



# Receptor modelling of both particle composition and size distribution from a background site in London, UK – a two-step approach

David C. S. Beddows<sup>1</sup> and Roy M. Harrison<sup>1,a</sup>

<sup>1</sup>National Centre for Atmospheric Science, School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

<sup>a</sup>also at: Department of Environmental Sciences/Center of Excellence in Environmental Studies, King Abdulaziz University, P.O. Box 80203, Jeddah, 21589, Saudi Arabia

**Correspondence:** David C. S. Beddows (d.c.beddows@bham.ac.uk)

Received: 30 July 2018 – Discussion started: 30 August 2018

Revised: 16 January 2019 – Accepted: 7 February 2019 – Published: 11 April 2019

**Abstract.** Some air pollution datasets contain multiple variables with a range of measurement units, and combined analysis using positive matrix factorization (PMF) can be problematic but can offer benefits through the greater information content. In this work, a novel method is devised and the source apportionment of a mixed unit dataset (PM<sub>10</sub> mass and number size distribution, NSD) is achieved using a novel two-step approach to PMF. In the first step the PM<sub>10</sub> data are PMF-analysed using a source apportionment approach in order to provide a solution which best describes the environment and conditions considered. The time series *G* values (and errors) of the PM<sub>10</sub> solution are then taken forward into the second step, where they are combined with the NSD data and analysed in a second PMF analysis. This results in NSD data associated with the apportioned PM<sub>10</sub> factors. We exemplify this approach using data reported in the study of Beddows et al. (2015), producing one solution which unifies the two separate solutions for PM<sub>10</sub> and NSD data datasets together. We also show how regression of the NSD size bins and the *G* time series can be used to elaborate the solution by identifying NSD factors (such as nucleation) not influencing the PM<sub>10</sub> mass.

## 1 Introduction

It is unquestionable that worldwide, the scientific vista of air quality is expanding, whether it is the increasing num-

ber of observatories or the refinement of information mined from the increasing sophistication of measurements often incorporated in campaign work. The number of metrics being measured has increased from simple measurements of particulate matter (PM) mass and gas concentrations, and we can now probe the composition of the PM mass and the size distributions with mass spectrometers, mobility analysers and optical devices.

Studies using positive matrix factorization (PMF) as a tool for source apportionment of particle mass using multicomponent chemical analysis data are published frequently using datasets from around the world. However, they do not always provide consistent outcomes (Pant and Harrison, 2012), and one means by which source resolution and identification can be improved is by inclusion of auxiliary data, such as gaseous pollutants (Thimmaiah et al., 2009), particle number count (Masiol et al., 2017) or particle size distribution (Beddows et al., 2015; Ogulei et al., 2006; Leoni et al., 2018).

Harrison et al. (2011) analysed number size distribution (NSD) data (merged Scanning Mobility Particle Sizer (SMPS) and Aerodynamic Particle Sizer (APS) data) with PMF using auxiliary data (meteorology, gas concentration, traffic counts and speed). The study used particle size distribution data collected at the Marylebone Road supersite in London in the autumn of 2007 and put forward a 10-factor solution comprised of roadside and background particle source factors. Sowlat et al. (2016) carried out a similar analysis on number size distribution (13 nm–10 μm) data combined with several auxiliary variables collected in Los

Angeles. These included black carbon (BC), elemental carbon (EC) and organic carbon (OC), PM mass, gaseous pollutants and meteorological and traffic flow data. A six-factor solution was chosen comprising of nucleation, two traffic factors, an urban background aerosol, a secondary aerosol and a soil factor. The two traffic sources contributed up to above 60 % of the total number concentrations combined. Nucleation was also observed as a major factor (17 %). Urban background aerosol, secondary aerosol and soil, with relative contributions of approximately 12 %, 2.1 % and 1.1 %, respectively, overall accounted for approximately 15 % of PM number concentrations, although these factors dominated the PM volume and mass concentrations, due mainly to their larger mode diameters. Chan et al. (2011) considered extracting more source information from an aerosol composition dataset by including data on other air pollutants and wind data in the analysis of a small but comprehensive dataset from a 24-hourly sampling programme carried out during June 2001 in an industrial area in Brisbane. They chose multiple types of composition data (aerosols, volatile organic compounds (VOCs) and major gaseous pollutants) and wind data in source apportionment of air pollutants and found it to result in better defined source factors and better fit diagnostics, compared to when non-combined data were used. Likewise, Wang et al. (2017) report an improvement in source profiles when coupling the PMF model with  $^{14}\text{C}$  data to constrain the PMF run as a priori information.

However, while combining, for example, particle chemical composition and size distribution data in a single PMF analysis may assist source resolution, difficulties arise if the two datasets have different and/or ambiguous rotations (discussed in Sect. 2). This tends to result in factors with either mass contributions and small number contributions or number contributions and small mass contributions and rarely a meaningful contribution from both data types. Experimental design can of course circumnavigate this problem, for instance, using chemical data which are already size-segregated, measured using a cascade impactor (Contini et al., 2014). Such an approach is attractive by view of the fact that there is no question as to whether both datasets sufficiently overlap across the size bins. However, cascade impactors do not offer the high time resolution of particle counting instruments, with individual measurements lasting hours or days. Even so, for the case in which two or more instruments are available in a campaign to measure two or more different metrics, e.g. PM mass and particle number (PN), then a combined data analysis is useful. Emami and Hopke (2017) have shown that the effect of adding variables as auxiliary data (with potentially different units) to a NSD dataset is to decrease the rotational ambiguity of a solution from a one-step PMF analysis.

In this study, we present a method for analysing simultaneously collected  $\text{PM}_{10}$  composition and NSD data. In the work of Beddows et al. (2015), both particle composition and NSD data from a background site in London (2011 and



**Figure 1.** Venn diagram showing the summary of the findings of Beddows et al. (2015), applying PMF to  $\text{PM}_{10}$ -only, NSD-only and  $\text{PM}_{10}$ –NSD datasets. Table shows the apportionment of  $\text{PM}_{10}$  and NSD taken from Beddows et al. (2015).

2012) were analysed using positive matrix factorization. As part of the methodology development, it was concluded that it was preferable not to combine these two data types in a single analysis but to conduct separate PMF analyses for  $\text{PM}_{10}$  mass and particle number. This yielded a six-factor solution for the  $\text{PM}_{10}$  data (diffuse urban, marine, secondary, non-exhaust traffic and crustal (NET and crustal), fuel oil and traffic). Factors described as diffuse urban, secondary and traffic were identified in the four-factor solution for the NSD data, together with a nucleation factor not seen in the  $\text{PM}_{10}$  mass data analysis (see Fig. 1). When combining the  $\text{PM}_{10}$  and NSD data in a single PMF analysis, diffuse urban, nucleation, secondary, aged marine and traffic factors were identified, but the factors were not as clearly separated from each other as the factors derived from the separate datasets. For example, fuel oil was now mixed in with marine and called aged marine. This is summarized in Fig. 1. However, it would still be useful to obtain a number size distribution for each of the six  $\text{PM}_{10}$  factors and/or a chemical composition for the four NSD factors. As a continuation of this work, we present an alternative method for analysing the combined dataset in a so-called two-step methodology. In the first step, we analyse the mass data ( $\text{PM}_{10}$ ; units:  $\mu\text{g m}^{-3}$ ) according to the methodology of Beddows et al. (2015). This results in a time series factor  $\mathbf{G}$ , which is carried forward into a second PMF analysis of a combined dataset consisting of the  $\mathbf{G}$  time series and an auxiliary dataset (i.e. NSD; units:  $\text{cm}^{-3}$ ). The first step identifies sources and apportions the  $\mathbf{G}$  factors to their contribution to mass, and in the second step, an  $\mathbf{FKEY}$  matrix is chosen such that  $\mathbf{G}$  “drives” the model and the NSD data “follow”. This means that we have  $\text{PM}_{10}$  factors, each of which is augmented by its number size distribution. Furthermore, we also consider linear regression (LR) as a second step in a PMF–LR analysis to show that although the initial analysis is biased toward mass by analysing  $\text{PM}_{10}$  factors only, unseen factors influencing the NSD data (e.g. nucleation) can be identified in the data.

## 2 Experiment

With a population of 8.5 million in 2014 (ONS, 2017), the UK city of London is the focus of study in this work; the North Kensington (NK) site (lat. = 51°: 31': 15.780" N and long. = 0°: 12': 48.571" W) was considered. NK is part of both the London Air Quality Network and the national Automatic Urban and Rural Network and is owned and part-funded by the Royal Borough of Kensington and Chelsea. The facility is located within a self-contained cabin within the grounds of Sion Manning School. The nearest road, St. Charles Square, is a quiet residential street approximately 5 m from the monitoring site, and the surrounding area is mainly residential. The nearest heavily trafficked roads are the B450 (~ 100 m east) and the very busy A40 (~ 400 m south). For a detailed overview of the air pollution climate at North Kensington, the reader is referred to Bigi and Harrison (2010).

### 2.1 Data

As alluded to, this work is a continuation of the study carried out by Beddows et al. (2015), which analysed NSD and PM<sub>10</sub> chemical composition data collected at the NK receptor site. Number size distribution (NSD) data were collected continuously every 15 min using a Scanning Mobility Particle Sizer (SMPS), consisting of a CPC (TSI model 3775) combined with an electrostatic classifier (TSI model 3080) and air-dried according to the EUSAAR protocol (Wiedensohler et al., 2012). The particle sizes covered were 51 size bins ranging from 16 to 604 nm, the 15 min distributions were aggregated up to hourly averages (when there were at least three 15 min samples per hour) and all missing values were replaced using a value calculated using the method of Polissar et al. (1998). Further details of the SMPS settings are given in Table S1 in the Supplement, and the reader is also referred to Beccaceci et al. (2013a, b) for an extensive account of how the NSD data were collected and quality-assured.

Accompanying the NSD data from the study of Beddows et al. (2015) was the PMF output from the analysis of PM<sub>10</sub> chemical composition data. The latter data consisted of 24 h air samples taken daily over a 2-year period (2011 and 2012) using a Thermo Partisol 2025 sampler fitted with a PM<sub>10</sub> size-selective inlet. These filters were analysed for total metals, PM<sub>metals</sub> (Al, Ba, Ca, Cd, Cr, Cu, Fe, K, Mg, Mo, Na, Ni, Pb, Sn, Sb, Sr, V and Zn), using a Perkin Elmer/Sciex ELAN 6100DRC following hydrofluoric acid digestion of GN-4 Metrical membrane filters. Water-soluble ions, PM<sub>ions</sub> (Ca<sup>2+</sup>, Mg<sup>2+</sup>, K, NH<sub>4</sub><sup>+</sup>, Cl<sup>-</sup>, NO<sub>3</sub><sup>-</sup> and SO<sub>4</sub><sup>2-</sup>), were measured using a near-real-time URG-9000B (hereafter URG) ambient ion monitor (URG Corp). The data capture over the 2 years ranged from 48 % to 100 % as different sampling instruments varied in reliability. Data gaps were filled by measurements made on daily PM<sub>10</sub> filter samples collected continuously at this site using a Partisol 2025; laboratory-

based ion chromatography measurements were made for anions on Tissuquartz™ 2500 QAT-UP filters. No cation measurements were available from these filters, and this resulted in a lower data capture for the cations. Again, all missing data were replaced using a value calculated using the method of Polissar et al. (1998). A woodsmoke metric, CWOD, was also included. This was derived as PM woodsmoke from the methodology of Sandradewi et al. (2008) utilizing aethalometer and EC/OC data, as described in Fuller et al. (2014). Samples were also collected using a Partisol 2025 with a PM<sub>10</sub> size-selective inlet, and concentrations of elemental carbon (EC) and organic carbon (OC) were measured through collection on quartz filters (Tissuquartz™ 2500 QAT-UP) and analysis using a Sunset Laboratory thermal–optical analyser according to the QUARTZ protocol (which gives results very similar to EUSAAR 2: Cavalli et al., 2010) (NPL, 2013). We refer to CWOD, EC and OC as PM<sub>carbon</sub>. In addition, particle mass was determined on samples collected on Teflon-coated glass fibre filters (TX40HI20WW) with a Partisol sampler and PM<sub>10</sub> size-selective inlet.

This aforementioned PM<sub>10</sub> data were represented in this work as the PMF solution for PM<sub>10</sub>-only data, derived in Beddows et al. (2015) and consisting of six sources, namely diffuse urban, marine, secondary, non-exhaust traffic and crustal, fuel oil and traffic. The diffuse urban factor had a chemical profile indicative of contributions mainly from both woodsmoke (CWOD) and road traffic (Ba, Cu, Fe, Zn). The marine factor explained much of the variation in the data for Na, Cl<sup>-</sup> and Mg<sup>2+</sup>, and the secondary factor was identified from a strong association with NH<sub>4</sub><sup>+</sup>, NO<sub>3</sub><sup>-</sup>, SO<sub>4</sub><sup>2-</sup> and organic carbon. For the traffic emissions, the PM did not simply reflect tailpipe emissions, as it also included contributions from non-exhaust sources, i.e. resuspension of road dust and primary PM emissions from brake, clutch and tyre wear. The non-exhaust traffic and crustal factor explained a high proportion of the variation in the Al, Ca<sup>2+</sup> and Ti measurements consistent with particles derived from crustal material, derived either from wind-blown or vehicle-induced resuspension. There was also a significant explanation of the variation in elements such as Zn, Pb, Mn, Fe, Cu and Ba, which had a strong association with non-exhaust traffic emissions. As there was a strong contribution of crustal material to particles resuspended from traffic, this likely reflected the presence of particulate matter from resuspension and traffic-polluted soils. The last factor was attributed to fuel oil, characterized by a strong association with V and Ni together with significant SO<sub>4</sub><sup>2-</sup>. This output comprised the first-step solution in the two-step analysis of PM<sub>10</sub> and NSD data, and in this study we concentrate on the analysis of the NSD data in the second PMF step, with the aim of assigning a NSD to each of the six PM<sub>10</sub> factors.

## 2.2 Methods

### 2.2.1 PMF

Positive matrix factorization (PMF) is a well-established multivariate data analysis method used in the field of aerosol science. PMF can be described as a least-squares formulation of factor analysis developed by Paatero (Paatero and Tapper, 1994). It assumes that the ambient aerosol concentration  $\mathbf{X}$  (represented by  $m \times n$  matrix of  $m$  observations and  $n$  PM<sub>10</sub> constituents or NSD size bins), measured at one or more sites, can be explained by the product of a source profile matrix  $\mathbf{F}$  and source contribution matrix  $\mathbf{G}$  whose elements are given by Eq. (1):

$$x_{ij} = \sum_{k=1}^p g_{ik} \cdot f_{kj} + e_{ij} \quad i = 1 \dots m; j = 1 \dots n, \quad (1)$$

where the  $j$ th PM constituent (element, size bin or auxiliary measurement) on the  $i$ th observation (i.e. hour) is represented by  $x_{ij}$ . The term  $g_{ik}$  is the contribution of the  $k$ th factor (of a total of  $p$  factors) to the receptor on the  $i$ th hour,  $f_{kj}$  is the fraction of the  $j$ th PM constituent in the  $k$ th factor and  $e_{ij}$  is the residual for the  $j$ th measurement on the  $i$ th hour. The residuals (i.e. difference between measured and reconstructed concentrations) are accounted for in matrix  $\mathbf{E}$ , and the two matrices  $\mathbf{G}$  and  $\mathbf{F}$  are obtained using an iterative algorithm which minimizes the object function  $Q$  (see Eq. 2).

Using the data and uncertainty matrices for the model, Eq. (1) is optimized in the PMF algorithm by minimizing the  $Q$  value (Eq. 2):

$$Q = \sum_{i=1}^n \sum_{j=1}^m \left( \frac{e_{ij}}{s_{ij}} \right)^2, \quad (2)$$

where  $s_{ij}$  is the uncertainty in the  $j$ th measurement for hour  $i$ . All analyses were carried out in robust mode which reduces the impact of outliers (Paatero, 2002).

PMF is a weighted technique, and the value of  $Q$ , and hence the model fit, is determined by the input variables with the lowest values of uncertainty,  $s_{ij}$ , thus giving their variables a higher weighting in the analysis. Input variables with low weight have little effect upon the value of  $Q$ , even when their residuals are large. This can be used to the advantage of the operator; e.g. when apportioning total PM mass in a conventional one-step PMF, the total PM concentrations are normally input with artificially high uncertainty, so that they are essentially passive in the PMF analysis and do not influence its outcome. By doing so, the chemical composition data determine the apportionment of PM mass to the source-related factors identified by the PMF. A similar approach can be followed in the PMF analysis of a combined dataset, whereby higher weightings can be applied to the main dataset of interest such that it drives the analysis and the auxiliary dataset follows; i.e. the uncertainties are chosen such that the balance of total weights from the two datasets is *tipped* towards

the measurement of interest and highest reliability in regards of rotational unambiguity.

To assess the PMF model, the  $Q$  value is outputted by PMF and compared to a theoretical value  $Q_{\text{theory}}$ , which is approximately the difference between the product of the dimensions of  $\mathbf{X}$  and the product of the number of factors and the sum of dimensions of  $\mathbf{X}$  (i.e.  $m \times n - p(m+n)$ ). For a given number of factors, the whole uncertainty matrix is scaled by a factor  $b_{\text{scale}}$  until the ratio between  $Q$  and  $Q_{\text{theory}}$  is approximately 1 ( $rQ$  value =  $Q/Q_{\text{theory}} = 1 \pm 0.02$ ).

With regards to the final output from PMF, a scaling has to be applied in order to achieve quantitative results. This is done by scaling either  $\mathbf{G}$  or  $\mathbf{F}$  to unity such that the units from  $\mathbf{X}$  are carried over to either  $\mathbf{F}$  or  $\mathbf{G}$  respectively to complete the apportionment. However, different routes have to be considered depending on whether  $\mathbf{X}$  has homogeneous or heterogeneous units.

### 2.2.2 One-step method using data in the same units – homogeneous units

Given a PMF input data matrix  $\mathbf{X}$ , a solution  $\mathbf{GF} + \mathbf{E}$  can be computed, where  $\mathbf{G}$  represents the time series of the source profiles  $\mathbf{F}$ , with a residual matrix  $\mathbf{E}$ . Often  $\mathbf{X}$  comprises columns of PM<sub>10</sub> component concentrations (e.g. ICPMS values measured from acid-digested filters collected with a Partisol 2025), and it is common practice to also include a total variable (e.g. column of PM<sub>10</sub>, measured using a Tapered Element Oscillating Microbalance, TEOM) in the data matrix. The resulting PM<sub>10</sub> profile element value can then be used to scale  $\mathbf{G}$  and  $\mathbf{F}$  such that  $\mathbf{G}$  carries the units of  $\mathbf{X}$ , with  $\mathbf{F}$  unitless. Note that neither  $\mathbf{G}$  or  $\mathbf{F}$  is scaled to unity in this approach. Instead, scaling is done after the analysis using a constant  $a_k$ , determined by the time series of a total variable (e.g. PM<sub>10</sub>), downweighted by applying a high uncertainty, within the input data.

$$x_{ij} = \sum_{k=1}^p (a_k g_{ik}) \left( \frac{f_{kj}}{a_k} \right) \quad (3)$$

The resulting value for the PM<sub>10</sub> contribution for each factor within the  $\mathbf{F}$  matrix is then used as a scaling constant  $a_k$  in Eq. (3). Such scaling results in unitless factors  $\mathbf{F}$  which describe the characteristics of the sources and time series  $\mathbf{G}$  with units of  $\mu\text{g m}^{-3}$ . Apportionment can then be carried out by averaging the  $\mathbf{G}$  values for each source factor, or a fully quantified time series of each factor can be presented, e.g. in bivariate plots. Of course, the  $\mathbf{G}$  and  $\mathbf{F}$  can be normalized such that  $\mathbf{G}$  is unitless and  $\mathbf{F}$  carries units, an approach necessary when  $\mathbf{X}$  contains heterogeneous units. This approach, using the PMF, requires the average of each column of  $\mathbf{G}$  to be scaled to unity, by using the PMF setting  $\text{mean}|\mathbf{G}| = 1$ .

### 2.2.3 One-step method using data with different units – heterogeneous units

If the analysis of  $\mathbf{X}$  was to be enhanced by the inclusion of data  $\mathbf{Z}$  from a second instrument with different units, then a different approach to the one-step method with homogeneous units would be required to analyse the joint data matrix  $[\mathbf{X}, \mathbf{Z}] = \mathbf{G}[\mathbf{X}, \mathbf{Z}] \cdot \mathbf{F}[\mathbf{X}, \mathbf{Z}] + \mathbf{E}[\mathbf{X}, \mathbf{Z}]$ . If the previous method was applied where  $\mathbf{F}$  was normalized, then it would not be clear what units to assign to  $\mathbf{G}$ , whether the units from  $\mathbf{X}$  or  $\mathbf{Z}$ . To get around this problem,  $\mathbf{G}$  is scaled to unity. This results in a unitless time series  $\mathbf{G}$  and a quantified  $\mathbf{F}$  matrix. For each source profile the sum of the species associated with either data type gives the average total apportionment, e.g. of  $\text{PM}_{10}$  or number concentration PN. Of course, this requires the complete mass or number closure of the elements making up either  $\text{PM}_{10}$  or PN respectively, although inclusion of measurements of total  $\text{PM}_{10}$  or PN can be used instead, if available.

In the ideal case, if the individually computed factors for both datasets result in  $\mathbf{G}(\mathbf{X})$  and  $\mathbf{G}(\mathbf{Z})$  being identical, then a straightforward joint model  $[\mathbf{X}, \mathbf{Z}]$  is successful and  $\mathbf{G}[\mathbf{X}, \mathbf{Z}] = \mathbf{G}(\mathbf{X}) = \mathbf{G}(\mathbf{Z})$ . However, if  $\mathbf{G}(\mathbf{X})$  and  $\mathbf{G}(\mathbf{Z})$  are significantly different, then the joint model will fail, identified by too large a  $Q$  value. A solution to this problem is to set the total weights of the better dataset  $\mathbf{X}$  significantly higher than the total weights of the auxiliary dataset  $\mathbf{Z}$  such that  $\mathbf{X}$  will drive the model and  $\mathbf{G}[\mathbf{X}, \mathbf{Z}]$  will be approximately equal to  $\mathbf{G}(\mathbf{X})$ , and a reasonable  $Q$  value is obtained for the  $\mathbf{Z}$ . However, care is required to ensure that  $\mathbf{X}$  or  $\mathbf{Z}$  do not contain rotational ambiguity because such rotation for  $\mathbf{X}$  may not be suitable for  $\mathbf{Z}$ . For such cases, equal total weights for both  $\mathbf{X}$  and  $\mathbf{Z}$  are applied in the hope that the best rotation for both  $\mathbf{X}$  and  $\mathbf{Z}$  can be found.

### 2.2.4 Two-step method using data with different units – heterogeneous units

The method proposed in this work separates the analysis of the two datasets  $\mathbf{X}$  and  $\mathbf{Z}$  into two different PMF analyses. Dataset  $\mathbf{X}$  is first analysed, and an unambiguous rotation is selected, which gives computed factors  $\mathbf{G}(\mathbf{X})$ . These are then carried over into a second PMF step in which  $\mathbf{G}(\mathbf{X})$  are combined with  $\mathbf{Z}$  to form a joint matrix for analysis. By using **FKEY** (described below) factors,  $\mathbf{G}(\mathbf{X}, \mathbf{Z})$  are forced to be equal to  $\mathbf{G}(\mathbf{X})$  from step 1. So, for example, if in the first step we analyse  $\text{PM}_{10}$  data and carry forward the output  $\mathbf{G}(\text{PM}_{10})$  into a second step combined with the NSD data, i.e.  $[\mathbf{G}(\text{PM}_{10}), \text{NSD}]$ , this results in profiles  $\mathbf{F}[\mathbf{G}(\text{PM}_{10}), \text{NSD}]$ . In other words, we force out of the NSD data source profiles which have the same  $\mathbf{G}$  factors as the  $\text{PM}_{10}$  data and extend the list of components of the sources identified in the first step, thus improving characterization of the source. Note that this is equivalent to non-negative weighted regression of matrix  $\mathbf{Z}$  by columns of matrix  $\mathbf{G}$  for which other tools exist.

Furthermore, by using a two-step method, we can continue to use the scaling method described in Sect. 2.2.2 to apportion the sources using a quantified time series  $\mathbf{G}(\mathbf{X})$  rather than normalizing the  $\mathbf{G}(\mathbf{X}, \mathbf{Z})$  matrix sums to 1 and relying on the summation of the elements in the rows of  $\mathbf{F}(\mathbf{X}, \mathbf{Z})$  to give the apportionment of  $\mathbf{X}$  and  $\mathbf{Z}$ .

### 2.2.5 Application of PMF

Positive matrix factorization was carried out in this work using the DOS-based executable file PMF2 v4.2 compiled by Pentti Paatero and released on 11 February 2010 (available from Dr. Pentti Paatero). This is used by the author in preference to a GUI version of PMF (e.g. US EPA PMF 5.0; Norris et al., 2014) because of the ease with which it can be incorporated into a Cran R procedure script using shell commands, thus facilitating automation of the analysis and any optimization. R script can be written to manipulate and organize input data for PMF2, run PMF2, collect the output and produce the necessary output for consideration as text, table or plot. The main strength of this approach is to improve the repeatability and transference of a method between practitioners within our group.

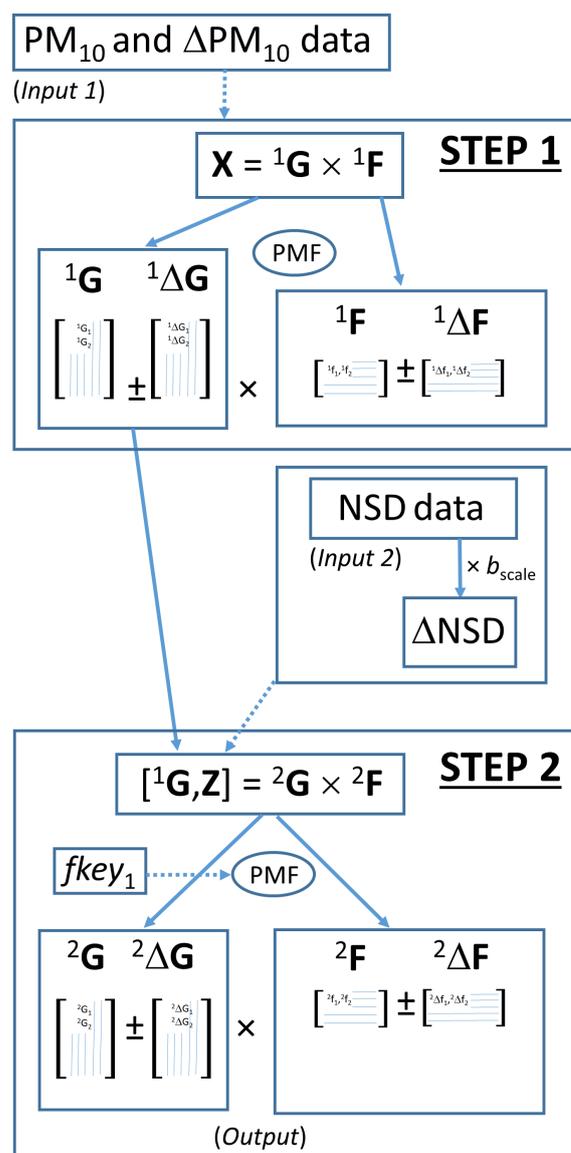
The two-step method is shown schematically in Fig. 2. Matrix  $\mathbf{X}$  yields factors  ${}^1\mathbf{G}$  and  ${}^1\mathbf{F}$  in the first step. The time series  ${}^1\mathbf{G}$  matrix is carried through to the second step, where it is combined with an auxiliary dataset  $\mathbf{Z}$  to give the step 2 input matrix  $[{}^1\mathbf{G}, \mathbf{Z}]$ . This in turn is analysed to produce factors  ${}^2\mathbf{G}$  and  ${}^2\mathbf{F}$ . In the current example, the dataset of Beddows et al. (2015) is used as a starting matrix  $\mathbf{X}$  and comprises the  $\text{PM}_{10}$  chemical composition dataset. This yields time series  ${}^1\mathbf{G}$  and source profile  ${}^1\mathbf{F}$ , and the reader is referred to Beddows et al. (2015) for a description of the analysis and output. Figure 1 shows the output from the first step which was found to be the optimum solution after considering three- to eight-factor solutions. The scaled time series matrix  ${}^1\mathbf{G}$  from this analysis was combined with the NSD data – concurrently measured with the  $\text{PM}_{10}$  data – to form the input matrix  $[{}^1\mathbf{G}, \mathbf{Z}]$  for step 2. The uncertainties of the  ${}^1\mathbf{G}_1$  matrix,  ${}^1\Delta\mathbf{G}$ , are transferred from the output of the first step and entered as input uncertainties for the second step. The hourly NSD data were aggregated into daily values to match the daily  ${}^1\mathbf{G}$  factors outputted from the PMF analysis of the daily  $\text{PM}_{10}$  data sampled. This reduced the data matrix down to 590 rows by 57 columns ( ${}^1\mathbf{G}_1 \dots {}^1\mathbf{G}_6$ ,  $\text{NSD}_1^{16\text{nm}} \dots \text{NSD}_{51}^{640\text{nm}}$ ) for which we have a  $Q_{\text{theory}}$  value of 29 748 for a six-factor solution. For the NSD data, the uncertainties are taken as the NSD values multiplied by the value of an arbitrary parameter  $b_{\text{scale}}$  (see Fig. 2). Initially,  $b_{\text{scale}}$  was set to 4 to ensure that the model was weighted such that it was driven by the  $\text{PM}_{10}$  data. However, this operation becomes somewhat redundant by the use of the **FKEY** matrix discussed in the next section. However, in order to find the optimal NSD uncertainties the value of the parameter  $b_{\text{scale}}$  (typically 0.2) was optimized in Cran R so that

the ratio of  $Q/Q_{\text{theory}} = 1 \pm 0.02$ , indicating a relative percentage uncertainty in the region of 20%. In retrospect – by taking into account the decrease in reliability of the size bin counts towards the edges of the size bin range – an improvement would be to gradually increase the uncertainties from 5% in the middle range of sizes to a predefined larger value, e.g. 50%, over the lower and upper size bins. The uncertainties were entered directly into the model using the PMF matrix  $\mathbf{T}$ , with  $\mathbf{U}$  and  $\mathbf{V}$  redundant.

### 2.2.6 Pulling down with GKEY and FKEY

**GKEY** and **FKEY** are matrices with the same dimensions as  $\mathbf{G}$  and  $\mathbf{F}$  respectively, for incorporating a priori information into a PMF analysis. They are used in the second step of the PMF analysis to “pull” elements of the source profiles to zero. **GKEY** and **FKEY** indicate the location of suspected zeros in source profiles  ${}^2\mathbf{F}$  or contributions  ${}^2\mathbf{G}$  (Fig. S1). Since we are concerned with the profiles, this information is given in the form of integer values in **FKEY**. The greater the certainty that an element of a source profile is zero, the larger the integer value that is specified. In this case, in the second step for the input dataset  $[{}^1\mathbf{G}, \mathbf{NSD}]$ , it is certain that only one unique contribution will be strong for each row of the profile  ${}^2\mathbf{F}$ , outputted from the second PMF analysis; e.g. only  ${}^1\mathbf{G}_1$  and not  ${}^1\mathbf{G}_2 \dots {}^1\mathbf{G}_6$  will contribute to the first position in output factor  ${}^2\mathbf{F}_1$  (Fig. S1). All “non-zero” elements within the output of  ${}^2\mathbf{F}$  take a *fkey* value of zero, whereas all elements of  ${}^2\mathbf{F}$  which are pulled to zero take a non-zero value of *fkey*<sub>1</sub>. This leads to a **FKEY** matrix which can be understood in two parts. The first part is a square matrix of dimensions equal to the number of columns of  ${}^1\mathbf{G}$ , with all its entries equal to *fkey*<sub>1</sub> except for the leading diagonal (set to zero); this part ensures that  ${}^1\mathbf{G}$  is the same as  ${}^2\mathbf{G}$ . The second part of the matrix consists of all the elements as zero and represents the NSD input data. An *fkey*<sub>1</sub> value of 7 to 9 is considered a medium to strong pull, and in this work, we used a value of 24, which in comparison is very aggressive ensuring only one rotational solution is available ensuring  ${}^1\mathbf{G} \approx {}^2\mathbf{G}$ .

To extend the analysis from six factors to seven factors, an extra row was added to **FKEY**. This was done in order to investigate any factors missed in the NSD data which the first analysis using  $\text{PM}_{10}$  would not be sensitive to. For example, a nucleation mode would be detected in NSD data but not in  $\text{PM}_{10}$  data. In order to give the model freedom to factorize out a nucleation factor, the seventh row of **FKEY** values consisted of  $\{fkey_2, \dots, fkey_2, nsd_1, nsd_2, \dots, nsd_{51}\}$ ;  $fkey_2 = 20$ . This ensured that all the  ${}^2\mathbf{G}$  contributions were allocated to the first six factors, only leaving the seventh factor to account for the remaining unfactorized NSD data. There is no reason why more than seven factors could not be used to investigate possible unresolved NSD factors. However, we constrained the scope of our investigation to reidentifying those in Fig. 1.



**Figure 2.** Flow diagram showing the flow of data through the two-step PMF–PMF analysis. The PMF analyses of a single dataset  $\mathbf{X}$  are considered in step 1, and output is indicated by factors and uncertainties  ${}^1\mathbf{G}$ ,  ${}^1\Delta\mathbf{G}$ ,  ${}^1\mathbf{F}$  and  ${}^1\Delta\mathbf{F}$ . The second PMF analysis is carried out on the joint dataset  $[{}^1\mathbf{G}, \mathbf{Z}]$  and yields factors and uncertainties  ${}^2\mathbf{G}$ ,  ${}^2\Delta\mathbf{G}$ ,  ${}^2\mathbf{F}$  and  ${}^2\Delta\mathbf{F}$ . In our analysis,  $\mathbf{X}$  and  ${}^1\mathbf{G}$  are the  $\text{PM}_{10}$  and resulting time series from the analysis of Beddows et al. (2015), and  $\mathbf{Z}$  is the auxiliary NSD data concurrently measured using a SMPS.

### 2.3 Regression

As an alternative to using PMF in the second step, a regression was carried out. Each column of data for each of the 51 size bins  $j$  within the NSD was regressed against the six  ${}^1\mathbf{G}$  time series using Eq. (4):

$$\text{NSD}_j = \alpha_{0,j} + \alpha_{1,j} {}^1\mathbf{G}_1 + \alpha_{2,j} {}^1\mathbf{G}_2 + \dots + \alpha_{6,j} {}^1\mathbf{G}_6, \quad (4)$$

where  $\alpha_0$  is the population intercept and  $\alpha_{1-6}$  are the population slope coefficients. This results in a 7 by 51 matrix of values. Each column represents a size bin of the NSD data, and each row represents the slope coefficients associated with six of the factors (giving an indication of how each size bin scales with each of the six factors) and an intercept. When  $\alpha_{1-6,j}$  is plotted against the size bin, six plots showing the dependence of each size bin  $j$  on each of the six  $\text{PM}_{10}$  factors are produced. It is also assumed that these (referred to here as NSD regression source profiles) will be comparable to the actual NSD PMF source profile. Similarly, the  $\alpha_{0,j}$  values are expected to give a background value due possibly to noise; however, it is more likely to yield a source (such as nucleation) to which the  $\text{PM}_{10}$  mass analysis is insensitive.

## 2.4 Peak fitting

If it is assumed that the factors derived from the daily NSD data are the same as those present in the hourly data, i.e. the factors are conserved when averaging the data from hourly to daily data before PMF analysis, then daily NSD profiles can be fitted to the hourly NSD spectra to recover a diurnal cycle for the factors. However, it is worth noting that the process of aggregating hourly data to daily NSD data may cause loss of information, implying that minor factors (e.g. due to event episodes) might well be averaged out of the data. For the elements of the  $i$ th number size distribution  $\text{NSD}_{ij}$  (of dimensions  $m \times n$ ), the factors can be fitted using Eq. (5), which is the difference across the size bins of the  $i$ th row of the NSD data and the linear sum of the  $p$  NSD source profiles ( $p = 7$  in this case) scaled with respect to the scalar values  $c_{ik}$ , representing the time series of each fitted NSD source profile.

$$d_i = \begin{cases} \sum_{j=1}^N \{ \text{NSD}_{ij} - \sum_{k=0}^p c_{ik} \times f_{kj} \}, & c_{ik} \geq 0 \\ 1 \times 10^{10}, & c_{ik} < 0 \end{cases} \quad (5)$$

The Cran R package non-linear minimization (nlm) (R Core Team, 2018) was used to minimize the value of  $d_i$  with respect to the scalar value  $c_{ik}$  with a non-negative constraint on  $c_{ik}$  placed in the function. If a negative value is returned by any of the  $c_k$  values, then  $d_i$  returns an excessively large value. Furthermore, in order to extract an apportionment to number concentration ( $\text{cm}^{-3}$ ) the fitted values were scaled using a scalar  $\beta_k$ . Seven values were derived for  $\beta_k$  by regressing the total particle number (total hourly SMPS) against each of the fitted values  $c_k$  (Eq. 6).

$$PN = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \dots + \beta_7 c_7 \quad (6)$$

The resulting scaled-fitted values were then used to calculate the PN concentration for each of the regression source profiles (Eq. 7), allowing subsequent plotting of the seven diurnal cycles.

$$PN_k = \beta_k c_k \quad (7)$$

## 2.5 Bivariate plot

Identification of the sources responsible for the factors outputted from PMF can be assisted by meteorological data. Time series of the  $k$ th factor (or  $g_k$  values) can be plotted against wind direction and wind speed using either the polarPlot or polarAnnulus functions provided in the Openair package. Polar plots are simply used for plotting the factor contribution on a polar coordinate plot with north, east, south and west axes. Mean concentrations are calculated for wind speed direction “bins” (e.g. 0–1, 1–2  $\text{m s}^{-1}$ , ... and 0–10, 10–20°, etc.) and smoothed using a generalized additive model. Each bin concentration is plotted as a group of pixels (coloured according to a concentration colour scale) and positioned a distance away from the origin according to the magnitude of wind speed and along an angle from the north axis according to the wind direction. Such plots are useful when identifying the nature of the source. A diffuse source will tend to have its highest concentration showing as a “hotspot” at the origin of the polar plot, whereas a point source will cause a hotspot both away from the origin and in the direction pointing towards the source. On the other hand wind blown sources tend to be recognized by their relation to wind speed and hence do not necessarily produce hotspots. Instead, they produce a minimum to maximum gradual gradient of colour from the origin, spreading radially out towards the edge of the plot in the direction of the source, e.g. for a marine source. Likewise, annulus plots plot the mean factor concentration on a colour scale by wind direction and as a function of hour of the day as an annulus, represented by the distance of the coloured pixels from the origin. The function is good for visualizing how concentrations of pollutants vary by wind direction and hour of the day. For example, for the North Kensington site – positioned west of the city centre – we might well expect most of the anthropogenic sources (traffic, diffuse urban, etc.) to show an easterly direction with the appropriate diurnal cycle (e.g. rush hour traffic patterns). Similarly, we might expect cleaner air (marine, nucleation, etc.) to occur from a westerly direction and at times of the day when the solar strength is highest.

## 3 Results and discussion

The aim of this work has been to show how a given PMF result can be complemented with concurrently measured auxiliary data. We exemplify this using  $\text{PM}_{10}$  and NSD data collected from the North Kensington receptor site in London and start with the premise that we are completely satisfied with the  $\text{PM}_{10}$  analysis and are using a rotation which gives quantified factors (quantified **G** and scaled **F**) which best represent the urban atmosphere sampled, i.e. the output from Beddows et al. (2015). For each  $\text{PM}_{10}$  factor we wish to assign a NSD distribution. Rather than repeat the PMF analysis using a combined  $\text{PM}_{10}$ –NSD dataset which can be complicated

if the rotations of the individual PMF analyses of PM<sub>10</sub> and NSD data are mismatched or ambiguous, we can carry out a second PMF analysis or a regression.

Furthermore, because of the nature of any factor analysis, we also have to make the assumption that each source chemical profile and size distribution not only remains unchanged between source and receptor but that they remain constant throughout the measurement campaign. This of course limits our capacity to fully understand the aerosol within the atmosphere we are considering. Chemical reactions during the transit of the air masses will of course modify the chemical composition. It might be assumed that a fully aged aerosol remains unchanged and is identified as a background component, but, for example, we would expect progressive chlorine depletion within a fresh marine aerosol passing over a city. Likewise, we also have to appreciate that different particle sizes will have different atmospheric transit efficiencies, with large particles settling out of the air mass before smaller ones. Similarly, particles nucleate and grow from 1 nm up to 20–30 nm over a short time period of time. It is these finer details which are missed when making an overall assessment of the chemical and physical composition of air mass measured over a long (e.g. 2 years) dataset using PMF.

### 3.1 Two-step PMF–PMF analysis

Figure 3 presents the profiles  $^1F_k$  and  $^2F_k$  from the first and second PMF analysis respectively. The plots of  $^1F_k$  were carried over from Beddows et al. (2015) to complete the assignment of the source profiles.

The time series  $^1G_k$  and uncertainties  $^1\Delta G_k$  from the first PMF analysis of PM<sub>10</sub> data were carried over into the second step, where they are combined with the NSD data for PMF analysis (Fig. 2). The uncertainties of the NSD data are taken as an optimized multiple of the NSD values themselves ( $\sim 5\%$  uncertainty, yielding a  $Q$  value of 30 333 in the robust mode; see Table S2 for PMF settings). Also in order to encourage  $^2G_k$  to be proportional to  $^1G_k$  for  $k = 1\text{--}6$  (see Table S4), the **FKEY** matrix is applied to pull elements in the source matrix to zero, as described in Sect. 2.2.6. This ensured that the PMF analysis of the NSD data was driven by the  $^1G$  time series and resulted in a six-factor output in which there were unique contributions from the  $k$ th factor  $^1G_k$  from the first analysis to the  $k$ th factor  $^2F_k$  in the second analysis. This is mainly due to the aggressive pulling of the factor element in  $^2F$  applied using **FKEY**.

When inspecting Fig. 3 it is notable that the source profiles are surprisingly similar to those calculated for the NSD-only and PM<sub>10</sub>–NSD data in Beddows et al. (2015). The diffuse urban factor has a modal diameter just below 0.1  $\mu\text{m}$ , which is comparable to the same factor in the NSD-only analysis. The marine factor is comparable to the aged marine factor derived from the PM<sub>10</sub>–NSD analysis. The secondary factor is again the factor with the largest modal diameter (between 0.4 and 0.5  $\mu\text{m}$ ), and traffic has as expected a modal diameter

between 30 and 40 nm. The fuel oil factor appears to be a combination of a nucleation factor and a mode comparable to diesel exhaust seen in the traffic factor.

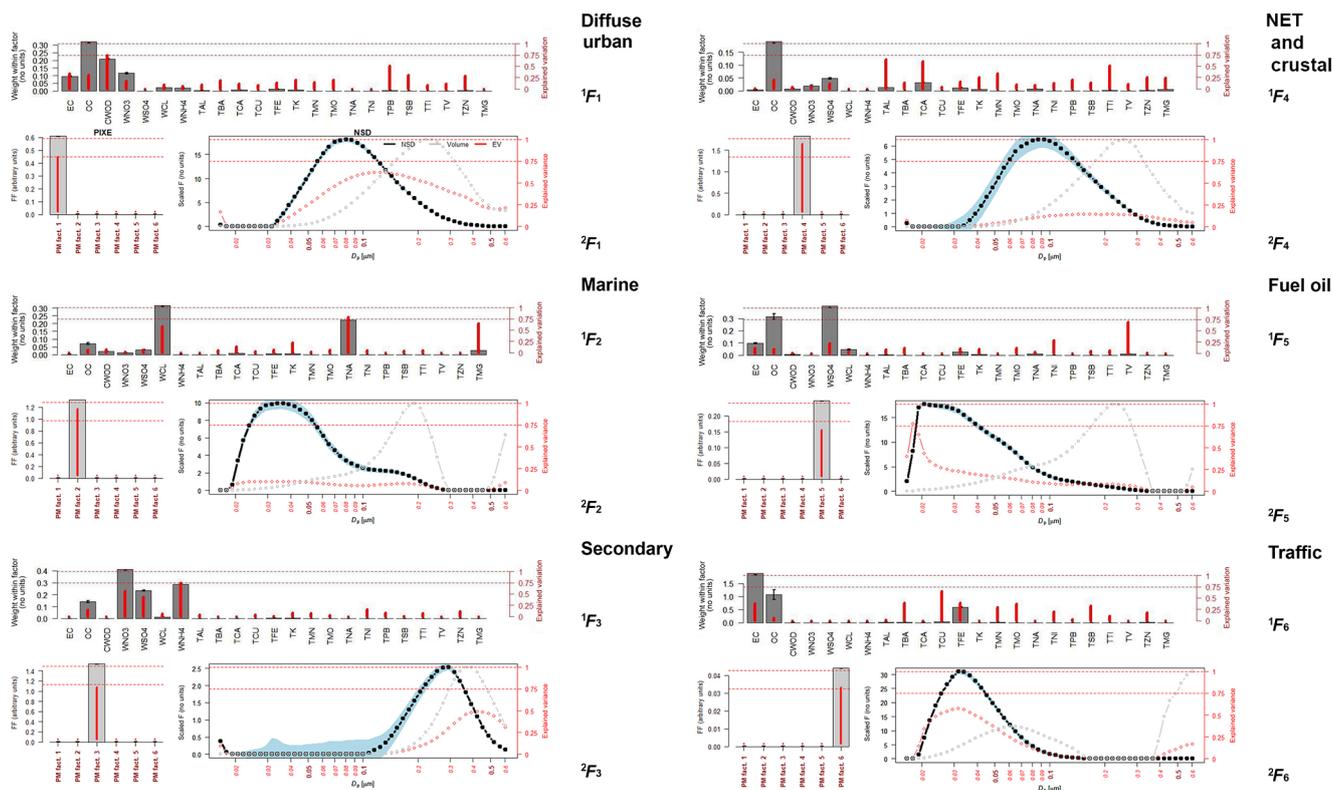
### 3.2 Two-step PMF–LR analysis

Figure S2 shows the results of the linear regression of the NSD data and the PM<sub>10</sub>  $^1G_k$  scores, and again what is remarkable is the similarity between these regression source profiles and both the factors derived in Beddows et al. (2015) and those from the two-step PMF–PMF analysis.

This PMF–LR analysis was carried out using daily averaged data, and to obtain hourly information – and thus obtain the diurnal patterns (Fig. S2) – the resulting regression source profiles were refitted to the original NSD data. On inspection of these source profiles and diurnal plots, the negative values make interpretation a struggle, reinforcing one of the four conditions (Hopke, 1991) in the analysis if it is to make sense. We can however fit non-negative gradients using non-negative regression. However, the surprising consequence of applying this constraint is that the same profiles are derived, but they are clipped so that all negative values are replaced by zero values – hence, information is lost. One interpretation of the negative values is that these are particle sinks, but this contradicts the PMF–PMF findings, and hence it is concluded that the PMF–LR analysis only serves as an indication of how the PM<sub>10</sub> factors are augmented by the NSD data. If all profiles are shifted to above the zero line, then comparisons to the PMF–PMF data can be made. However, what is interesting to note in this result is the intercept NSD, which is comparable in profile and diurnal pattern to the nucleation mode identified in Beddows et al. (2015). This is a seventh regression source profile, in addition to the six PM<sub>10</sub> factors and suggests that although the PMF analysis of the PM<sub>10</sub> data alone misses a nucleation factor, this can be recovered in a second analysis as a remainder or bias in the data. Furthermore, this result indicates that the composition of the nucleation NSD factor has no link to the chemical PM<sub>10</sub> composition and cannot be used to infer a composition. This is unsurprising given the very small mass contributed by the nucleation-mode particles.

Returning to the PMF–PMF analysis and extending the analysis from six factors to seven factors, an extra row in the **FKEY** matrix was added to pull all of the  $^1G_7$  contributions to  $^2F_7$  to zero in the solution (Fig. S1). The same **FKEY** matrix of  $fkey_1$  and 0 values was used, but this time it was augmented with a seventh row of  $fkey_2$  and zero values. In this case, the  $fkey_2$  values were set to a value of 20.

The same six-factor solution is obtained with the additional seventh factor (Figs. 4 and S3), and, as expected, this seventh factor was a nucleation factor. It was suspected that in the six-factor solution, the nucleation factor was combined with the fuel oil factor. This does not suggest any link between the nucleation and fuel oil factor other than that there was an insufficient number of factors within the model for



**Figure 3.** Source profiles  ${}^1F$  and  ${}^2F$  from both the first and second PMF step using six factors. (Grey bars and black line indicate the values of  $F$ ; red lines and dots indicate the explained variations; and the grey dotted line indicates the  $dV/d\log D_p$  values.)

the two to factorize out of the data, giving the fuel oil NSD profile a more reasonable modal peak between 50 and 60 nm rather than 20, 30 and 60 nm.

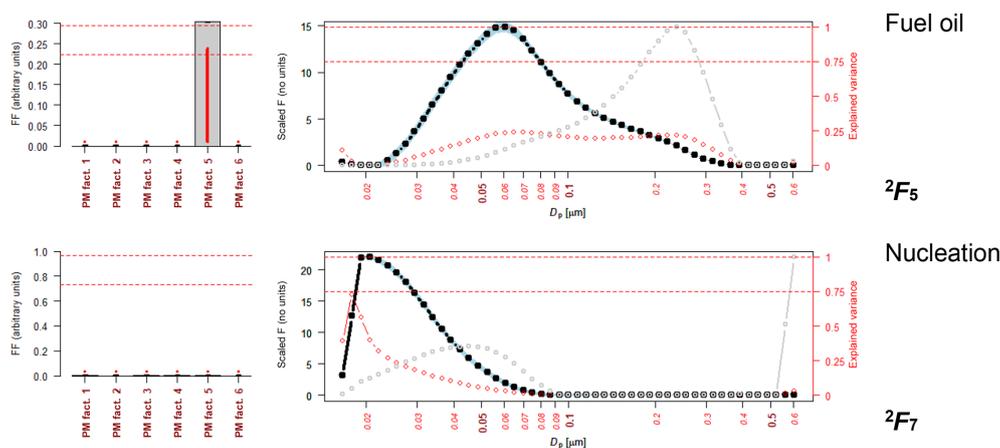
Beddows et al. (2015) applied a one-step analysis to three different datasets:  $PM_{10}$ -only; NSD-only and  $PM_{10}$ -NSD. The analyses of the  $PM_{10}$ -only and NSD-only – both with homogeneous units – produced quantitative time series  $G$ . This was unlike the analysis of the  $PM_{10}$ -NSD with heterogeneous units, which could not apportion its five factors using  $G$  but was able to factorize out a nucleation factor from the data, seen also in the four sources in the PMF solution for the NSD-only data. A  $PM_{10}$ -only seven-factor solution did not reveal this factor, presumably because the mass associated with nucleation-mode particles is too small to affect composition significantly. Furthermore, fuel oil was not factorized out of the  $PM_{10}$ -NSD data and was more likely divided across all five factors.

Another interesting observation is that although only four factors were derived from the PMF analysis of NSD-alone (diffuse urban; secondary; traffic and nucleation), when extra information is included from the PMF analysis of the  $PM_{10}$  data, more information can be extracted from the PMF analysis of the NSD data in the form of the marine, fuel oil and NET and crustal factors. The nucleation factor is only revealed when performing a regression between the NSD size

bins and the  $G$  scores of the  $PM_{10}$  PMF analysis, which leads to increasing the factor number from six to seven and in turn yields the nucleation profile. It is also reassuring that the bivariate plots for the seven factors (discussed in the next section) correspond to the bivariate plots given in Beddows et al. (2015). Also note that there is no reason why any further investigation might not explore this using more than seven factors. In fact the nucleation factor appears at first sight to be multimodal. However, we restricted our analysis to seven factors, considering it complete in terms of identifying the sources obtained by Beddows et al. (2015).

### 3.3 Diurnal and bivariate plots

The original PMF was carried out on daily  $PM_{10}$  data, and in order to make diurnal and bivariate plots, a higher time resolution is desirable. It is assumed that the factors derived in the hourly NSD data are the same as those derived from the daily averaged data; i.e. the factors are conserved when averaging the data from hourly to daily data before PMF analysis. Then the hourly NSD data can be fit with the PMF profiles derived from the daily data (see Sect. 2.4). Figure 5 shows the resulting diurnal profiles. The diurnal trends of the parameter  $c_k$  (Eq. 6), required to fit the seven daily NSD factors to the hourly NSD data, are shown. These have been scaled to PN (measured in  $cm^{-3}$ ) using the integral of the NSD (Eq. 7).



**Figure 4.** Nucleation and fuel oil factors derived when extending the second PMF analysis from the six factors (shown in Fig. 3) to seven factors. Source profiles  ${}^2F_1$  to  ${}^2F_6$  are given in Fig. S3. Each plot shows the output  ${}^2F_k$ . (Grey bars and black line indicate the values of  $F$ ; red lines and dots indicate the explained variations; and the grey dotted line indicates the  $dV/d\log D_p$  values.)

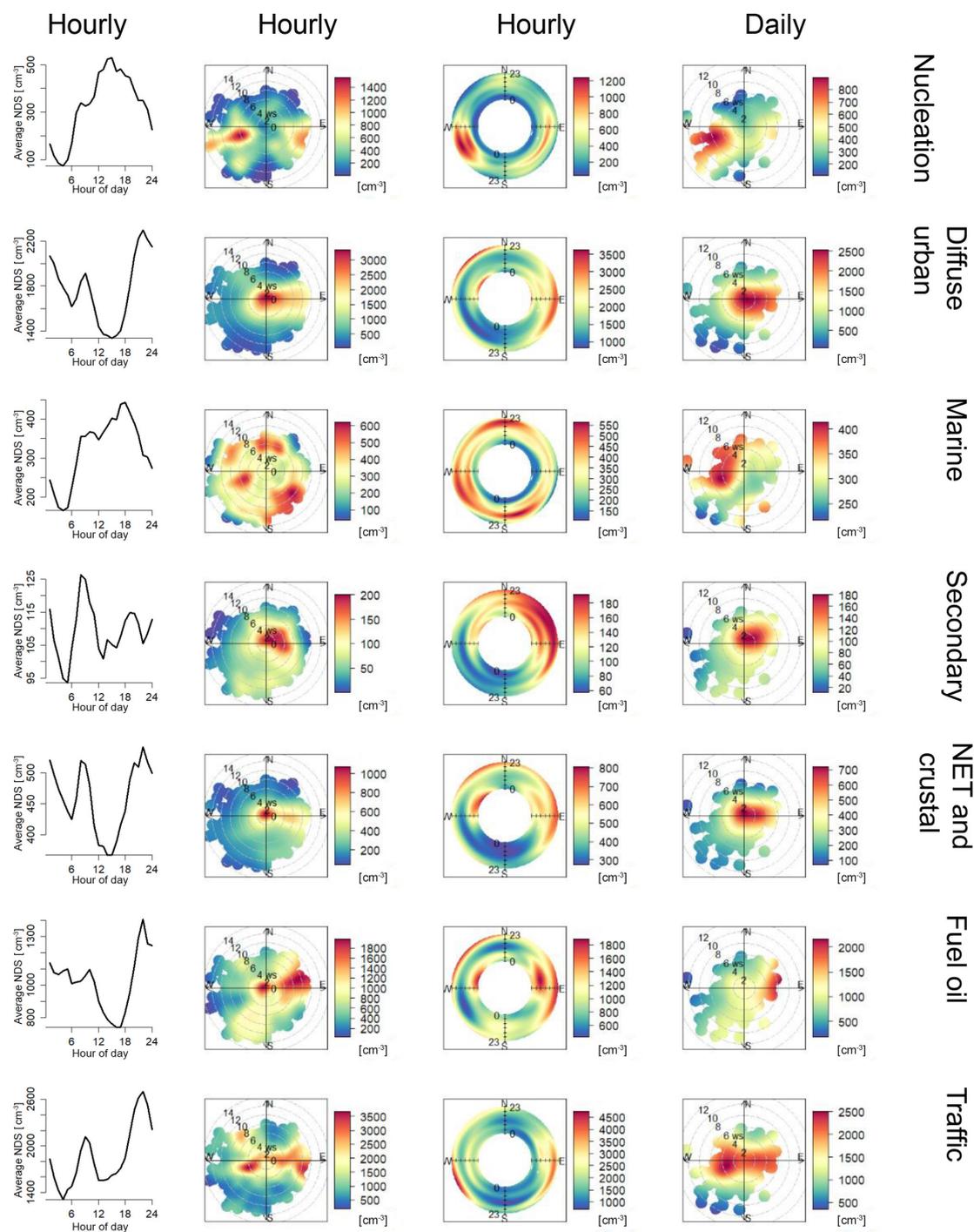
The nucleation factor diurnal trend behaves as expected, rising to a maximum during the day and then falling back down to a minimum at night. This corresponds to the intensity of the sun during the day and the increased likelihood of nucleation on clean days when there is sufficient precursor material to form particles with a low particle condensation sink. The marine factor is also high during the day, presumably due to higher wind speeds. Diffuse urban, NET and crustal and traffic factors all follow a trend which is synchronized to the daily cycle of anthropogenic activity and traffic as influenced by greater atmospheric stability at night. The secondary factor shows a small diurnal range. Fuel oil is highest during the evening and night and may correspond to home heating rather than shipping emissions. The particle size distributions associated with the marine and NET and crustal sources are of limited value as these sources are dominated by coarse particles, beyond the range of the SMPS data, although there is a sharp increase in the volume of the particles above  $0.5\ \mu\text{m}$  in the marine factor. As pointed out in Beddows et al. (2015), the marine factor is identified by its chemical profile of sodium and chloride and is accompanied by an aged nucleation mode at around  $30\ \text{nm}$ . This can be either viewed simply as clean marine air being “polluted” by traffic emission and/or as the consequence of nucleation occurring over at city in clean maritime air masses (Brines et al., 2015). The key point here is that the factors derived in this work are comparable to those factorized in Beddows et al. (2015) using the combined dataset, and the advantage of the two-step approach is that now we have quantified hourly time series  $\mathbf{G}$ .

The hourly contributions are aggregated into daily values and plotted as bivariate plots in Fig. 5 to assist comparison with the daily plots in Beddows et al. (2015). In that work, the same PMF analysis of the NSD data yielded four factors which are named identically to those in the bivariate

plots. The similarity of both of the polar and annular plots for each of the six factors supports our previous factor identification. The secondary and diffuse urban factors are background sources with strongest contributions in the evening and morning. Traffic is strongest for all wind speeds from the east, which makes sense since North Kensington is to the west of the city centre of London where traffic is expected to be most dense. Nucleation is also seen to be strongest for wind from the west, which is expected to be cleaner and have a lower condensation sink. NET and crustal and fuel oil factors are similar to the diffuse urban factor, suggesting a similar predominant source location in the centre of London. The marine factor is observed to be strongest for elevated wind speeds for all wind directions, which is consistent with the expected strong contribution for all high wind speeds from the south-west, as observed in the daily polar plots in Beddows et al. (2015).

### 3.4 Composition associated with the nucleation factor

The nucleation factor was extracted from the two-step PMF–PMF analysis, which included pulling the  ${}^1\mathbf{G}_1$ – ${}^1\mathbf{G}_6$  values to zero of factor  ${}^2F_7$ . It might be reasonable to suggest that if the two-step PMF–PMF analysis is repeated and the order of analysis of  $\text{PM}_{10}$  and NSD datasets reversed that it would be possible to derive the chemical conditions within the atmosphere which were conducive to nucleation. For this, the time series of the four NSD factors ( ${}^1\mathbf{G}_1$ – ${}^1\mathbf{G}_4$ ) reported in Beddows et al. (2015) were combined with the  $\text{PM}_{10}$  data. We again assume that the first PMF step has been carried out and that we are satisfied with how the final solution represents the urban environment of the receptor site and that there are no rotational ambiguities. We then carry out the second step PMF analysis on the  $34 \times 591$  input matrix ( $[{}^1\mathbf{G}_1 \dots {}^1\mathbf{G}_4], \text{PM}_{10}[\text{PM}, \text{PM}_{\text{carbon}}, \text{PM}_{\text{ions}}, \text{PM}_{\text{metals}}]$ ). The hourly output uncertainties from the first PMF analysis



**Figure 5.** Diurnal cycles of the derived  $PN_k$  calculated by the fitting of the daily PMF factor profiles to the hourly NSD data fitted (see Eq. 7 and Sect. 2.4). (Left-left column – diurnal trends of  $PN_k$ ; left-middle column – bivariate plot of  $PN_k$ ; middle-right – annular plot  $PN_k$ ; right-right – bivariate plot of  $PN_k$ , plotted using the Openair program. Polar plots show a point coloured according to the key, the number concentration at that point on the plot whose distance from the origin represents wind speed and angle wind direction. Likewise for the angular plots, the number concentration is shown for a wind direction at an hour of the day between 00:00 and 23:00.) Note that the diurnal plots do not start at zero.

of the NSD data  ${}^1\Delta\mathbf{G}_1\dots{}^1\Delta\mathbf{G}_4$  were carried forward into the second PMF analysis by adding them in quadrature to give daily uncertainties. As with the analysis of the auxiliary data in the PM<sub>10</sub>–NSD data, the measurement uncertainties for the PM<sub>10</sub> data (this time the auxiliary data) were naively taken to be 4 times the PM<sub>10</sub> matrix. Extra care could have been taken in assigning the PM<sub>10</sub> uncertainties, but since we force the output using **FKEY**, a simpler approach was taken. In fact, the **FKEY** consisted of a 4 × 4 diagonal matrix of zero values, with an *fkey*<sub>1</sub> of 20 for all the off-diagonal positions joined to a 4 × 30 matrix of zeros. Furthermore, the uncertainty values of the PM<sub>10</sub> were scaled until  $Q/Q_{\text{theory}} = 0.99$  using parameter  $b_{\text{scale}} = 0.35$  (see Table S3 for more details).

Ideally, the chemical data would be limited to the composition of the particles in the same size range as the SMPS data. However, since we are using the PM<sub>10</sub> composition data, we can at best describe the composition of the aerosol which accompanied each factor (Fig. S4). For the NSD secondary factor, with its strongest contribution (indicated by the explained variation) ~ 400 nm, we have a strong contribution to PM<sub>10</sub> and PM<sub>2.5</sub> together with nitrate, sulfate and ammonium. The diffuse urban factor, with its strongest contribution at 100 nm, is accompanied by contributions from elemental carbon and wood smoke, indicative of traffic and recreational wood burning. There are also contributions from barium, chromium, iron, molybdenum, antimony and vanadium, all indicative of non-exhaust traffic emissions and the burning of fuel oil. Similarly, the traffic factor has a modal diameter of roughly 30 nm, which is indicative of exhaust emissions, and this is accompanied by contributions to aluminium, barium, calcium, copper, iron, manganese, titanium and various other metals attributed to vehicles, albeit from tyre or brake wear or resuspension.

The nucleation factor, with its peak ~ 20 nm, was associated with marine air, as indicated by the strong contributions to Na, Cl and Mg (Fig. S4). There are also traces of V, Cr and Ni and a high contribution to PM<sub>10</sub> mass, which are all associated with marine air. This is explained by an association with the south-westerly wind sector, which brings strong winds and marine aerosol rather than reflecting the composition of the nucleation particles themselves. Marine air is considered to provide the conditions required of an air mass conducive to nucleation, i.e. cleaner air with particles with a low condensation sink. As these air masses pass over the land and eventually into London, anthropogenic precursor gases are added to this air, which then nucleate particles seen at the receptor site as a nucleation mode. This also goes some way to explain the earlier observation of aged nucleation particles observed in the marine factor in Fig. S3. There are also strong contributions to vanadium, which is most likely from an unresolved fuel oil source being mixed into the marine and diffuse urban factors.

## 4 Conclusions

A two-step PMF analysis method is presented, whereby existing PMF profiles can be extended to incorporate auxiliary data concurrently measured and with different units. This is exemplified using PM<sub>10</sub> and NSD data.

When analysing PM<sub>10</sub> composition data, the inclusion of auxiliary data such as meteorological, gas and particle number data has proved to give a clearer separation of factors. However, for a successful output, there must be no rotational ambiguity in either the PM<sub>10</sub> data or in the auxiliary data. In the ideal case, the individually computed factors  $\mathbf{G}(\mathbf{X})$ ,  $\mathbf{G}(\mathbf{Z})$  and  $\mathbf{G}(\mathbf{X}, \mathbf{Z})$  need to be similar if the joint model is to be successful and not produce large residuals and hence too large a  $Q$  value. In the best case, the total weight of the PM<sub>10</sub> data can be set higher than the auxiliary data so that the PM<sub>10</sub> data drive the analysis. In this work, we present an alternative method called the two-step PMF method. In the first step the PM<sub>10</sub> data are PMF-analysed using the standard approach without the inclusion of additional data. An appropriate solution is derived using the methods described in the literature in order to give an initial separation of source factors. The time series  $\mathbf{G}$  (and errors) of the PM<sub>10</sub> solution are then taken forward into the second step, where they are combined with the NSD data. The PMF analysis is then repeated using the combined and mixed unit  $\mathbf{G}$  time series and NSD dataset. In order to ensure that unique factors are obtained for the  $\mathbf{G}$  scores, **FKEY** is used to pull off-diagonal values to zero, thus driving the NSD data. This ensures that the NSD factors are specific to the PM<sub>10</sub> solution and the PM<sub>10</sub> analysis is not affected by any rotational ambiguity of the NSD data. For our demonstration using the Beddows et al. (2015) analysis, this results in six PM<sub>10</sub> factors whose time series are not only apportioned in mass, but the source profiles are identified for the NSD data. Comparisons of the factor profiles, diurnal trends and bivariate plots to those of Beddows et al. (2015) show that this technique produces one solution linking the two separate solutions for PM<sub>10</sub> and NSD datasets together. This generates confidence that the NSD and PM<sub>10</sub> factors ascribed to one source are in fact attributable to that same source.

Hence, the process starts with a dataset which produces a solution which is sensitive to mass, but the factors more sensitive to number can be accessed using a second step. Furthermore, by exploring a higher number of factors, NSD factors which are insensitive to PM<sub>10</sub> mass can be identified as in the case of the nucleation factor. This information can also be extracted using a linear regression, PMF–LR, whereby the size bins of the NSD data are regressed against the PM<sub>10</sub> PMF time series. For this dataset, the nucleation factor profile is identified as an intercept within the fitted model, leading to an increase in the number of PMF factors from six to seven.

*Data availability.* Data supporting this publication are openly available from the UBIRA eData repository at <https://doi.org/10.25500/edata.bham.00000306> (Beddows and Harrison, 2019).

*Supplement.* The supplement related to this article is available online at: <https://doi.org/10.5194/acp-19-4863-2019-supplement>.

*Author contributions.* DCSB conceived the two-step method, ran the PMF and other data analyses and wrote the first draft of the paper. RMH provided constructive criticism, contributed to the data interpretation and wrote sections of the final paper.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* The collection of data used for this paper was funded by the Natural Environment Research Council Clean Air for London (ClearFlo) Programme through grant number NE/H003142/1. The contributions of Gary Fuller and David Green (King's College, London) are gratefully acknowledged. The National Centre for Atmospheric Science is funded by the UK Natural Environment Research Council (grant number R8/H12/83/011). Figures were produced using CRAN R and Openair (R Core Team, 2016; Carslaw and Ropkins, 2012).

The authors are grateful to three anonymous reviewers and to Pentti Paatero for their detailed and constructive comments on the first version of this paper.

Edited by: Ari Laaksonen

Reviewed by: Pentti Paatero and three anonymous referees

## References

- Beccaceci, S., Mustoe, C., Butterfield, D., Tompkins, J., Sarantidis, D., Quincey, D., Brown, R., Green, D., Grieve, A., and Jones, A.: Airborne Particulate Concentrations and Numbers in the United Kingdom (phase 3), Annual Report 2011, NPL Report as 74, available at: [https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1306241448\\_Particles\\_Network\\_Annual\\_Report\\_2011\\_\(AS74\).pdf](https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1306241448_Particles_Network_Annual_Report_2011_(AS74).pdf) (last access: 21 February 2019), 2013a.
- Beccaceci, S., Mustoe, C., Butterfield, D., Tompkins, J., Sarantidis, D., Quincey, D., Brown, R., Green, D., Fuller, G., Tremper, A., Priestman, M., Font, A. F., and Jones, A.: Airborne Particulate Concentrations and Numbers in the United Kingdom (phase 3), Annual Report 2012, NPL Report as 83, available at: [https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1312100920\\_Particles\\_Network\\_Annual\\_report\\_2012\\_AS\\_83.pdf](https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1312100920_Particles_Network_Annual_report_2012_AS_83.pdf) (last access: 21 February 2019), 2013b.
- Beddows, D. C. S. and Harrison, R. M.: Two step PMF data from London Studies, Dataset, University of Birmingham, <https://doi.org/10.25500/edata.bham.00000306>, 2019.
- Beddows, D. C. S., Harrison, R. M., Green, D. C., and Fuller, G. W.: Receptor modelling of both particle composition and size distribution from a background site in London, UK, *Atmos. Chem. Phys.*, 15, 10107–10125, <https://doi.org/10.5194/acp-15-10107-2015>, 2015.
- Bigi, A. and Harrison, R. M.: Analysis of the air pollution climate at a central urban background site, *Atmos. Environ.*, 44, 2004–2012, 2010.
- Brines, M., Dall'Osto, M., Beddows, D. C. S., Harrison, R. M., Gómez-Moreno, F., Núñez, L., Artíñano, B., Costabile, F., Gobbi, G. P., Salimi, F., Morawska, L., Sioutas, C., and Querol, X.: Traffic and nucleation events as main sources of ultrafine particles in high-insolation developed world cities, *Atmos. Chem. Phys.*, 15, 5929–5945, <https://doi.org/10.5194/acp-15-5929-2015>, 2015.
- Carslaw, D. C. and Ropkins, K.: openair – an R package for air quality data analysis, *Environ. Model Softw.*, 27–28, 52–61, 2012.
- Cavalli, F., Viana, M., Yttri, K. E., Genberg, J., and Putaud, J.-P.: Toward a standardised thermal-optical protocol for measuring atmospheric organic and elemental carbon: the EUSAAR protocol, *Atmos. Meas. Tech.*, 3, 79–89, <https://doi.org/10.5194/amt-3-79-2010>, 2010.
- Chan, Y.-C., Hawas, O., Hawker, D., Vowles, P., Cohen, D. D., Stelcer, E., Simpson, R., Golding, G., and Christensen, E.: Using multiple type composition data and wind data in PMF analysis to apportion and locate sources of air pollutants, *Atmos. Environ.*, 45, 439–449, 2011.
- Contini, D., Cesari, D., Genga, A., Siciliano, M., Ielpo, P., Guascito, M. R., and Conte, M.: Source apportionment of size-segregated atmospheric particles based on the major water-soluble components in Lecce (Italy), *Sci. Total Environ.*, 472, 248–261, 2014.
- Emami, F. and Hopke, P. K.: Effect of adding variables on rotational ambiguity in positive matrix factorization solutions, *Chemometr. Intell. Lab.*, 162, 198–202, 2017.
- Fuller, G. W., Tremper, A. H., Baker, T. D., Yttri, K. E., and Butterfield, D.: Contribution of wood burning to PM<sub>10</sub> in London, *Atmos. Environ.*, 87, 87–94, 2014.
- Harrison, R. M., Beddows, D. C. S., and Dall'Osto, M.: PMF analysis of wide-range particle size spectra collected on a major highway, *Environ. Sci. Technol.*, 45, 5522–5528, 2011.
- Hopke, P. K.: A guide to Positive Matrix Factorization, *J. Neuroscience*, 2, 1–16, 1991.
- Leoni, C., Pokorna, P., Hovorka, J., Masiol, M., Topinka J., Zhao, Y., Krumal, K., Cliff, S., Mikuska, P., and Hopke, P. K.: Source apportionment of aerosol particles at a European air pollution hot spot using particle number size distributions and chemical composition, *Environ. Pollut.*, 234, 45–154, 2018.
- Masiol, M., Hopke, P. K., Felton, H. D., Frank, B. P., Rattigan, O. V., Wurth, M. J., and LaDuke, G. H.: Source apportionment of PM<sub>2.5</sub> chemically speciated mass and particle number concentrations in New York City, *Atmos. Environ.*, 148, 215–229, 2017.
- Norris, G., Duvall, R., Brown, S., and Bai, S.: EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide, U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-14/108 (NTIS PB2015-105147), 2014.
- Ogulei, D., Hopke, P. K., Zhou, L., Pancras, J. P., Nair, N., and Ondov, J. M.: Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data, *Atmos. Environ.*, 40, S396–S410, 2006.

- ONS (Office for National Statistics): Population estimates: Persons by single year of age and sex for local authorities in the UK, mid-2014, available at: <https://www.ons.gov.uk/peoplepopulationandcommunity> (last access: 27 February 2019), 2017.
- Paatero, P.: User's Guide to Positive Matrix Factorization Programs PMF2 and PMF3, Part 2, 2002.
- Paatero, P. and Tapper U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 5, 111–126, 1994.
- Pant, P. and Harrison, R. M.: Critical review of receptor modelling for particulate matter: A case study of India, *Atmos. Environ.*, 49, 1–12, 2012.
- Polissar, A. V., Hopke, P. K., and Paatero, P.: Atmospheric aerosol over Alaska – 2. Elemental composition and sources, *J. Geophys. Res.-Atmos.*, 103, 9045–19057, <https://doi.org/10.1029/98JD01212>, 1998.
- R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016, available at: <https://cran.r-project.org/web/packages/openair/index.html> (last access: 21 February 2019), 2016.
- R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018, available at: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nlm.html> (last access: 21 February 2019), 2018.
- Sandradewi, J., Prevot, A. S. H., Weingartner, E., Schmidhauser, R., Gysel, M., and Baltensperger, U.: A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength Aethalometer, *Atmos. Environ.*, 42, 101–112, 2008.
- Sowlat, M. H., Hasheminassab, S., and Sioutas, C.: Source apportionment of ambient particle number concentrations in central Los Angeles using positive matrix factorization (PMF), *Atmos. Chem. Phys.*, 16, 4849–4866, <https://doi.org/10.5194/acp-16-4849-2016>, 2016.
- Thimmaiah, D., Hovorka, J., and Hopke, P. K.: Source apportionment of winter submicron Prague aerosols from combined particle number size distribution and gaseous composition data, *Aerosol Air Qual. Res.*, 9, 209–236, 2009.
- Wang, X., Zong, Z., Tian, C., Chen, Y., Luo, C., Li, J., Zhang, G., and Luo, Y.: Combining Positive Matrix Factorization and radiocarbon measurements for source apportionment of PM<sub>2.5</sub> from a national background site in north China, *Sci. Rep.*, 7, 10648, <https://doi.org/10.1038/s41598-017-10762-8>, 2017.
- Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., Fjåraa, A. M., Asmi, E., Sellegri, K., Depuy, R., Venzac, H., Villani, P., Laj, P., Aalto, P., Ogren, J. A., Swietlicki, E., Williams, P., Roldin, P., Quincey, P., Hüglin, C., Fierz-Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Gröning, C., Faloon, K., Beddows, D., Harrison, R., Monahan, C., Jennings, S. G., O'Dowd, C. D., Marinoni, A., Horn, H.-G., Keck, L., Jiang, J., Scheckman, J., McMurry, P. H., Deng, Z., Zhao, C. S., Moerman, M., Henzing, B., de Leeuw, G., Löschau, G., and Bastian, S.: Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions, *Atmos. Meas. Tech.*, 5, 657–685, <https://doi.org/10.5194/amt-5-657-2012>, 2012.