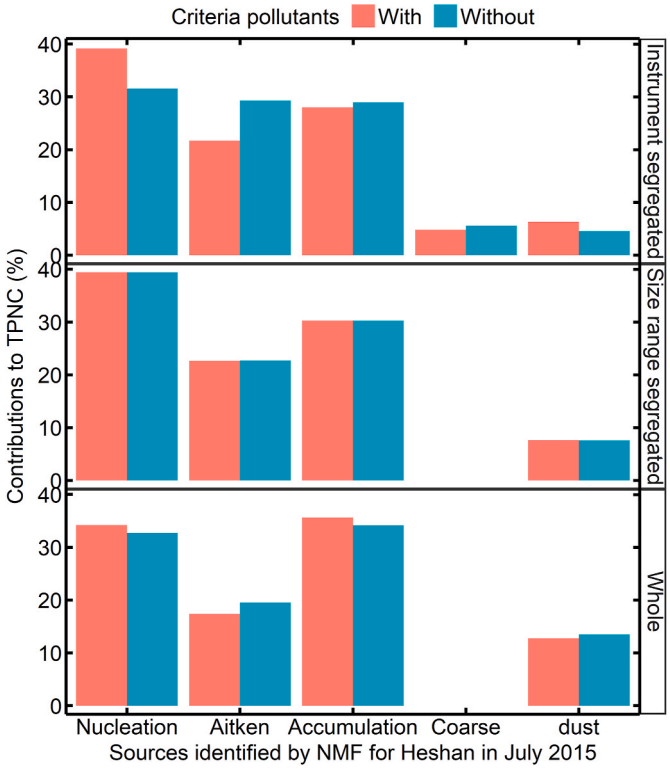

Fig. 9. Source apportionment by NMF for TPNC with and without criteria pollutants on a whole apportionment basis in July 2015.

is Nucleation (process source, the same below), N2 is Nucleation, N3 is Aitken (process source, the same below), N4 is Accumulation (process source, the same below), N5 is Accumulation, N6 is Aitken, N7 is dust, N8 is Coarse (process source, the same below), and N9 is Accumulation.

3.3.2. *Size range segregated apportionment*

Similarly (Liang et al., 2020), the factors N1 to N6 (Fig. S6) in the size range segregated apportionment are named as: N1 is Aitken, N2 is Nucleation, N3 is Nucleation, N4 is Accumulation, N5 is dust, and N6 is Accumulation.

3.3.3. *Whole apportionment*

Likewise (Liang et al., 2020), the factors N1 to N6 (Fig. 3) in the whole apportionment are named as: N1 is Accumulation, N2 is Nucleation, N3 is Aitken, N4 is dust, N5 is Nucleation, and N6 is Accumulation.

The following Sections 3.3.4 through 3.3.7 are based on the data of July 2015 only.

3.3.4. *Comparisons of different apportionments*

The contribution percentages of TPNC calculated from the three different PNC source apportionments are shown in Fig. 4. The absolute difference between the instrument segregated and the whole apportionments is: Nucleation (4.9 %), Aitken (4.3 %), Accumulation (7.6 %), Coarse (4.8 %), dust (6.4 %), with a mean of 5.6 %. Likewise, the absolute difference between the size range segregated and the whole apportionments is: Nucleation (5.2 %), Aitken (5.3 %), Accumulation (5.3 %), Coarse (0 %), dust (5.2 %), with a mean of 4.2 %. If the percentages are calculated by using weighted sum of contribution percentages of all size bins, similar result will be gotten (Fig. S7).

3.3.5. *Comparisons of NMF and weighted NMF*

Figs. 5 and 6 show the 0–1 (minimum to maximum) normalization and binary (0 for outlier or interpolated entry and 1 for other values) weighted NMF results of the whole apportionment in July 2015 respectively. More information of weighted NMF can be found elsewhere (Ho, 2008). The characteristic peaks of concentrations and percentages of weighted NMF are more mixed (with multiple peaks, W3 and W4 in Fig. 5a and W5 in Fig. 6a) than those of NMF (Fig. 3a). A mixed factor obviously contains multiple main sources and hence is hard to apportion (be assigned to one certain source).

The sources and their percentages apportioned by weighted NMF are shown in Fig. 7a and Fig. 7b. The difference between the outcomes of NMF and weighted NMF mainly lies in the mixed sources (Fig. 7c). Mixed sources are obscure and not so meaningful in practical use that needs explicit results, so the NMF rather than the weighted NMF is chosen to be the receptor model in this work. Although an ideal weight method of NMF is not found in this work, weighted NMF is worth of exploring and developing in the future for PNC source apportionment, especially based on optimizing the binary weighted NMF which is more promising (with one less mixed source) than the 0–1 normalization weighted NMF.

3.3.6. *Comparisons of calculations of contribution percentages of TPNC*

Specifically, the contribution percentages of TPNC calculated by both using TPNC contribution percentages alone and weighted sum of contribution percentages of all size bins are shown in Fig. 8. The absolute difference between these two calculations is small, ranging from 0 % to 3.4 % and averaging at 1.1 %.

3.3.7. *Influence of criteria pollutants on the matrix factorization by NMF*

To examine the influence of criteria pollutants (O_3, SO_2, NO_2, CO, and $PM_{2.5}$) on the matrix factorization by NMF, PNC source

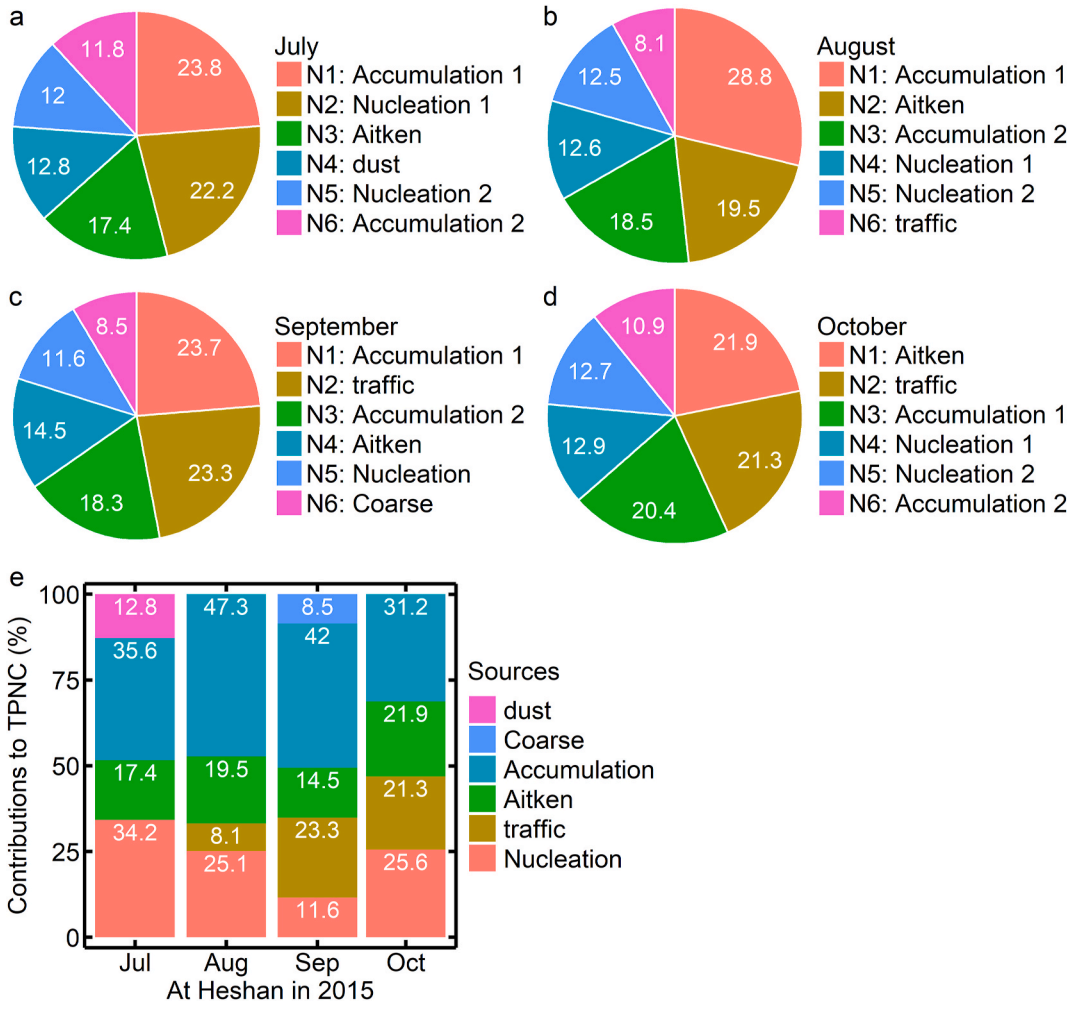

Fig. 10. Sources and their percentages and variations.

Table 2
Samples of complete and linearly interpolated cases.

Month-year	Number of samples	
	Interpolated	Complete
Jul-15	178	178
Aug-15	720	692
Sep-15	553	398
Oct-15	297	294

apportionments without criteria pollutants were also conducted for July 2015 (Figs. S8–S10), relative to the results above (Figs. S5, S6, and 3). The comparison of different apportionment methods (instrument segregated, size range segregated, and whole) can be further made here. The absolute difference between the instrument segregated and the whole apportionments is: Nucleation (1.2 %), Aitken (9.8 %), Accumulation (5.2 %), Coarse (5.6 %), dust (8.9 %), with a mean of 6.1 % (Fig. S11). Likewise, the absolute difference between the size range segregated and the whole apportionments is: Nucleation (6.7 %), Aitken (3.2 %), Accumulation (3.9 %), Coarse (0 %), dust (5.9 %), with a mean of 3.9 % (Fig. S11). For the influence of criteria pollutants, in general, the absolute difference of the contribution percentages of TPNC between with and without criteria pollutants is small, ranging from 0 % to 7.6 % and averaging at 1.9 % (Fig. 9).

3.3.8. *PNC source apportionments of all months*

Similar to that of July 2015 (Fig. 3), the PNC source apportionments of other months August, September, and October 2015 were conducted by using NMF on a whole apportionment basis. The results can be seen in Figs. S12–S14. The sources and their percentages are summarized in Fig. 10. The 6 apportioned sources (normalized averages of available contribution percentages in the 4 months) include Accumulation (32.4 %), Nucleation (20.0 %), Aitken (15.2 %), traffic (14.6 %), dust (10.6 %), and Coarse (7.1 %).

3.4. *Effect of the interpolation method on the NMF performance*

The effect of the linear interpolation on the NMF performance can be investigated by a comparison between complete cases (omitting samples with missing values) and linearly interpolated cases. An obvious difference is the number of samples. The complete case of July has the same number of samples with that of the linearly interpolated case while the complete samples are many (155) less than the interpreted one in September (Table 2), which are also reflected by the missing values gaps in the time series (Fig. 1). However, the numbers of 5-min samples in July 2015 are 2304 and 2198 for the interpolated and complete cases respectively, though they have the same number of hourly samples.

The two months July and September, with the smallest and biggest differences in number of samples among the complete and linearly interpolated cases respectively, were chosen to investigate the influence of linear interpolation on NMF on a whole apportionment basis. The NMF results of July (Fig. S15) and September (Fig. S16) without linear

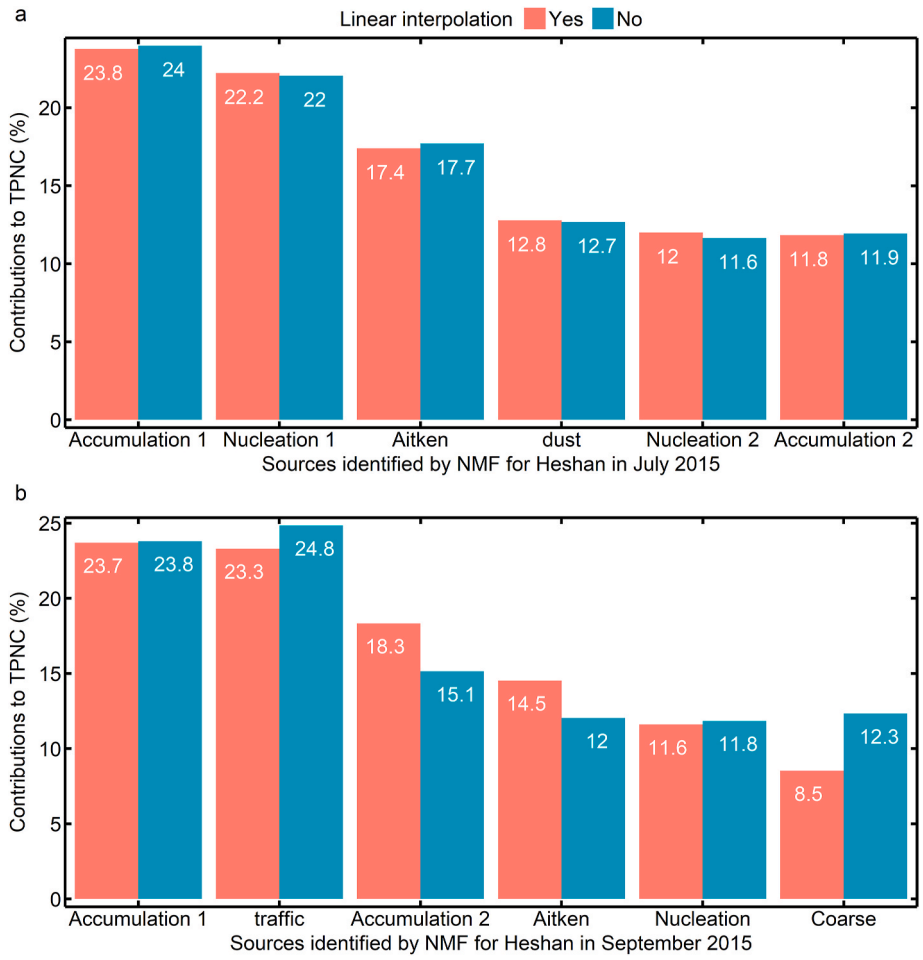

Fig. 11. Influence of linear interpolation on sources and their percentages.

interpolation for PNC show small and big differences from those (Fig. 3 and S13) with linear interpolation for PNC respectively, responding to the differences in number of samples. They have the same factor names (sources), but their mean absolute differences in contribution percentages are 0.2 % (0.1–0.4 %) and 1.9 % (0.1–3.8 %) for July and September respectively (Fig. 11). These indicate the linear interpolation would not change the kinds of sources (quality) but would change the contribution percentages (quantity) via the retained samples.

3.5. *Further test of the NMF based PNC source apportionment method*

Since the duration of the 4-month sampling period is not so long, some more data gotten from the Internet were used to test the above proposed NMF based PNC source apportionment method. The data are from a European station named Ispra (lat = 45.8, lon = 8.633) in Varese, Italy. It is a rural station but with many sport clubs (such as comprehensive sports center, golf course, rowing club, and football field), restaurants, and small roads around. The data of PNC (10–20,000 nm) and criteria pollutants (O_3, SO_2, NO_2, and CO) of Ispra in 2017 were downloaded from EBAS (a database storing observation data of atmosphere) managed by the Norwegian Institute for Air Research (NILU) (Torseth et al., 2012). Their accessory data meteorological parameters (WS, WD, P, T, and RH in Milan Malpensa Airport, 20.2 km away from the Ispra station) were obtained by using R package worldmet (Carslaw, 2021) and backward trajectories were obtained by using R package splitr (Iannone, 2021) from the NOAA. The PNC source apportionment was operated successfully by using the above proposed method on a whole apportionment basis (Fig. 12). N1 is aged traffic, N2 is fresh traffic, N3 is Accumulation 1, N4 is Accumulation 2, N5 is Nucleation 1, and N6 is Nucleation 2.

The contribution percentages of the NMF factors (sources) to TPNC range from 9 % to 25.3 % (Fig. 13a). Traffic, Accumulation, and Nucleation contributed 46.1 %, 32.7 %, and 21.2 % to the TPNC (Fig. 13b), respectively. Compared with the relatively remote station Heshan (on a hill) where traffic only contributed 8.1–23.3 % during August to October in 2015 (Fig. 10), this rural station Ispra is much more influenced by traffic. The closer a station is to roads, the higher contribution of traffic to PNC might be. PNC source apportionment is very sensitive to human activities such as transport, depending on the distance between a receptor site and the anthropogenic sources.

4. Conclusions

Particle number concentrations (PNC) with size range 2–20,000 nm at the Guangdong Atmospheric Supersite (Heshan) during July–October 2015 were observed by using NSMPS, SMPS, and APS. After preprocessing, the data of PNC with size range 19–20,000 nm were retained and used for source apportionments by using NMF. There are three main conclusions from the source apportionment for the exemplary July 2015. First, different tests show that the instrument segregated, size range segregated, and whole apportionments are similar in contribution percentages of factors, with average absolute difference around 5 %. Second, the TPNC contribution percentages themselves are very close to the weighted sum of contribution percentages of all size bins, with average absolute difference of 1.1 %. Third, the absolute difference of the contribution percentages of TPNC between with and without criteria

N1 to N6: NMF factors; N/A/A/C: concentration ratio of Nucleation, Aitken, Accumulation, and Coarse modes; NC: normalized contributions
CM_{WS}, CMtp, and CMth: contribution moments of wind speed, trajectory pressure, and trajectory height; CWT: Concentration Weighted Trajectory

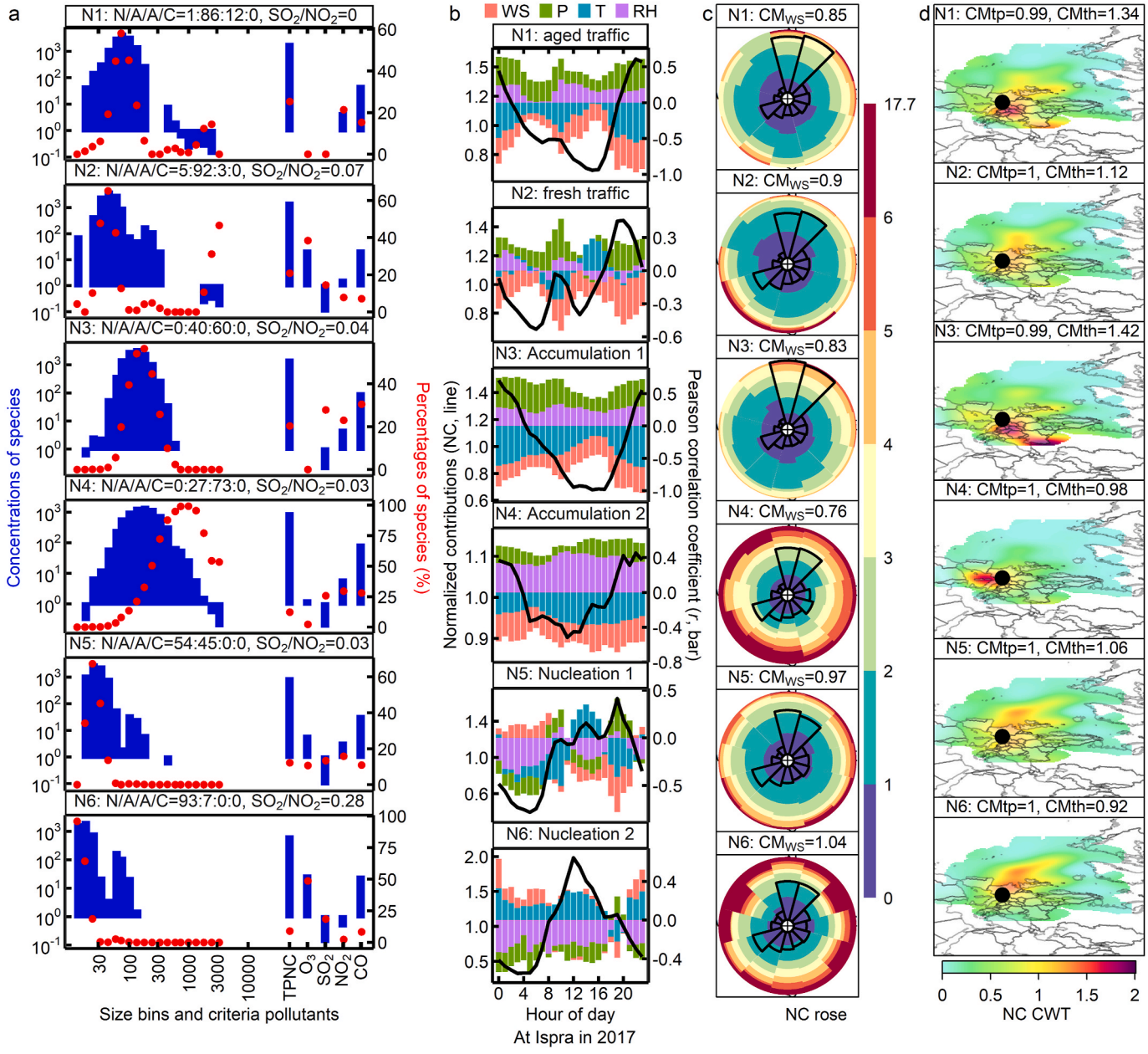

Fig. 12. NMF results of Ispra in 2017 on a whole apportionment basis.

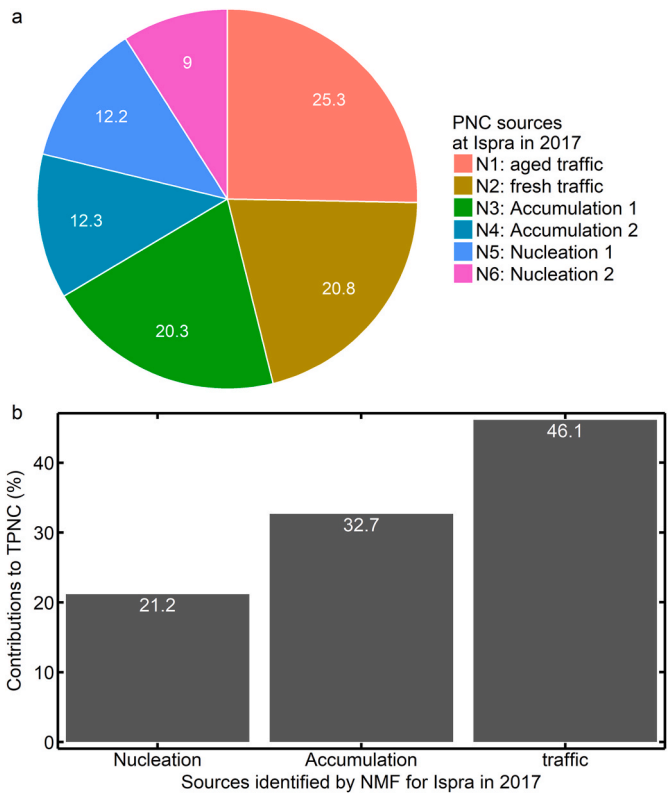

Fig. 13. Sources and their percentages at Ispra in 2017 on a whole apportionment basis.

pollutants is small (averaging at 1.9 %). In other words, conducting the wide size PNC source apportionment in one step as a whole rather than one by one separately is feasible, using the TPNC percentage is alternatively enough instead of necessarily calculating the weighted sum, and combing the criteria pollutants with size bins is helpful with few disturbances. The method also proved successful in the case of Ispra in 2017. The promising binary weighted NMF is worth of optimizing to develop weight method for NMF based PNC source apportionment in the future. In all, NMF is decent and efficient for source apportionment of PNC with wide size range. And, the reduced and integrated steps make the analysis of particle sources (or other datasets with a large number of numeric variables and observations) via NMF simpler.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Gansu Provincial Special Fund Project for Guiding Scientific and Technological Innovation and Development (2019ZX-06) and Fundamental Research Funds for the Central Universities (lzujbky-2020-kb31 and lzujbky-2021-kb12). We thank the data originators (Putaud, JP; Martins dos Santos, Sebastiao; Lagler, F; and Jensen, NR at European Commission's Joint Research Centre) for the Ispra data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2021.117846.

Author statement

Chun-Sheng Liang: data analysis and writing. **Dingli Yue**: observation and writing, **Hao Wu**: writing and review. **Jin-Sen Shi**: funding acquisition. **Ke-Bin He**: conceptualization and supervision.

References

Almeida, S.M., Manousakas, M., Diapouli, E., Kertesz, Z., Samek, L., Hristova, E., Sega, K., Alvarez, R.P., Belis, C.A., Eleftheriadis, K., Grp, I.E.R.S., 2020. Ambient particulate matter source apportionment using receptor modelling in European and Central Asia urban areas. Environ. Pollut. 266.

Bache, S.M., Wickham, H., 2014. Magrittr: A Forward-Pipe Operator for R. The R Foundation, Vienna, Austria. R package version 1.5. https://CRAN.R-project.org/package=magrittr.

Beddows, D.C.S., Dall'Osto, M., Harrison, R.M., 2009. Cluster Analysis of rural, urban, and curbside atmospheric particle size data. Environ. Sci. Technol. 43, 4694–4700.

Beddows, D.C.S., Harrison, R.M., Green, D.C., Fuller, G.W., 2015. Receptor modelling of both particle composition and size distribution from a background site in London, UK. Atmos. Chem. Phys. 15, 10107–10125.

Cai, J., Chu, B.W., Yao, L., Yan, C., Heikkinen, L.M., Zheng, F.X., Li, C., Fan, X.L., Zhang, S.J., Yang, D.Y., Wang, Y.H., Kokkonen, T.V., Chan, T., Zhou, Y., Dada, L., Liu, Y.C., He, H., Paasonen, P., Kujansuu, J.T., Petaja, T., Mohr, C., Kangasluoma, J., Bianchi, F., Sun, Y.L., Croteau, P.L., Worsnop, D.R., Kerminen, V.M., Du, W., Kulmala, M., Daellenbach, K.R., 2020. Size-segregated particle number and mass concentrations from different emission sources in urban Beijing. Atmos. Chem. Phys. 20, 12721–12740.

Carslaw, D., 2021. Worldmet: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD). R package version 0.9.5. https://CRAN.R-project.org/package=worldmet.

Carslaw, D., Ropkins, K., 2019. Openair: Tools for the Analysis of Air Pollution Data. R package version 2.7-0. https://cran.r-project.org/web/packages/openair/.

Carslaw, D.C., Ropkins, K., 2012. Openair - an R package for air quality data analysis. Environ. Model. Software 27–28, 52–61.

Chang, D., Wang, Z., Guo, J., Li, T., Liang, Y.H., Kang, L.Y., Xia, M., Wang, Y.R., Yu, C., Yun, H., Yue, D.L., Wang, T., 2019. Characterization of organic aerosols and their precursors in southern China during a severe haze episode in January 2017. Sci. Total Environ. 691, 101–111.

Charron, A., Birmili, W., Harrison, R.M., 2008. Fingerprinting particle origins according to their size distribution at a UK rural site. J. Geophys. Res. Atmos. 113.

Coquelin, L., Le Brusquet, L., Fischer, N., Gensdarmes, F., Motzkus, C., Mace, T., Fleury, G., 2018. Uncertainty propagation using the Monte Carlo method in the measurement of airborne particle size distribution with a scanning mobility particle sizer. Meas. Sci. Technol. 29.

Delmaire, G., Omidvar, M., Puigt, M., Ledoux, F., Limem, A., Roussel, G., Courcot, D., 2019. Informed weighted non-negative matrix factorization using αβ-divergence applied to source apportionment. Entropy 21, 253.

Delmaire, G., Roussel, G., Hleis, D., Ledoux, F., 2010. Une version pondérée de la Factorisation Matricielle Non Negative pour l'identification de sources de particules atmospheriques. Application au littoral de la Mer du Nord. Apii Jesa J. Eur. Syst. Automat. 44, 547–566.

Demin, G., 2019. Expss: Tables, Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics. R package version 0.10.1. https://CRAN.R-project.org/package=expss.

Fellows, I., Stotz, J.P., 2019. OpenStreetMap: Access to Open Street Map Raster Images. R package version 0.3.4. https://CRAN.R-project.org/package=OpenStreetMap.

Gaujoux, R., Seoighe, C., 2010. A flexible R package for nonnegative matrix factorization. BMC Bioinf. 11.

Gaujoux, R., Seoighe, C.N.M.F., 2020. Algorithms and Framework for Nonnegative Matrix Factorization (NMF). R package version 0.23.0. https://cran.r-project.org/web/packages/NMF/.

Grolemond, G., Wickman, H., 2011. Dates and times made easy with lubridate. J. Stat. Software 40, 1–25.

Hagan, D.H., Gani, S., Bhandar, S., Patel, K., Habib, G., Apte, J.S., Hildebrandt Ruiz, L., Kroll, J.H., 2019. Inferring aerosol sources from low-cost air quality sensor measurements: a case study in Delhi, India. Environ. Sci. Technol. Lett. 6, 467.

Harrison, R.M., Beddows, D.C.S., Dall'Osto, M., 2011. PMF analysis of wide-range particle size spectra collected on a major highway. Environ. Sci. Technol. 45, 5522–5528.

Ho, N.-D., 2008. Nonnegative Matrix Factorization Algorithms and Application. PhD thesis. Université catholique de Louvain.

Hopke, P.K., Dai, Q.L., Li, L.X., Feng, Y.C., 2020. Global review of recent source apportionments for airborne particulate matter. Sci. Total Environ. 740.

Horii, Y., Ohtsuka, N., Minomo, K., Takemine, S., Motegi, M., Hara, M., 2021. Distribution characteristics of methylsiloxanes in atmospheric environment of Saitama, Japan: diurnal and seasonal variations and emission source apportionment. Sci. Total Environ. 754.

Hussein, T., Molgaard, B., Hannuniemi, H., Martikainen, J., Jarvi, L., Wegner, T., Ripamonti, G., Weber, S., Vesala, T., Hameri, K., 2014. Fingerprints of the urban particle number size distribution in Helsinki, Finland: local versus regional characteristics. Boreal Environ. Res. 19, 1–20.

Iannone, R., 2021. Use the HYSPLIT Model from inside R. R package version 0.4.0.9000. https://github.com/rich-iannone/SplitR.

Kangasluoma, J., Ahonen, L.R., Laurila, T.M., Cai, R.L., Enroth, J., Mazon, S.B., Korhonen, F., Aalto, P.P., Kulmala, M., Attoui, M., Petaja, T., 2018. Laboratory verification of a new high flow differential mobility particle sizer, and field measurements in Hyytiala. J. Aerosol Sci. 124, 1–9.
Kfoury, A., Ledoux, F., Roche, C., Delmaire, G., Roussel, G., Courcot, D., 2016. PM2.5 source apportionment in a French urban coastal site under steelworks emission influences using constrained non-negative matrix factorization receptor model. J. Environ. Sci. 40, 114–128.
Khan, M.F., Latif, M.T., Amil, N., Juneng, L., Mohamad, N., Nadzir, M.S.M., Hoque, H.M. S., 2015. Characterization and source apportionment of particle number concentration at a semi-urban tropical environment. Environ. Sci. Pollut. Res. 22, 13111–13126.
Kulkarni, P., Baron, P.A., 2011. An approach to performing aerosol measurements. Aerosol Measurement: Principles, Techniques, and Applications, 3th edition. John Wiley & Sons, Inc, pp. 55–65.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.
Lee, H.K., Lee, H., Ahn, K.H., 2020. Development of a new nanoparticle sizer equipped with a 12-channel multi-port differential mobility analyzer and multi-condensation particle counters. Atmos Meas Tech 13, 1551–1562.
Liang, C.-S., Wu, H., Li, H.-Y., Zhang, Q., Li, Z., He, K.-B., 2020. Efficient data preprocessing, episode classification, and source apportionment of particle number concentrations. Sci. Total Environ. 744, 140923.
Liang, C.-S., Wu, H., Li, H.-Y., Zhang, Q., Li, Z., He, K.-B., 2021. Dataprep: Efficient and Flexible Data Preprocessing Tools. R package version 0.1.3. https://cran.r-project.org/web/packages/dataprep.
Liang, C.S., Yu, T.Y., Chang, Y.Y., Syu, J.Y., Lin, W.Y., 2013. Source apportionment of PM2.5 particle composition and submicrometer size distribution during an Asian dust storm and non-dust storm in Taipei. Aerosol Air Qual Res 13, 545–554.
Liu, H.T., Tian, C.G., Zong, Z., Wang, X.P., Li, J., Zhang, G., 2019. Development and assessment of a receptor source apportionment model based on four nonnegative matrix factorization algorithms. Atmos. Environ. 197, 159–165.
Liu, Q.L., Liu, D., Chen, X.T., Zhang, Q., Jiang, J.K., Chen, D.R., 2020. A cost-effective, miniature electrical ultrafine particle sizer (mini-eUPS) for ultrafine particle (UFP) monitoring network. Aerosol Air Qual Res 20, 231–241.
Masiol, M., Harrison, R.M., Vu, T.V., Beddows, D.C.S., 2017. Sources of sub-micrometre particles near a major international airport. Atmos. Chem. Phys. 17, 12379–12403.
Masiol, M., Vu, T.V., Beddows, D.C.S., Harrison, R.M., 2016. Source apportionment of wide range particle size spectra and black carbon collected at the airport of Venice (Italy). Atmos. Environ. 139, 56–74.
Müller, K., Wickham, H., 2019. Tibble: Simple Data Frames. R package version 2.1.3. https://CRAN.R-project.org/package=tibble.
Njalsson, T., Novosselov, I., 2018. Design and optimization of a compact low-cost optical particle sizer. J. Aerosol Sci. 119, 1–12.
Ogulei, D., Hopke, P.K., Zhou, L.M., Pancras, J.P., Nair, N., Ondov, J.M., 2006. Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data. Atmos. Environ. 40, S396–S410.
Ouaret, R., Ionescu, A., Ramalho, O., Candau, Y., 2017. Indoor Air Pollutant Sources Using Blind Source Separation Methods. In: 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium.
Pey, J., Querol, X., Alastuey, A., Rodriguez, S., Putaud, J.P., Van Dingenen, R., 2009. Source apportionment of urban fine and ultra-fine particle number concentration in a Western Mediterranean city. Atmos. Environ. 43, 4407–4415.
Robinson, D., Hayes, A., 2019. Broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.3. https://CRAN.R-project.org/package=broom.
Rodins, V., Lucht, S., Ohlwein, S., Hennig, F., Soppa, V., Erbel, R., Jockel, K.H., Weimar, C., Hermann, D.M., Schramm, S., Moebus, S., Slomiany, U., Hoffmann, B., 2020. Long-term exposure to ambient source-specific particulate matter and its components and incidence of cardiovascular events - the Heinz Nixdorf Recall study. Environ. Int. 142, 14.
Sarkar, D., 2008. In: Lattice: Multivariate Data Visualization with R, ed. Springer, New York.
Sarkar, D., 2018. Lattice: Trellis Graphics for R. R package version 0.20-38. https://cran.r-project.org/web/packages/lattice/.
Scerri, M.M., Genga, A., Lacobellis, S., Delmaire, G., Giove, A., Siciliano, M., Siciliano, T., Weinbruch, S., 2019. Investigating the plausibility of a PMF source apportionment solution derived using a small dataset: a case study from a receptor in a rural site in Apulia - South East Italy. Chemosphere 236.
Shang, X., Zhang, K., Meng, F., Wang, S.H., Lee, M., Suh, I., Kim, D., Jeon, K., Park, H., Wang, X.Z., Zhao, Y.X., 2018. Characteristics and source apportionment of fine haze aerosol in Beijing during the winter of 2013. Atmos. Chem. Phys. 18, 2573–2584.
Slowikowski, K., 2019. Ggrepel: Automatically Position Non-overlapping Text Labels with 'ggplot2'. R package version 0.8.1. https://CRAN.R-project.org/package=ggrepel.
Spielvogel, J., Keck, L., Guo, X.A., Pesch, M., 2010. Comprehensive measurement of atmospheric aerosols with a wide range aerosol spectrometer. In: Ranzi, E. (Ed.), Aaas10: Advanced Atmospheric Aerosol Symposium. Aidic Servizi Srl, Milano.
Spinu, V., Grolemund, G., Wickham, H., 2018. Lubridate: Make Dealing with Dates a Little Easier. R package version 1.7.4. https://cran.r-project.org/web/packages/lubridate/.
Stolzenburg, M.R., McMurry, P.H., 2018. Method to assess performance of scanning mobility particle sizer (SMPS) instruments and software. Aerosol Sci. Technol. 52, 609–613.
Sun, Q., Hu, J.T., Wang, H.Q., Chen, M.J., Yu, F.J., Gui, H.Q., Liu, J.G., Lu, L., 2019. Iop. Design and Evaluation of a Mini-Faraday Cup for Portable Ultrafine Particle Sizer. Third International Conference on Energy Engineering and Environmental Protection.
Suzuki, Y., Matsunaga, K., Yamashita, Y., 2021. Assignment of PM2.5 sources in western Japan by non-negative matrix factorization of concentration-weighted trajectories of GED-ICP-MS/MS element concentrations. Environ. Pollut. 270.
The R Core Team. R: A Language and Environment for Statistical Computing, 2021. R Foundation for Statistical Computing, Vienna, Austria, Version 4.0.5. https://www.R-project.org/.
Torseth, K., Aas, W., Breivik, K., Fjaeraa, A.M., Fiebig, M., Hjellbrekke, A.G., Myhre, C.L., Solberg, S., Yttri, K.E., 2012. Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972-2009. Atmos. Chem. Phys. 12, 5447–5481.
Ushey, K., 2018. RcppRoll: Efficient Rolling/Windowed Operations. R package version 0.3.0. https://CRAN.R-project.org/package=RcppRoll.
Vo, E., Horvatin, M., Zhuang, Z.Q., 2018. Performance comparison of field portable instruments to the scanning mobility particle sizer using monodispersed and polydispersed sodium chloride aerosols. Ann. Work Expos. Health 62, 711–720.
Vu, T.V., Delgado-Saborit, J.M., Harrison, R.M., 2015. Review: particle number size distributions from seven major sources and implications for source apportionment studies. Atmos. Environ. 122, 114–132.
Wåhlin, P., Palmgren, F., Van Dingenen, R., 2001. Experimental studies of ultrafine particles in streets and the relationship to traffic. Atmos. Environ. 35, S63–S69.
Wickham, H., 2007. Reshaping data with the reshape package. J. Stat. Software 21, 1–20.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis, eds. Springer-Verlag, New York.
Wickham, H., 2017. reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. R package version 1.4.3. https://cran.r-project.org/web/packages/reshape2/.
Wickham, H., 2019a. Forcats: Tools for Working with Categorical Variables (Factors). R package version 0.4.0. https://CRAN.R-project.org/package=forcats.
Wickham, H., 2019b. Stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.
Wickham, H., Bryan, J., 2019. Readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl.
Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., 2019a. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.2.1. https://cran.r-project.org/web/packages/ggplot2/.
Wickham, H., Francois, R., Henry, L., Müller, K., 2019b. Dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr.
Wickham, H., Henry, L., 2019. Tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr.
Wickham, H., Pedersen, T.L., 2019. Gtable: Arrange 'Grobs' in Tables. R package version 0.3.0. https://CRAN.R-project.org/package=gtable.
Wickham, H., Seidel, D., 2019. Scales: Scale Functions for Visualization. R package version 1.1.0. https://CRAN.R-project.org/package=scales.
Wilke, C.O., 2019. Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot.
Wu, T., Boor, B.E., 2020. Urban aerosol size distributions: a global perspective. Atmos. Chem. Phys. Discuss. 2020, 1–83.
Xia, T., Catalan, J., Hu, C., Batterman, S., 2020. Development of a mobile platform for monitoring gaseous, particulate, and greenhouse gas (GHG) pollutants. Environ. Monit. Assess. 193, 7.
Yang, J., Wang, H.Q., Zhou, J.T., Chen, D.R., Kong, D.Y., Yu, F.J., Gui, H.Q., Liu, J.G., Chen, M.J., 2021. A build-in data inversion method to retrieve aerosol size distributions for a portable ultrafine particle sizer (PUPS). Ieee Access 9, 2879–2889.
Yu, G., 2019. Ggplotify: Convert Plot to 'grob' or 'ggplot' Object. R package version 0.0.4. https://CRAN.R-project.org/package=ggplotify.
Yue, D.L., Zhong, L.J., Zhang, T., Shen, J., Zhou, Y., Zeng, L.M., Dong, H.B., Ye, S.Q., 2015. Pollution properties of water-soluble secondary inorganic ions in atmospheric PM2.5 in the pearl river delta region. Aerosol Air Qual Res 15, 1737–1747.
Zeileis, A., Grothendieck, G. zoo, 2005. S3 infrastructure for regular and irregular time series. J. Stat. Software 14, 1–27.
Zeileis, A., Grothendieck, G., Ryan, J.A., 2019. Zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations). R package version 1.8-6. https://cran.r-project.org/web/packages/zoo/.
Zhang, K., Shang, X., Herrmann, H., Meng, F., Mo, Z., Chen, J., Lv, W., 2019. Approaches for identifying PM2.5 source types and source areas at a remote background site of South China in spring. Sci. Total Environ. 691, 1320–1327.
Zhang, W.H., Liu, B.S., Zhang, Y.F., Li, Y.F., Sun, X.Y., Gu, Y., Dai, C.L., Li, N., Song, C.B., Dai, Q.L., Han, Y., Feng, Y.C., 2020. A refined source apportionment study of atmospheric PM2.5 during winter heating period in Shijiazhuang, China, using a receptor model coupled with a source-oriented model. Atmos. Environ. 222.
Zhou, L.M., Kim, E., Hopke, P.K., Stanier, C.O., Pandis, S., 2004. Advanced factor analysis on Pittsburgh particle size-distribution data. Aerosol Sci. Technol. 38, 118–132.