Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

Atmospheric Chemistry and Physics Discussions

RECEPTOR MODELLING OF BOTH PARTICLE COMPOSITION AND SIZE DISTRIBUTION FROM A BACKGROUND SITE IN LONDON, UK – THE TWO STEP APPROACH

David C.S. Beddows and Roy M. Harrison*†

National Centre for Atmospheric Science
School of Geography, Earth and Environmental Sciences
University of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom

*To whom correspondence should be addressed.
Tele: +44 121 414 3494; Fax: +44 121 414 3708; Email: r.m.harrison@bham.ac.uk

†Also at: Department of Environmental Sciences / Center of Excellence in Environmental Studies, King Abdulaziz University, PO Box 80203, Jeddah, 21589, Saudi Arabia

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

Atmospheric Chemistry and Physics Discussions

ABSTRACT

Some air pollution datasets contain multiple variables with a range of measurement units, and combined analysis by Positive Matrix Factorization (PMF) is problematic, but can offer benefits from the greater information content. In this work, a novel method is devised and the source apportionment of a mixed unit data set (PM_{10} mass and Number Size Distribution NSD) is achieved using a novel two-step approach to PMF. In the first step the PM_{10} data is 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_{10} 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 apportioned NSD data associated with the PM_{10} 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_{10} 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_{10} mass.

Keywords: PM_{10}; London; PMF; source apportionment; receptor modelling

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

Atmospheric Chemistry and Physics Discussions

1. INTRODUCTION

It is unquestionable that worldwide, the scientific vista of air quality is expanding; whether it is the increasing number 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 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 PMF as a tool for source apportionment of particle mass using multicomponent chemical analysis data are published almost daily 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). However, while combining, for example, particle chemical composition and size distribution data in a single PMF analysis may assist source resolution, it does not allow quantitative attribution of either particle mass or particle number to the source factors.

Comero et al. (2009) alluded to the problem of including more than one metric with different units when citing Hopke (1991). In order to obtain a physically realistic PMF solution some natural constraints must be satisfied, one being, "*Only for chemical elements or compounds, where the unit of measurement are the same, the sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts (only in the case of chemical elements or compounds)".* This underlies the necessity to have a consistency of

units throughout the input dataset in order to make a quantified apportionment. To exemplify this point, in Harrison et al. (2011), NSD data (merged SMPS and APS data) was analysed 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. The size distribution profiles, bivariate plots and diurnal cycles were presented but the contributions of each factor were limited to percentage contributions simply because of the mixed units which were inputted into the analysis, and there can be no confidence as to whether the sources are apportioned by units of number concentration ($1/cm^3$) or any of the other units used in the auxiliary data. Chan et al. (2011) identified this "*as a matter of debate within the community concerned"* when considering the use of multiple types of composition data for source apportionment. They 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, 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}C data to constrain the PMF run as *a priori* information.

The potential for an improved factor solution obtained by mixing data types in PMF provides a motivation in the community to develop a methodology which can overcome the aforementioned difficulties. In this study, we present such a method for analysing

simultaneously collected PM_{10} composition and NSD data. In the work of Beddows et al. (2015), both particle composition and number size distribution (NSD) data from a background site in London (2011 and 2012) was 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 the analysis but to conduct separate PMF analyses for PM_{10} mass and particle number. This yielded a 6 factor solution for the PM_{10} data (Diffuse Urban; Marine; Secondary; Non-Exhaust Traffic / Crustal (NET/Crustal); Fuel Oil; and Traffic. Factors described as Diffuse Urban; Secondary; and Traffic were identified in the 4 factor solution for the NSD data. A further factor was the Nucleation factor. When combining the 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 Figure 1. However in the analysis, it would still be useful to obtain a number size distribution for each of the 6 PM_{10} factors and/or a chemical composition for the 4 NSD factors.

In this work, we present a continuation of the analysis of Beddows et al. (2015) describing a two-step methodology in which we use the first step to analyse a primary dataset (PM_{10}; units: $\mu g/m^3$) and a then combine the output with a second dataset (NSD; units: $1/cm^3$). The first step identifies sources and apportions their contribution to mass. Then in the second step, PM_{10} factors are augmented by number size distribution factors. We show that a more complete picture of the sources can be obtained using a 2-step (PMF-PMF) analysis. Furthermore, we also consider linear regression as a second step in a PMF-LR analysis to show how this can reveal hidden factors.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

Atmospheric Chemistry and Physics Discussions

2. EXPERIMENTAL

With a population of 8.5 million in 2014 (ONS, 2017), the UK city of London is the focus of study in this work where the London *North Kensington* (NK) Site (*LAT* = 51º : 31' : 15.780" N and *LONG* = 0º : 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 metres 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

For this study, the same datasets considered, and PMF analysis outputs generated, by Beddows et al. (2015) were used. For this, 24h air samples were taken daily over a two year period (2011 and 2012) using a Thermo Partisol 2025 sampler fitted with a PM_{10} size selective inlet, and alongside, Number Size Distribution (NSD) data were collected continuously every ¼ hour using a Scanning Mobility Particle Sizer (SMPS) consisting of a CPC (TSI model 3775) combined with an electrostatic classifier (TSI model 3080) in air dried according to the EUSAAR protocol (Wiedensohler et al., 2012). The particle sizes covered were 51 size bins ranging from 16.55 nm to 604.3 nm. Analysis of this data resulted in PMF source profiles F and source time series G of the PM_{10} and NSD data sets which were carried forward into this work. Further details of the data, collection methods, coverage and first analysis are given in Beddows et al. (2015).

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

2.2 Proxy Data

Besides the PM_{10} mass, estimates of PM mass can be derived using the NSD assuming sphericial particles of a fixed density. For the SMPS settings, a particle size range between 16 and 604 nm is collected which can be used to estimate a $PM_{0.6}$ value using equation 1.

$$PM_{0.6} = \rho_{eff} \times \frac{\pi}{6} \sum_{sizeBins} d^3 \quad (1)$$

Where ρ_{eff} is set to 2 g/cm^3 for a Diffuse Urban background (based upon 1.8-2.5 g/cm^3 for an urban background aerosol; Beddows et al., 2010).

Figure S1 plots the total apportioned PM_{10} mass against the $PM_{0.6}$ estimates and shows that although the SMPS does not account for the whole mass, it does track with the total PM, with a fitted gradient of 0.65, i.e. accounting for 65% of the mass. To account for the particles greater than 600 nm in the PMF analysis, a proxy was used created by using the difference between the total daily apportioned PM mass in the step 1 of the PMF analysis and the mass estimated from the SMPS data. This difference was then converted back into a number and added to the NSD matrix of counts as $PN_{0.6\text{-}10}$ to improve the match of the NSD matrix to the PM_{10}.

2.3 Methods

2.3.1 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

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

Atmospheric Chemistry and Physics Discussions

ambient aerosol concentration *X* (represented by *n x m* matrix of *n* observations and *m* PM_{10} constituents or NSD size bins), measured at one or more sites can be explained by the product of a source profile matrix *F* and source contribution matrix *G* whose elements are given by equation 1:

$$x_{ij} = \sum_{k=1}^{p} g_{ik} \cdot f_{kj} + e_{ij} \qquad \text{i=1…n; j=1…m} \tag{1}$$

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 to the receptor on the i^{th} hour, f_{kj} is the fraction of 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 *E* and the two matrices *G* and *F* are obtained by an iterative algorithm which minimises the object function Q (see equation 2).

Given, the data and uncertainty to matrices for the model, equation 1 is optimised in the PMF2 algorithm by minimising the Q value (equation 2),

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

were s_{ij} is the uncertainty in the j^{th} measurement for hour *i*.

It may be seen from equation (2) that PMF is a weighted technique; the value of Q, and hence the model fit, is determined by the input variables with the lowest values of uncertainty, S_{ij}. Input variables with high uncertainty have little effect upon the value of Q,

even when their residuals are large. This can be used to the advantage of the operator. When apportioning total PM mass in a conventional one-stage 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. In this work, the primary aim was to define a particle size distribution associated with each factor derived from the PMF of PM_{10} composition. Consequently, in the second stage of the PMF, large uncertainties were input for the particle number data, combined with realistic uncertainties for the PM G-values, so that the latter determine the outcome of ("drive") the PMF analysis.

In this work, the *Q* value is outputted by PMF2 and compared to a theoretical value Q_{theory} which is approximately the difference between the product of the dimensions of *X* and the product of the number of factors and the sum of dimensions of *X* (i.e. n x m – p(n + m)) pk x m. For a given number of factors, the whole uncertainty matrix is scaled by a factor X_{scale} until the ratio between Q and Q_{theory} is approximately one (rQ value = $Q/Q_{theory} \approx 1$).

2.3.2. Application of PMF

The two step method is shown schematically in Figure 2. In the current example, it uses the PMF output of Beddows et al. (2015) as a starting point and assumes that a PMF analysis of the PM_{10} chemical composition dataset (Step One) has already been carried out and dealt with as in the previous study. In this current work, a second step which takes the output from the first step and uses it as an input for the second step is developed. This is done by using the G1 time series from the PMF analysis of PM_{10} and combining this with secondary

data, (i.e. NSD data). The uncertainties of the G1 matrix are transferred from the output of the first step and entered as input uncertainties for the second step. For the NSD data, the uncertainties are taken as X times the NSD values in order to be large and ensure that the PMF is driven by the G1 matrix (see Figure 2). The value of X was optimised in Cran R so that the ratio of Q/Q_{theory} ~ 1.

2.3.3 Fkey

Fkey is a feature in for incorporating a priori information into a PMF analysis and is used in the second step of the PMF-PMF analysis. It is used to "pull" elements of the source profiles to zero. This method uses a matrix that indicates the location of suspected zeros in source profiles or contributions (Figure 3). Since here it is concerned with the profiles, this information is given in the form of integer values in an 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, it is certain that only one PM G score from one of the sources will be strong, e.g. the traffic source will be the only contributing to the PM G value in the Traffic NSD profile, and likewise for the other sources: Diffuse Urban; Secondary; Marine; Fuel Oil; and NET & Crustal (Figure 3).

2.4 Regression

The output of the regression of a dependent variable Y regressed against independent variables X1, X2, X3, … Xn is n gradients and one intercept. When n = 1 it yields a line, when n = 2 it is a fitted plane. But when n > 2 or in this case n = 6, it is a multidimensional fitted model. Each of the n gradients show how Y varies with the n X values given that the other X values are fixed and the intercept provides a bias value. If Y is allowed to take on

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

each value of the NSD size bin and X variables are set to the 6 G time series from the first step of PMF analysis, then it can be seen how the NSD are correlated to the 6 G time series and infer an associated NSD for each of the factors derived in the first step of the PMF-LR analysis.

As an alternative second step, each size bin within the NSD was regressed with the six G1 time series, Equation 3.

$$NSD[a,j] = \sum_{a=\ ^1G1 \ldots\ ^1G6} grad_{a,j} + int_j \quad (3)$$

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 gradients associated with 6 of the factors (giving an indication of how each size bin correlates with each of the 6 factors) and an intercept. When $grad_{a,j}$ is plotted against the size bin, 6 plots showing the dependence of each size bin on each of the 6 PM_{10} factors are produced. It is also assumed that these will be comparable to the actual source profile. Similarly, the int_j values are expected to give a background value, possibly noise. However, this method can extract information known as a remainder factor, shown later in this paper.

2.5 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. Given the i^{th} number size distribution,

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

NSD_i, the difference $D_{i,j,k}$ (equation 3), between the k^{th} element and the linear superposition of the k^{th} element of the seven factors $f_{j,k}$ is minimised.

$$D_{i,j,k} = f(x) = \begin{cases} NSD_{i,k} - \sum_{j=1}^{7} a_j \times f_{j,k}, & aj \geq 0 \\ 1 \times 10^{10}, & aj < 0 \end{cases} \quad (3)$$

The Cran R package Non-Linear Minimization (nlm) (R Core Team, 2018) was used to minimise equation 3. A non-negative constraint is placed in the function. If a negative value is returned by any of the a_j values then $D_{i,j,k}$ returns an excessively large value. Furthermore, in order to extract an apportionment to number concentration (1/cm^3) the fitted values were scaled using a factor SA_j. Six values were derived for SA_j by regressing the total particle number (total hourly SMPS) against each of the fitted values a_j.

3. RESULTS AND DISCUSSION

The aim of this work is to take the results from the first step of a PMF analysis where a successful source apportionment study has been completed and then complement the results with a second step to derive further information about the sources. This can be done using a second PMF analysis or a regression.

3.1 2-Step PMF-PMF Analysis

Figure 4 presents our results from the second PMF analysis of a combined dataset. The G1 time series and uncertainties from the first PMF analysis of PM_{10} data are carried over into the second step where they are combined with the NSD data for PMF analysis. The uncertainties of the NSD data are taken as an optimised multiple of the NSD values themselves. Also in order to maintain the solution from step 1 in step 2 the Fkey matrix is

applied to pull elements in the source matrix to zero as described. This ensures that PMF analysis of the NSD data is driven by the G1 time series. This results in a 6 factor solution in which there are unique contributions from one of the G1 scores and an associated NSD source profile, and it is notable that they are surprisingly similar to those calculated for the just-NSD and PM_{10}+NSD data in Beddows et al. (2015). The Diffuse Urban factor has a modal-diameter just below 0.1 μm which is comparable to the NSD factor in the just-NSD analysis. Marine is comparable to the Aged Marine factor derived from the PM_{10}+NSD analysis. The Secondary factor is again the factor with the largest modal diameter (between 0.4 and 0.5 μm) and traffic has as expected a modal diameter between 30 and 40 nm. Fuel Oil is interesting as it appears to be a combination of a nucleation factor and a mode comparable to diesel exhaust seen in the Traffic factor.

3.2 2-Step PMF-LR Analysis

Figure S2 shows the results of the linear regression of the NSD data plotted against the PM_{10} G1 scores and again what is remarkable is the similarity between these correlation plots and both the factors derived in Beddows et al. (2015) and those from the 2-step PMF-PMF analysis. This analysis was carried out using daily averaged data. To obtain hourly information and thus obtain the diurnal patterns, the resulting correlation factors were re-fitted 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 4 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 by doing this. One interpretation 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_{10} 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 factor in addition to the 6 PM_{10} factors and suggests that although the PMF analysis of the PM_{10} 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_{10} composition and cannot be used to infer a composition.

Returning to the PMF-PMF analysis and extending the analysis from 6 factors to 7 factors and adding an extra row in the Fkey matrix which pulls all of the G1 scores to zero in the solution, the same 6 factor solution is obtained with the additional 7^{th} factor (Figure 5 and Figure S3). As expected, this seventh factor is a Nucleation factor by separating out of the fuel oil factor a nucleation mode leaving a mode with a modal-diameter between 50 and 60 nm. In the results of Beddows et al. (2015), the Nucleation factor was only seen when applying PMF to the just-NSD and PM_{10}+NSD data, and in the PM_{10}+NSD results, Fuel Oil was not separated and appeared to be smeared across all 5 factors. A seven factor solution to PMF of the PM_{10} chemical composition data did not reveal this factor, presumably because the mass associated with nucleation mode particles is too small to affect composition significantly.

Another interesting observation is that although only 4 factors were derived from the PMF analysis of NSD data 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 & Crustal factors. The Nucleation factor is only revealed when performing a regression between the NSD size bins and the G scores of the PMF analysis which leads to increasing the factor number from 6 to 7 which yields the Nucleation profile. It is also reassuring that the bivariate plots for of the 7 factors (discussed in the next section) correspond to the bivariate plots given in 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 required. 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. Figure 6 shows the resulting diurnal profiles.

The diurnal trends of the fitted peaks show the values required in equation 3 to fit the 7 daily NSD factors to the hourly NSD data. These have been scaled in these plots according to the integral of the NSD factor measured in $1/cm^3$. The nucleation 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 the is sufficient precursor material to form particles with a low particle condensation sink. Marine is also high during the day presumably due to higher wind speeds. Diffuse Urban, NET and Crustal, and Traffic all

354 follow a trend which is synchronised to the daily cycle of anthropogenic activity and traffic
355 as influenced by greater atmospheric stability at night. Secondary also follows a similar
356 anthropogenic cycle and would be expected to be strongest at night. Fuel Oil is highest
357 during the evening and night and may correspond to home heating rather than marine
358 activity. The particle size distributions associated with the Marine and NET and Crustal
359 sources are of limited value as these sources are dominated by coarse particles, beyond
360 the range of the SMPS data.

361

362 The hourly contributions are aggregated into daily values and plotted as bivariate plots in
363 Figure 7 to assist comparison with the daily plot in Beddows et al. (2015). In that work, the
364 same PMF analysis of the NSD data yielded 4 factors which are represented here again in
365 the bivariate plots. The similarity of both of the polar and annular plots for each of the 4
366 factors justifies our aformentioned factor-fitting assumption. The Secondary and Diffuse
367 Urban are background sources with strongest contributions in the evening and morning.
368 Traffic is strongest for all wind speeds from the East which makes sense since North
369 Kensington is to the West of the city centre of London where traffic is expecting to be most
370 dense. Nucleation is also seen to be strongest for those wind direction from the West which
371 are expected to be cleaner, and have a lower condensation sink. NET & Crustal and Fuel
372 Oil are similar to Diffuse Urban suggesting a similar predominant source location in the
373 centre of London. Marine is observed to be strongest for elevated wind speeds for all wind
374 directions which is consistent with the expected strong contribution for all high wind speeds
375 from the South West, as observed in the daily polar plots in Beddows et al. (2015).

376

377

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

3.4 Composition of Hidden Factor

The Nucleation factor was extracted from the two-step PMF-PMF analysis when forcing the condition of no PM_{10} contribution through G1 to G6. It might be reasonable to suggest that if the two-step PMF-PMF analysis is repeated and the order of analysis of PM_{10} and NSD datasets reversed that it would be possible to derive the chemical conditions within the atmosphere which were conducive to nucleation. Ideally, for this the chemical data would be more informed with regards to the composition of the particles below 100 nm. However, when using the PM_{10} data the Nucleation factor was associated with marine air with strong contributions to Na, Cl and Mg (Figure S4). There are also traces of V, Cr, Ni and a high PM level 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. Secondary shows a strong association with ammonium, nitrate and sulphate but there are also traces of organics, Al, Cd, Mn, Pb, Ti and Zn and high $PM_{2.5}$ and PM_{10}. Diffuse Urban makes the smallest contribution to PM but shows strong elemental carbon, wood smoke, Ba, Cr, Fe, Mo, Sb, V and Zn; indications of recreational wood burning and brake dust. Traffic has strong associations with Ba, Al, Ca, Cu, Mn, Ti and Zn which have sources in tyre and brake dust and resuspension.

4. CONCLUSIONS

It is recommended when applying PMF to atmospheric PM data that only metrics with the same unit are input in order to make a meaningful quantitative apportionment. However, the inclusion of meteorological and particle number data has proved to give a clearer separation of factors. Mixed unit datasets limit the PMF to a qualitative analysis and the quantitative step of apportioning the sources to a mass or number concentration has to be

403 omitted. This problem is overcome in this work by using a novel Two-Step PMF approach.
404 In the first step the PM_{10} data is PMF analysed using the standard approach without the
405 inclusion of additional data. An appropriate solution is derived using the methods described
406 in the literature in order to give an initial separation of source factors. The time series G
407 (and errors) of the PM_{10} solution are then taken forward into the second step where they are
408 combined with the NSD data. The PMF analysis is then repeated using the combined and
409 mixed unit G time series and NSD dataset. In order to ensure that unique factors are
410 obtained for the G scores, Fkey is used to pull off diagonal values to zero thus driving the
411 NSD data. This ensures that the NSD factors are specific to the PM_{10} solution. This results
412 in 6 PM_{10} factors which are not only apportioned in mass but are augmented by the NSD
413 data. Comparisons of both the factor profiles, diurnal trends and bivariate plots to those of
414 Beddows et al. (2015), show that this technique produces one solution linking the two
415 separated solutions for PM_{10} and NSD data datasets together. This generates confidence
416 that the NSD and PM_{10} factors ascribed to one source are in fact attributable to that same
417 source.

418

419 Hence, the process starts with a dataset which produces a solution which is sensitive to
420 mass but the factors more sensitive to number can be accessed using a second step.
421 Furthermore, by exploring a higher number of factors, NSD factors which are insensitive to
422 PM_{10} mass can be identified as in the case of the Nucleation factor. This information can
423 also be extracted using a linear regression PMF-LR where the size bins of the NSD data are
424 regressed against the PM_{10} PMF time series. For this dataset, the Nucleation factor profile
425 is identified as an intercept within the fitted model leading to an increase in the number of
426 PMF factors from 6 to 7.

427

5. ACKNOWLEDGEMENTS

The National Centre for Atmospheric Science is funded by the U.K. Natural Environment Research Council. Figures were produced using CRAN R and Openair (R Core Team, 2016; Carslaw and Ropkins, 2012).

REFERENCES

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, 2015.

Beddows, D. C. S., Dall'Osto, M. and Harrison, R. M.: An enhanced procedure for the merging of atmospheric particle size distribution data measured using electrical mobility and time-of-flight analysers, Aerosol Sci. Technol., 44, 930-938 (2010).

Bigi, A. and Harrison, R. M.: Analysis of the air pollution climate at a central urban background site, Atmos. Environ., 44, 2004-2012, 2010.

Carslaw, D. C. and Ropkins, K.: openair - an R package for air quality data analysis, Environ. Model Softw. 27-28, 52-61, 2012.

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.

Comero, S., Capitani, L. and Gawlik, B. M.: Positive Matrix Factorisation (PMF) An introduction to the chemometric evaluation of environmental monitoring data using PMF, JRC Scientific and Technical Reports, 2009. https://pdfs.semanticscholar.org/24d3/7d67993228d05513f877b007a16d9a677ff0.pdf?_ga=2.219748825.1591505292.1532007606-1458741279.1523619082.

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., Masio, 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, Envorn. 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 $PM_{2.5}$ chemically speciated mass and particle number concentrations in New York City, Atmos. Environ., 148, 215-229, 2017.

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 compositon data, Atmos. Environ., 40, S396-S410, 2006.

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.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. Available at: https://www.r-project.org/.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. Available at: https://www.R-project.org/.

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 $PM_{2.5}$ from a national background site in north China, Sci. Rep., 7, 10648, 2017. doi: 10.1038/s41598-017-10762-8.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., Fjaraa, 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., Huglin, C., Fierz-Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Gruning, C., Faloon, K., Beddows, D., Harrison, R. M., 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., Loschau, 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, 2012.

FIGURE LEGENDS:

Figure 1. Venn Diagram showing the summary of the findings of Beddows et al. (2015) applying PMF to PM_{10}-only, NSD-only and PM_{10}+NSD datasets.

Figure 2. Flow diagram showing the flow of data through the 2-step PMF-PMF analysis.

Figure 3. Entries in the Fkey matrix used in step 2 of the PMF-PMF analysis. An extremely strong value of 24 was chosen for Fkey.

Figure 4. Second step PMF result. PM_{10} G score driven PMF with quadruple NSD uncertainties. Also Fkey applied to G score part of F factors to pull off-diagonal elements to zero.

Figure 5. Second step PMF result. 7th factor. PM_{10} G score driven PMF with quadruple NSD uncertainties. Also Fkey applied to G score part of F factors to pull off-diagonal elements to zero.

Figure 6. Diurnal cycles derived from the houly NSD data fitted by the daily Factor profiles. Left-left column – diurnal trends of the fitted peaks; left-middle column – bivariate plot of the hourly fitted peaks; middle-right – annular plot of the hourly fitted peaks; right-right – bivariate plot of the daily averaged fitted peaks, plotted using the Openair program.

532

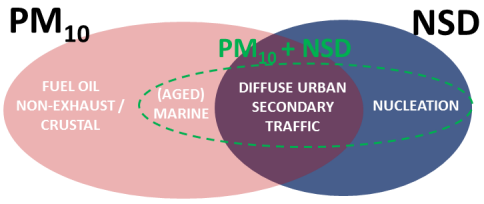


Figure 1. Venn Diagram showing the summary of the findings of Beddows et al. (2015) applying PMF to PM_{10}-only, NSD-only and PM_{10}+NSD datasets.

533

534

535

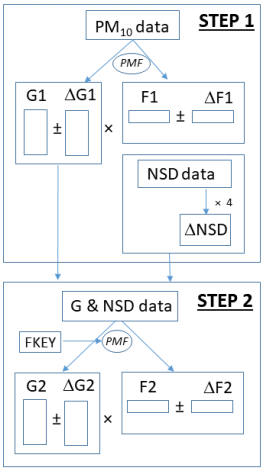
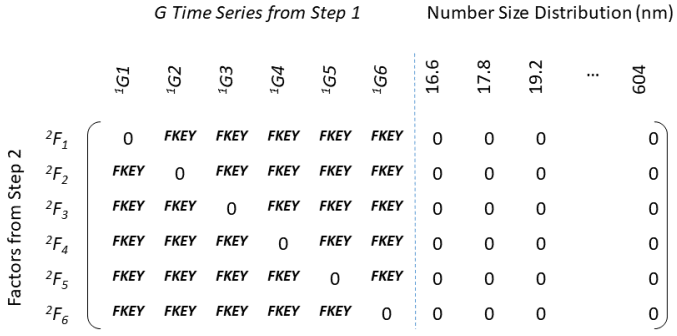


Figure 2. Flow diagram showing the flow of data through the 2-step PMF-PMF analysis.

536
537

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

538
539



540 **Figure 3.** Entries in the Fkey matrix used in step 2 of the PMF-PMF analysis. An extremely
541 strong value of 24 was chosen for Fkey.
542

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

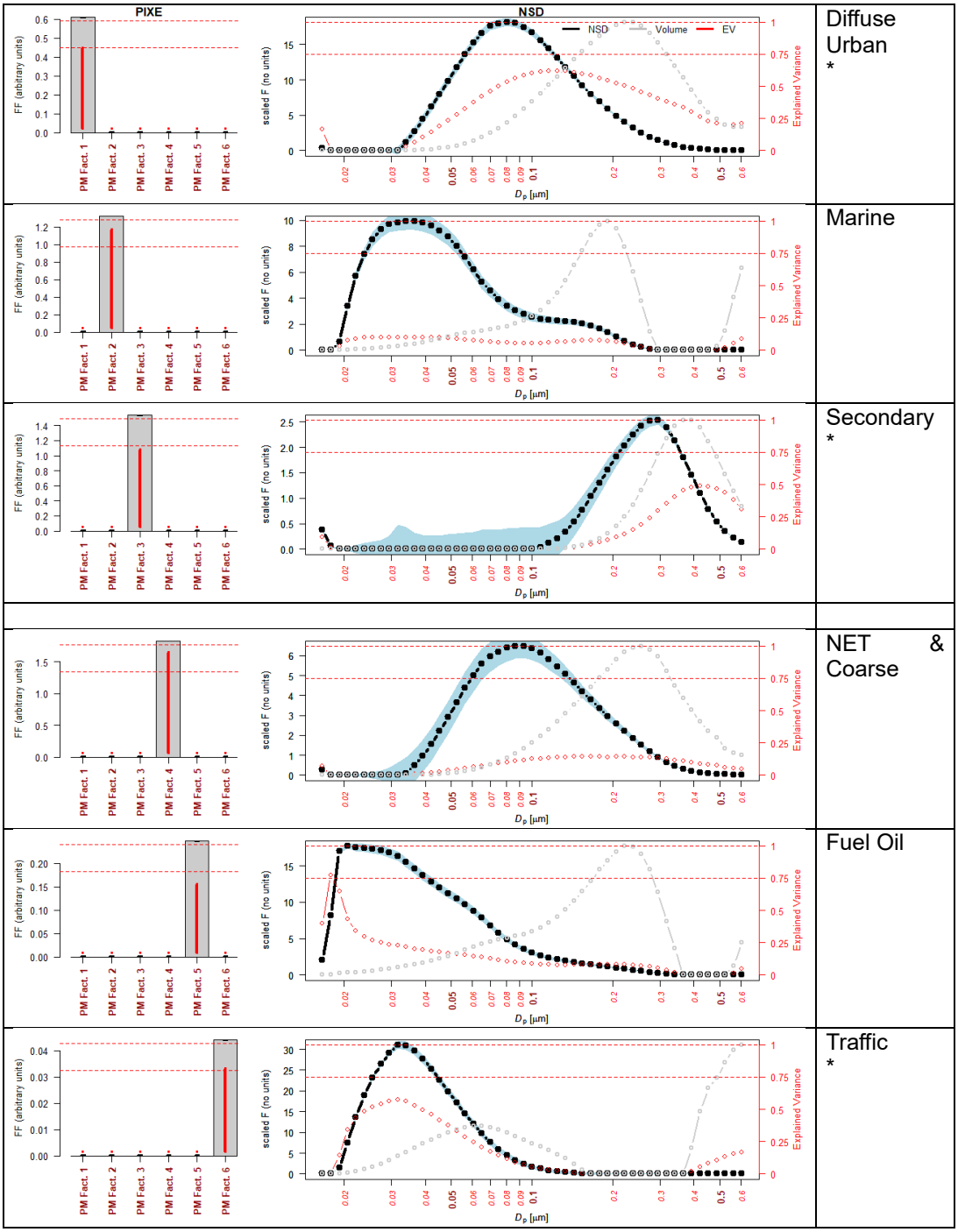


Figure 4. Second step PMF result. PM10 G score driven PMF with quadruple NSD uncertainties. Also Fkey applied to G score part of F factors to pull off-diagonal elements to zero.

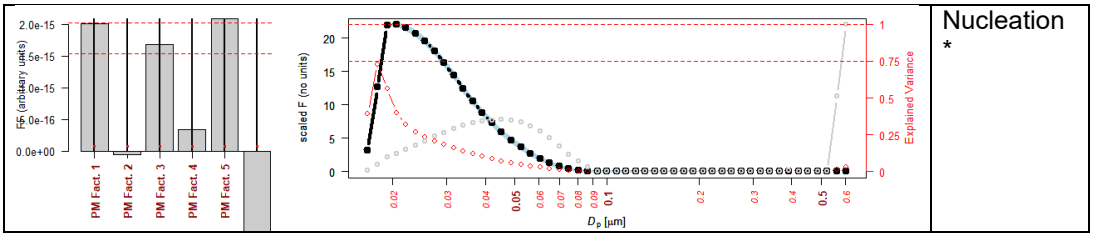


Figure 5. Second step PMF result. 7th factor. PM10 G score driven PMF with quadruple NSD uncertainties. Also Fkey applied to G score part of F factors to pull off-diagonal elements to zero.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 30 August 2018

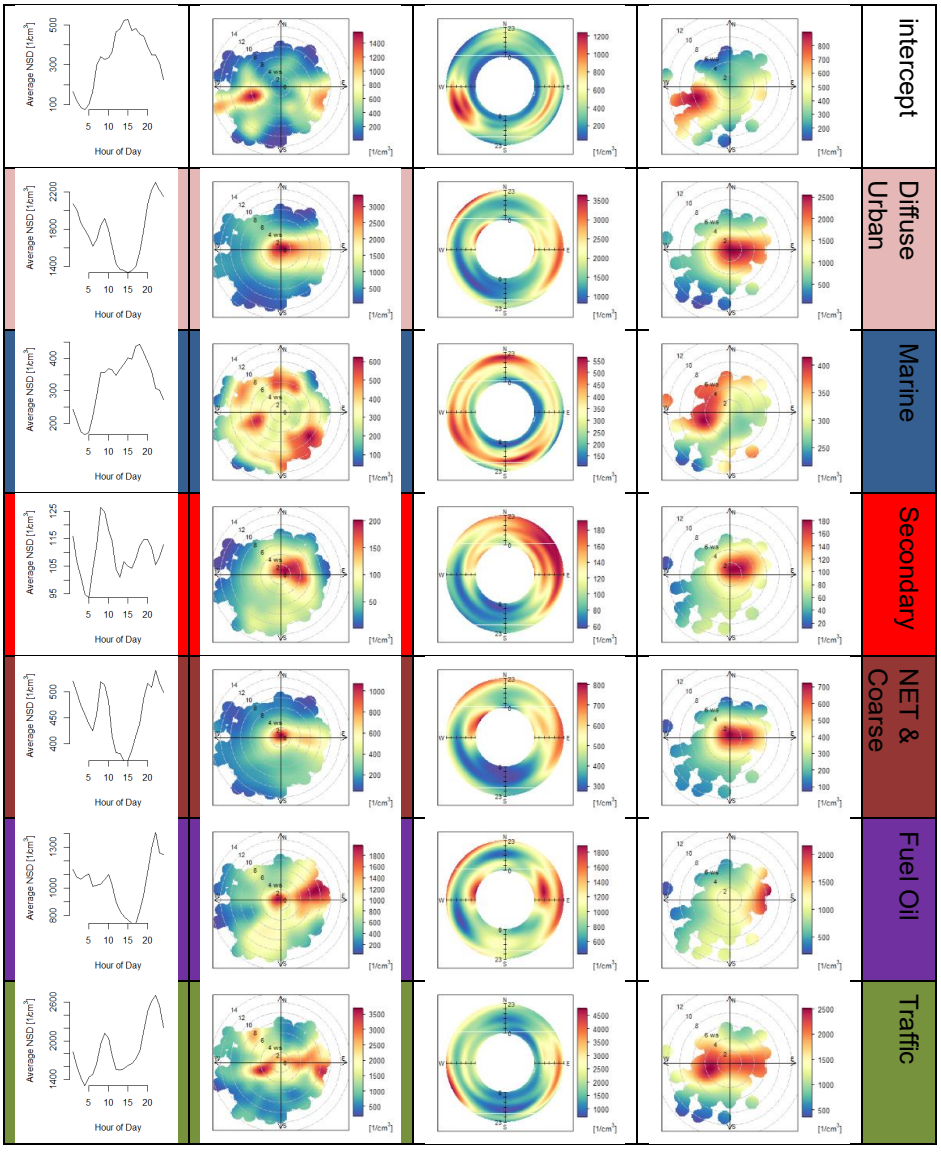


Figure 6. Diurnal cycles derived from the houly NSD data fitted by the daily Factor profiles. Left-left column – diurnal trends of the fitted peaks; left-middle column – bivariate plot of the hourly fitted peaks; middle-right – annular plot of the hourly fitted peaks; right-right – bivariate plot of the daily averaged fitted peaks, plotted using the Openair program.