**Journal: ACP**
**MS No.: acp-2018-784**
**Title: Receptor modelling of both particle composition and size distribution from a background site in London, UK – the two step approach**
**Author(s): David C. S. Beddows and Roy M. Harrison**

# RESPONSE TO REVIEWERS

## REVIEWER #1

General comment: The paper regards the description of a two-step approach for performing source apportionment using the PMF receptor model and an input composed by variables having different measurement units. The approach has elements of originality and potentially several applications. The topic is interesting considering that source apportionment is a major topic in nowadays research and the possibility to use an approach that use input variables having heterogeneous measurement units is certainly appealing. I also believe that the topic is suitable for the Journal and the paper generally well written and understandable. However, I found that some aspects are not completely clear (see my specific comments) and the paper need a revision before publication.
**RESPONSE:** We thank the reviewer for these positive overall remarks.

Specific comments:
Lines 59-62. This sentence is not completely true and I would suggest to modify it. This happens only if number size distributions are mixed with chemical composition (in mass), however, there are examples in which size-segregated chemical composition is used in PMF analysis to obtain quantitative evaluation of size distribution of sources (see for example Contini et al., 2014 Science of the Total Environment 472, 248–261 and references therein).
**RESPONSE:** Text has been added to this effect, saying that by careful experimental design the issue of datasets with heterogeneous units can be avoided, for example using a Cascade Impactor to measure size-fractionated chemical PM mass composition rather than two measurements: one for particle number size and the other for total PM mass.

Lines 140-141. Better to write 16-604 nm (like in line 149) because two decimal digits for size is an illusory precision.
**RESPONSE:** Correction made.

Lines 159-161. The conversion of mass difference in PN0.6-10 is likely quite uncertain. Some details should be given because I believe that some assumptions have been done regarding size distribution in the range 0.6-10 micron and the result of the conversion would be strongly influenced by these assumptions. A comment on this aspect is needed.
**RESPONSE:** This section has been re-written to include more details and a statement added that a large uncertainty is applied to this measurement so as not to influence the final results.

Lines 312-314. This could happen because nanoparticles have a limited mass to influence significantly PM10 mass composition, however, it could be different if NSD are mixed with PM1 chemical composition for example. A comment on this aspect would be useful.
**RESPONSE:** Within the context of response to comments on lines 59-62, we have commented on this to say that with a different measurement, $PM_1$, the NSD data would give a better overlap. However, having said this, PN measurements have a sensitivity bias towards the smaller nucleation particles whereas PM measurements have a bias towards the more coarse particles.

There is a particular reason for using PMF2 and not the more advanced PMF5 that is becoming the standard version of source apportionment with PMF?

**RESPONSE:** PMF2 is not version 2 of the US EPA PMF. PMF2 is the ordinary 2-way factor analysis as opposed to the 3-way factor analysis PMF3 or ME-2 for solving arbitrary (quasi) multilinear models. This has been clarified in the Experimental part of the text.

Lines 215-216. How much is it the X value chosen? This should likely be reported for completeness.
**RESPONSE:** We have added this information, which says that once fitted the NSD data have a relative uncertainty of 4-5%.

Line 259. I believe that the number of factors is six rather than seven.
**RESPONSE:** Yes, the correction has been made.

Lines 358-360. Looking at figures 4 and S3, it seems that the marine source is dominated by nanoparticles. Considering that this is a source generally made of coarse particles, and also authors mention this aspect, this result appears unusual and some discussion and explanations are needed.
**RESPONSE:** This apparent contradiction has already been addressed in Beddows et al. (2015) in the five-factor solution from the combined composition–NSD data set. In this, a factor which can be clearly assigned on the basis of its chemical association is that described as aged marine. This explains a large proportion of the variation in Na, Mg and Cl but shows a NSD with many features similar to that of the traffic factor, with which it has rather little in common chemically. Since the aged marine mass mode is expected to be in the super-micrometre region and hence well beyond that measured in the NSD data set, it seems likely that the size distribution associated is simply a reflection of other sources influencing air masses rich in marine particles. The main point to take away is that we get the same solution using the 2-step approach.

Minor comments:
Lines 148. "spherical"
**RESPONSE:** Yes, the correction has been made.

Line 351. "there is..." The source "NET and crustal" is reported in the text but repeated in the figures as "NET and coarse". I would suggest to use "NET and crustal" in all the paper that is more understandable and appropriate.
**RESPONSE:** Yes, the correction has been made to be consistent with our original work in Beddows et al. (2015).

Title Section 3.4. Why hidden?
**RESPONSE:** The work hidden has been replaced by unresolved. It was not resolved until a 7 factor solution was chosen using an FKEY matrix (as specified in Figure 3 with 6 x 6 zero diagonal FKEY matrix augmented with an 7$^{th}$ column and 7$^{th}$ row of zero entries).

Moreover, this section is dedicated to several factors...what is the hidden one the nucleation? An explanation or a change of the title is needed.
**RESPONSE:** Extra explanation is given.

What is the meaning of the "*" reported in figures 4 and 5?
**RESPONSE:** These have been removed.

**REVIEWER #3**
The manuscript has several elements of originality (to my knowledge, no similar methods have been already published) and directly hits a very controversial and up-to-date topic in the atmospheric sciences. Nowadays, source apportionment by PMF is amply used in both routine monitoring and research studies. Although most of them use "one-kind" variables (mostly PM chemical speciation data), an increasingly high number of studies (just a few have been cited in the manuscript, but the list should be improved) use variables with multiple units. Since the large number of available air quality measurement techniques, the merging of dataset(s) with different units is a suitable (and proven) way to better resolve the PMF source

profiles and to detect unresolved sources. Essentially, additional variables may help in better detecting the edges.
**RESPONSE:** We thank the reviewer for this positive perspective on the work.

Under this view, a recent paper (Emami and Hopke, Chemometr. Intell. Lab. 162 (2017) 198–202, which findings are unfortunately not considered in this manuscript), showed the effect of adding variables with different units to decrease the rotational ambiguity of PMF solutions.
**RESPONSE:** This paper has now been cited.

Thus, the topic is suitable for the journal ACP. However, the manuscript needs revisions before to be accepted for publication. Major points. Essentially, the rationale behind the whole manuscript is based on the statement reported in lines 59-62: "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.". Later, the authors also presented a case study where they mixed variables with different units without giving quantitative results. Even if one can agree with this statement, the authors have not exhaustively explained it. Since this is a methodological manuscript, I strongly encourage the authors to better support these statements.
**RESPONSE:** The comment regarding the unapportioned factor analysis of data with heterogeneous units from the supporting output of Beddows et al. (2015), is referred to in Section 3.2 entitled 2-Step PMF-LR Analysis. We have expanded this section of text to report more clearly, what was carried out in the supporting study.

Another major weakness of this manuscript is the lack of sufficient details on the PMF analyses. This point can be easily solved by the authors, who have an extended experience with PMF analysis. This manuscript presents a new approach, so particular care should be given to details so that anyone can easily reproduce what the authors did (and test with their own data). However, details of the PMF are generally missing or they are reported in the companion paper (Beddows et al., 2015).

- For example, the authors should describe the method(s) used to compute the uncertainties in the 1st step (including PN0.6-10, see next point).

**RESPONSE:** This section has been rewritten to give more detail.

- Also, the authors should report how the raw data have been handled (if any correction was done) and the number of variables and cases inputted into the models. For example, they should report the outliers detection and how they managed the missing values (SMPS sampled every 15 min, what is the minimum number of 15 min records to have a valid 1-hour NSD value?).

**RESPONSE:** The details of the SMPS setup are now in Table S1 and a note is added to say that the raw data was quality assured by the National Physical Laboratory (NPL), and to see Beccaceci et al. (2013a,b) for an extensive report on how the data was collected. Furthermore, we addressed these points by carrying forward the descriptions in Beddows et al. (2015) of how the $PM_{10}$ data was collected and prepared for this study.

In addition, since the Q values are used (lines 199-203 and 215-216), they should be reported as well. Furthermore, it is unknown if the authors dealt with the rotational ambiguity of the models.
**RESPONSE:** This is addressed in the response to P. Paatero's comments.

The authors used R to "optimize" X to have Q/Qtheory ~ 1. More details should be reported. In particular, what does "~ 1" mean? It can be every number, but having it from 0.5 to 1.5 or from 0.99 to 1.01 makes a big difference. Please explain.
**RESPONSE:** We have set a criterion of within 1 ± 0.02.

Basic information on the PMF set-up is important to report. This information will allow the reader to completely understand what the authors did and (possibly) to reply the methods. It would be useful to

have a quick overview of such details in the main text with the deepest description in the supplementary information.
**RESPONSE:** This point has been addressed within the rewritten mathematical description of the PMF analysis.

Another unclear point is related to the "proxy-data" used to assess the PN0.6-10 variable. This is an artificial variable: it was not directly measured, but it was computed on the basis of two (three?) main assumptions: (i) particles are assumed to be spherical, and (ii) particles have fixed density. But it is not completely clear if the density is assumed constant over the time or over the whole (16-604 nm) size spectra (or both, as it should be). The authors used a density of 2 g/cm3 over all the study period, but they report a 1.8-2.5 g/cm3 range for an urban background aerosol. Consequently, the PN0.6-10 variable will be affected by a large uncertainty that cannot be well assessed. I suggest to add more details and provide an estimate of the uncertainty of this new variable.
**RESPONSE:** Clarification of this has been made by adding a fuller and more mathematical description to explain how the proxy variable is calculated and how the density value is used.

This latter point raises another question. Why the authors did not plan to also use an APS to complete the size range to 10 µm? One can argue that the sampling campaign was not planned to have an APS included (or the merging of SMPS and APS was unreliable). However, my opinion is that this point should be at least mentioned in the text, so colleagues who want to pursue the same approach are advised on the possible use of wide range particle size spectra.
**RESPONSE:** We have added this point to a list of alternative approaches to using the proxy-data at the end of Section 2.2.

Minor comments:
Line 103. Missing bracket ")"
**RESPONSE:** Corrected.

Subsection 2.1: Please add more details on the SMPS set-up. For example, sheath and sample flows, the status of the CPC and electrostatic classifier (serviced, calibrated?), the type of neutralizer (X-ray, 85Kr?), software/algorithm used for the data inversion (or version of the AIM software), use of multiple charge and/or diffusion loss corrections, etc. These details need to be added as supplementary information.
**RESPONSE:** This has been added in Table S1, although it does seem like too much information for what the referee correctly identifies as a PMF methodology paper; it is not a data collection paper.

Line 138: 1 4 hour -> 15 min
**RESPONSE:** Change made although this is considered to be a personal preference.

There are two equations numbered as (3), see pages 11 and 12. This should be fixed, as most of the discussion on the method refers to these equations.
**RESPONSE:** Correction made.

Figure 3 can be easily moved to the supplementary material file.
**RESPONSE:** Figure moved.

Figure 4. NET & coarse should be NET & crustal.
**RESPONSE:** Correction made.

Figure 6. Once printed, the labels and axes of the single plots will be likely unreadable. Please increase the font size and (if possible) please uniform the font style and size among the figures. Also, it is advisable to use a color scale that is also easily readable when the paper is printed with a black and white printer.
**RESPONSE:** We have increased the font size at the expense of the size of the plots which has improved the readability of the text in these plots. However, we have not found a palette which

looks good in colour and preserves the information in black and white. All we can suggest is that a grey scale is used for the option of black and white printing.

Figure 6 shows polarplots and polarannuli. These "openair" analyses are commonly reported in air quality studies and are very helpful to better interpret the data. However, a quick overview of the information provided by these two plots should be briefly reported into the materials and methods section.
**RESPONSE:** General descriptions of polarPlot and polarAnnulus have been added to the Methods Section.

**REVIEWER #4**
The paper by Beddows et al. described a two-step source apportionment methodology on a combined database of both PM mass and number size distribution measurements carried out in London. A previous source apportionment study using the same database had been reported by Beddows et al. (2015). Thus, the novelty of this study could be represented by the methodology development. The topic is interesting, and the methodology would be useful in deal with mixing data types as input in PMF, which provide a better defined source factor and better fit diagnostics compared to when non-combined data were used. However, I found that some aspects are not clear and improvements should be made before the work be published in ACP.
**RESPONSE:** We thank the referee, and we welcome the opportunity to provide greater clarity.

Major comments:
1. The motivation of this study is to clarify the source contribution when a combined database was used in PMF. As the authors state, the combined PM chemical composition and size distribution data in a single PMF analysis could not allow quantitative attribution of either particle mass or particle number to the source factors. However, one could calculate the source contributions either by PM mass or by NSD base on the output results of PMF. The following reference is an example described the source contribution using combined database in PMF. Please clarify this item. Sowlat et al., 2016. Source apportionment of ambient particle number concentrations in central Los Angeles using positive matrix factorization (PMF).Atmos. Chem. Phys., 16, 4849-4866,
**RESPONSE:** We do calculate the source contributions either by PM mass or by NSD based on the process of using output from the PMF results in Beddows et al. (2015). Those results are carried through into this work, so we are already carrying out a 1-step analysis resulting in an apportionment. To address this oversight of the referee, we have added a line to Figure 2 saying "[The PMF analyses of Beddows et al. (2015) are considered as Step 1]." We have also added a table of apportionment values from Beddows et al. (2015) into Figure 1 as an insert showing the apportionment of the factors, and the reference to Sowlat et al. (2016) which is very similar to Harrison et al. (2010), which reports PMF of merged SMPS-APS data and chemical and meteorological data.

2. The two-step PMF-PMF method is new but the results maybe questionable. The G1 time series from the PMF analysis of PM10 chemical composition (Step One) could be considered as a constraint in Step Two, which means that six factors identified by PM mass was also applied to NSD. I think this is why the results from two step PMF-PMF method was different from results using combined dataset of PM and NSD in PMF reported by Beddows et al. (2015).
**RESPONSE:** The reviewer is correct in this interpretation.

Thus, what about the results if using the G1 time series from the PMF analysis of NSD as step one? Please clarify this item.
**RESPONSE:** The aim here is to assign a NSD description to the $PM_{10}$ mass sources, so we are not sure why we would consider a 1-step PMF analysis of the combined G1 + NSD data set without applying an 'FKEY constraint'. When removing the FKEY constraint, there is no clear separation of the G1 scores and we can no longer match the NSD of the resulting factors to the original source. Instead we have to introduce new descriptions based on the 6 factor names: Diffuse Urban; Marine; Secondary; NET / Coarse; Fuel Oil and Traffic. Furthermore, a conclusion from Beddows et al. (2015) was that a better result was obtained when analysing the datasets

separately. This work continues with this recommendation by heavily biasing the analysis to the data analysed in Step 1.

Specific comments:

1. Line 157-160. The particle number greater than 600nm is calculated from the difference between PM10 and PM0.6 estimated from SMPS. Except PM0.6-10, particle density, particle shape (spherical) and size distribution should be know when calculated the PN0.6-10. Please provide more description about the calculation process.

**RESPONSE:** This point has been address in line with the comment of Referee #1.

2. Line355-356. Why the secondary factor be expected to be strongest at night? 3. Line 362-363. These is not Fig.7 in the text.

**RESPONSE:** Typo: Figure 7 is Figure 6. This has been corrected. Furthermore, the secondary factor is expected to be stronger at night when compared to the secondary NSD factor derived in Beddows et al. (2015). In Beddows et al. (2015), both the secondary component derived from the $PM_{10}$ and NSD analysis are strongest at night, and in particular, the $PM_{10}$ secondary factor has a strong nitrate component which does grow to a maximum during the night due to reduced volatility of ammonium nitrate. Clarification has been given.

**REFEREE: P. PAATERO pentti.paatero86@gmail.com**

This manuscript deals with PMF analyses of "combined" data matrices such as [X Z] where X contains elemental composition profiles of aerosol samples and Z contains aerosol number size distributions measured simultaneously with composition profiles. This is an important problem that occurs often in modern aerosol research. There are specific problems in this task; these problems have not been studied in depth in literature so far. This manuscript studies one specific combined data matrix and reports a PMF model for this matrix. Thus the ms might deserve publication despite of certain serious problems. These problems are in part related to misunderstandings found in earlier papers that discuss this same topic. For this reason, the present review contains a lengthy general discussion of the task of modeling combined matrices. The specific questions regarding this ms are based on this general discussion. The ms might also be suitable for publication in the sister Journal AMT, tmospheric Measurement Techniques. My personal view is slightly in favour of AMT. However, both ACP and AMT seem possible, and this review considers publication in either Journal. The structure of this review is as follows:

**RESPONSE:** We recognise the immense contribution made by Professor Paatero to this field, and thank him for the critical insights which he provides.

==========

Recommendations
Notation used in this review
Background
Common mode errors
Joint matrices containing different units
Discussion of the manuscript
Two-stage PMF model vs. customary PMF model
The hidden factor, aka Nucleation factor
Miscellaneous

**Recommendations**

There are very many problems of different kinds in this manuscript. For this reason, I hesitantly recommend that this ms should NOT be published by ACP or AMT. However, if it is desired to publish this ms because of the importance of the problem, then a thorough rewriting of the text and mathematical details must be undertaken. I recommend that the following enhancements be performed: There has apparently been lack of communication between the person(s) who did the actual computations and those who wrote the paper.

**RESPONSE:** This first author is responsible for both the computation and the manuscript and we have endeavoured to follow these recommendations to avoid the appearance that an unnamed contributor has been involved. The paper has been extensively revised in response to the comments of all four reviewers.

For this reason, the mathematical description is erratic, chaotic and impossible to understand or replicate. In order to create an accurate description, the person(s) who did the computations should be included in the group of authors. Without such help, it may not be possible to achieve a satisfactory mathematical description of what was done. The entire mathematical discussion about problems attributed to PMF analysis of joint matrices containing different units is erroneous, based on a widespread misunderstanding. This discussion must be rewritten according to suggestions given below. It might be good to include in the author group somebody familiar with the quantitative mathematical structure of the PMF model.

**RESPONSE:** Thank you for highlighting this misunderstanding which we have addressed in the revised manuscript.

In particular, it seems that lines 79,80 are not based on quantitative understanding of the model. These lines, and other similar sentences, must be removed. Much of Conclusions must be rewritten so that the claims against using variables with different dimensions/units are replaced by opposite sentences stating e.g. that a joint analysis of matrices of variables with different dimensions/units is not harmed by these differences but unfortunately the opposite was believed to be true when the work was carried out.

**RESPONSE:** This correction has been carried out.

The mathematical description of what was done must be totally rewritten so that systematic matrix-form notation is used. Equations must be corrected and written in correct notation, using correct terminology and correct numbering. Details of PMF modeling must be reported, such as dimensions of matrices, used parameters such as uncertainties of data values, robust/nonrobust, obtained Q values, numbers of observed outliers, unique or multiple minima, and so on.

**RESPONSE:** This has been corrected following the guidance of all the other referees.

Rotational questions are an ever-present problem in factor analytic modeling, independently of what programs are used. It is alarming that the word "rotation" does not occur in this manuscript. Pay attention to rotational questions. There are certain weaknesses in the plan of this work, such as assuming that the rotational status of the original PMF model of X was correct or best possible (see below). These weaknesses cannot be corrected in an enhanced ms but they should be briefly discussed. This is important because otherwise, colleagues following the example of this work will feel the need to replicate everything that was done here, being unaware that some details may not have been optimal.

**RESPONSE:** Rotations are now briefly discussed.

Enhance figure captions so that readers do not need to guess what is shown. Have the enhanced ms proofread by colleagues. Check also the references. This ms illustrates, once again, how difficult it is to find ones own mistakes and typos.

**RESPONSE:** Enhancements of figure captions have been carried out.

**Notation used in this review**

The notation "[X Z]" indicates here attached or joined matrices, i.e. placing X and Z side by side so that they form one larger matrix. The notations G(X) and F(X) will indicate factor matrices (G and F) obtained from an individual PMF model of X only, and similarly G(Z) and F(Z) for Z only. The left and right parts of F, when modeling [X Z], are denoted by F[Xz] and F[xZ]. Q(X) and Q(Z) indicate Q values from separate analyzes of X and Z. Similarly, Q[Xz] and Q[xZ] denote Q sums computed over elements of X and over elements of Z in the joint analysis of [X Z]. Hence, Q[X Z] = Q[Xz] + Q[xZ]. Total weight of X means the sum of squares of $X_{ij}/s_{ij}$ over X, where $s_{ij}$ is the uncertainty assumed for $X_{ij}$. If both X and Z are equally important, and if X and Z are of different sizes, all $s_{ij}$ reported for the larger matrix should be increased so that total weights of X and Z become approximately equal. This implies a deviation from the general principle of determining weights from std-dev of values.

**RESPONSE:** An amended notation as suggested has now been implemented.

**Background**

Before examining this manuscript in detail, it is necessary to discuss the model that it tries to solve and the problems that make this task difficult. It is known that PMF of combined matrices often leads to disappointing results, such that some factors only (or mainly) fit X while other factors only/mainly fit Z. Such result is worthless in cases where X and Z are caused by the same emission sources whose emission profiles should be determined for X and Z. It is important to realize what advantages may be expected from the joint analysis of X and Z. Three Cases are possible: PMF models computed separately for X and for Z may be valid and rotationally unique for (A) both X and Z, (B) one of them (for X, say), or (C) neither one of them.

Case A: If individually computed factors G(X) and G(Z) are practically identical, then a straight-forward joint model is successful for this case. Then G_[X Z] = G(X) = G(Z). If G(X) and G(Z) are significantly different, however, then the joint model will fail, producing too large residual values and hence too large Q. Such result might be caused e.g. by "common-mode errors" (see below) in X and/or in Z.

Case B: Now a joint model should be specified so that total weight (see Notations, above) of better-analyzed matrix X is significantly higher than total weight of Z. Then X will "drive the model", and G_[X Z] will be approximately equal to G(X). If a reasonable Q[xZ] is obtained, then it indicates that X and Z are compatible, i.e. a joint PMF model is meaningful. Larger Z residuals and larger Q[xZ] would be obtained e.g. if X and Z do not have common sources or if there are common-mode errors. Then the joint PMF model is not meaningful for the chosen number of factors.

Case C: individual PMF models of both X and Z contain rotational ambiguity and/or other problems such as unidentifiable factors or missing factors. In this case, the approach of Case B cannot be used because the obtained ambiguous rotation, based mostly on X, may not be the best rotation for fitting Z. Ideally, equal total weights should be applied on X and Z, hoping that the best rotation for fitting both will be obtained when rotational information from Z is combined with information from X. Experience shows that quite often, such modeling fails. Few, if any, studies have been made about the reasons of such failures. It must be stressed that these failures must not be ascribed to "different units used in X and Z" (see below). As a first remedy, one might inspect the residuals in order to see if common mode errors are visible. Such errors might be corrected by hand, or by using an enhanced PMF model that automatically corrects for common mode errors. One might also inspect individual variables in order to see
if only few variables are causing incompatibility of X and Z. Such variables might be downweighted in order to obtain a better overall model. Of course, one must also consider the possibility that in addition to their joint sources, X and Z may also have one or several unique sources. An enhanced PMF model may be developed for analysing such joint matrices containing common and non-common sources.

Summary of Case C: too little is known about reasons why this case fails. Well documented case studies are needed. Singular value decompositions of G matrices computed for X, Z, and [X Z] may be useful for demonstrating the root of the problem. Reliable remedies may only be suggested when more is known about the reasons for failures in joint PMF modeling.

**RESPONSE:** An account of cases B and C has been added to the paper.

**Common mode errors**

Certain problems in measurements will cause so-called "common mode" errors. E.g. an error in air volume control in an aerosol sampler, when measuring sample i, causes that all aerosol concentrations on row i of X will change by the same fractional amount. Such common mode deviation does not contribute to residuals in customary PMF analysis of such aerosol data. Instead, common mode disturbance of sample i will change all elements of row i of matrix G. In a combined matrix, the other part Z is often measured using another instrument. Then Z may have its own common mode errors, different than those of X. In a joint analysis of X and Z, two independent sets of common mode errors will cause increased residuals when factors are common to X and Z. It appears highly probable that such common mode errors are an important

reason for those PMF results where individual factors tend to fit either X or Z but not both. Joint matrices containing different units This ms claims that quantitative PMF modeling of a joint matrix [X Z] is not possible if variables in X and Z are measured in different units, such as mass concentration (expressed in mass/airvolume) and particle number concentration (expressed in particles/airvolume). These claims are based on a widespread misunderstanding, as explained in this section. Customary aerosol PMF models are often scaled so that the sum of all elements in each row of matrix F equals unity. Then factor element F_pj indicates the fraction of species j in profile of source p. With joint matrices containing different units, summation over a row of F is not meaningful. The following workflow should be used instead in order to preserve the quantitative nature of the model: In PMF (or after PMF), scale factors so that the average of each column of G is scaled ("normalized") to unity. Then elements of F have the following quantitative meaning: F_pj indicates the average contribution of source p to observations in column j, both for species j in matrix X and for species j in Z. The average total amount of all aerosol species in source p is obtained by summing values F_pj over all species j in F[Xz], i.e. in the part of F corresponding to aerosol matrix X. In this way, the customary interpretation of F_pj as fractions of total may be obtained "off-line" after PMF computations by dividing the F_pj values by their sums taken over F[Xz]. The ms also suggests that presence of other variables (Z) in PMF model somehow makes the model non-quantitative or unreliable: ms lines 79-80: there can be no confidence as to whether the sources are apportioned by units of number concentration (1/cm3) or any of the other units used in the auxiliary data. Units may be entirely ignored in PMF modeling if all variables are represented in same units. If different units are present in different columns of matrix X, then the following practice is followed: elements of factor matrix G are pure numbers. Elements in column j of factor matrix F carry the same dimension and unit as column j of data matrix X. In the present case, all elements of left part F[Xz] of factor matrix F will be in mass/airvolume (same as X) while all elements of the right part F[xZ] are in units of number concentration (1/cm3) (same as Z). There is no confusion regarding dimensions or units.

Disturbance of quantitative modeling of X by "other variables" in Z may only be present if Z variables make the fit of X extremely poor, so that Q[Xz] increases to unacceptable levels in comparison to the original Q(X). This can be seen from Eq. (1) which defines PMF model: all values in column j of X are fitted using F factor elements from column j of F only. The "other columns" in F, corresponding to "other variables" in Z, do not enter in the fit of any X variables. If Q[Xz] remains normal, model of X remains quantitative even when Z is introduced in modeling. However, if introduction of Z requires that number of factors must be increased, then the two models are different. Then rotational uniqueness and interpretatability of the joint model of [X Z] may well be better or worse in comparison to the original model of X only. On the other hand, G(X) and G[X Z] may appear significantly different even when all Q values are normal. In this sense, including Z may interfere with the fit of X although the new fit of X remains as quantitative (or better) than the original fit of X. Such effect depends on rotational ambiguity of the original PMF fit of X: when Z is introduced, it may "rotate" a rotationally ambiguous model of X so that Z obtains a better fit while Q[Xz] does not increase from Q(X) or increases a little. Such rotation may only occurr if the original model of X is rotationally ambiguous, "non-quantitative". If such ambiguity is not understood by the scientist, it might appear that introduction of other variables "harms" the original model. In contrast, however, modifying the original model of X by a rotation is what is desired when using the joint model: both X and Z should be fitted as well as possible. This effect does not harm the quantitative nature of the model, as long as Q value of X does not grow too much. Summary of this section: if Q computed over X elements increases significantly when modeling [X Z] instead of X, this indicates that X and Z are not compatible (when assuming this number of factors). Then analysis of [X Z] should be rejected. In all other cases, the joint model of X is equally good or better than the original model of X. If original model is rotationally ambiguous, then factors usually change: G[X Z] is different from G(X) and similarly F[Xz] is different from F(X). These new factors fit X as well as the original factors, thus they are as quantitative as the original factors. The rotation of these new factors takes into account information from matrix Z. In some cases, the new factors are rotationally unique, without any ambiguity. More often, the ambiguity of new factors is less than the original ambiguity.

**RESPONSE:** This fullsome explanation is very valuable, and aspects of this background relevant to our paper have been added to the manuscript.

**Discussion of the manuscript**

This manuscript suffers badly from almost complete avoidance of equations and mathematical symbols and mathematical notation in general. Also, there are serious problems in the few equations that are present. A more compact and easier to read presentation is obtained if mathematical notation is used as the primary means of communication. It is possible that part of my criticism in this review is simply based on misunderstanding unclear and/or ambiguous verbal explanations of mathematical concepts.

**RESPONSE:** We have addressed this by a new, more mathematical description of our methods.

The ideal of scientific work is repeatability. This ms does not provide facts that might enable repeatability, even in principle. E.g., I could not find dimensions of data matrices or obtained Q values.

**RESPONSE:** Dimensions have been added.

How were NSD data preprocessed before PMF computations? Using averages or medians? How were outliers handled? How many factors were used in each case? And so on.

**RESPONSE:** This information has been added together with a reference to a report provided by the data provider, NPL.

The basic assumption of factor analytic modeling is that for each source, chemical profile and size distribution stay constant throughout the measurement campaign. On the other hand, it is well known that whenever nucleation happens, aerosol size distributions do vary. Also, largest particles tend to settle down more during longer transit times. In this work, constancy of size distributions was silently assumed. It might be good to discuss this fundamental question in future versions of this work.

**RESPONSE:** Point taken. This is something that we have mentioned in previous papers with reference to the assumption that the profile of the sources does not change between emission and arrival at the receptor site.

Two-stage PMF model vs. customary PMF model In the present ms, the goal was to determine the size distributions corresponding to the previously determined aerosol composition sources. It was assumed (on what grounds?) that the rotation of the original PMF result was correct, so that the originally obtained G matrix was deemed suitable for the PMF model of NSD matrix Z.

**RESPONSE:** We clarify the assumption that we were satisfied with the solution from the first analysis referring the reader to Beddows et al. (2015). This solution gave the best solution/rotation (agreed with all the authors during the work) to describe the urban atmosphere measured at the NK site. There are details in Beddows et al. (2015) justifying this.

In other words, it was desired that X "drive" the modeling of [X Z]. Essentially, this method corresponded to Case B, discussed above. Apparently, the authors were unaware of the one-stage method suggested for Case B.

**RESPONSE:** We were aware of this method but avoided it in view of having a united G factor which would not be possible with a joint matrix [X Z].

In hindsight, the best approach might have been to follow both Case B and Case C, especially if there was no positive information confirming that the original PMF model of X was rotationally unique and correct. An enhanced version of the ms should briefly discuss the one-stage possibilities of doing this work according to Case B and/or Case C.

**RESPONSE:** We have added two sections which describe Case B and case C.

The one-stage method, with suitably weighted X and Z, would be easier to explain and much easier to understand. However, it is not reasonable to expect that the work be redone using the one-stage approach. I understand step 2 so that the computed G factors from step 2 were forced to be practically identical to G factors from step 1. Is this right?

**RESPONSE:** Yes, this is correct.

If this is right, then step 2 appears to be equivalent to non-negative weighted regression (non-negative weighted linear least squares fit) of matrix Z by columns of matrix G. This should be mentioned.
**RESPONSE:** We have now mentioned this.

There are easy-to-use computer programs for computing such LS fits. Although PMF may also be used for this fit, using simpler tools would make the process more transparent, so avoiding unnecessary complications. Equations for defining the hidden factor should be given. The verbal definition is hard to understand and I did not manage to understand it.
**RESPONSE:** We hope that the clarification of the description and mathematics will mean that this is now expressed clearly.

**The hidden factor, aka Nucleation factor**
It is a good idea to assume that due to its higher time resolution, the NSD matrix Z may contain factors that are not visible in matrix X of chemical profiles. Unfortunately, the method for defining the hidden factor(s) in Z is questionable. First of all, why did you assume that there is only one hidden factor?
**RESPONSE:** The hidden factor was revealed as the intercept in the regression of the NSD values against the G1…G6 timeseries, after which, we then looked for it in the PMF analysis. Furthermore, from the results of Beddows et al. (2015) (for which optimum solutions were derived without factor splitting), we did not anticipate another factor to be present above 7 factors (see the Venn diagram shown in Figure 1) and this constrained our search to 7 factors, i.e. we had accounted for all of the factors in Beddows et al. (2015) and saw no need to go higher.

It seems that in stage 2, 6 factors were used. This is not defined (why not) but this is how I understand the ms. Why did you not use in 2nd stage PMF a 7th (and maybe an 8th) factor that may only fit the NSD part of the data matrix? This simple arrangement would determine hidden factor(s) avoiding the bias that non-negativity constraints may introduce in your method (see below). This alternative must be mentioned in a future version of the paper.
**RESPONSE:** We have clarified this. We initially used 6 and then 7 to find the nucleation factor and only went to 7 factors because of the response given to the previous point, i.e. we only looked for those factors in Beddows et al. (2015).

The second Equation (3) is incorrectly formulated. Symbol j is used as a summation index on the right side. Then it cannot appear on the left side. There is a symbol "x". It is not defined, what does it mean? The text says: "The Cran R package NonLinear Minimization (nlm) (R Core Team, 2018) was used to minimise equation 3." You must not say "minimize equation". You must specify the expression that is minimized, and also specify the free variable(s) that are varied in order to minimize. I cannot understand the expression to minimize nor the free variables. For this reason, I cannot comment more on determining the hidden factor. Maybe it is properly determined, maybe not. This part of the work is certainly not reproducible by others.
**RESPONSE:** This equation has been correctly formulated and numbered.

Bias: It seems that the second Equation (3) is not applied to all data because of nonnegativity constraints (however, there seems to be an error in the constraints, it is impossible to guess what was really intended). When some data are excluded, this creates a bias. It is impossible to know from the outside if this bias was negligible or if it distorted the results. The bias question must be documented.
**RESPONSE:** No data has been excluded. All data was fitted with non-negative constraints.

**Miscellaneous**
Lines 415-417 in Conclusion: "This generates confidence that the NSD and PM10 factors ascribed to one source are in fact attributable to that same source." This is a very important statement, good! There are two equations numbered (3). This caused a LOT of trouble when trying to understand the discussion of the "hidden profile" a.k.a. "nucleation profile". The first Equation 3 does not appear correctly on my computer. Possibly, it uses a symbol font that is not present on my computer so that one symbol is not visible. There is also another problem in this equation: symbol "a" is used as summation index, and symbol "a" appears also on left side. A summation index cannot be present on left side. Please check your equations before

submitting new versions of the ms. Make sure that the .pdf file contains all non-standard fonts that are used e.g. in equations.
**RESPONSE:** Changes made to clarify this matter.

The presentation should be helpful for the reader. The symbols used in text and in equations should be defined. Example: in first Eq. (3), there is symbol j. What does it mean? Is it the index of size bin? Why not help the reader and say so? In second Eq. (3), there is again a symbol j. What is it now? Please update the ms so that symbols are used in a systematic way, in order to help the reader. The following method is recommended in order to avoid confusion with symbols:
**RESPONSE:** Changes made to provide clarity.

For your own use, create a table where each symbol, however trivial, is entered. When needing more symbols, check first with the table if the symbol is already reserved for another use. When you are ready, include short definitions from the table into the ms, either in a table of notation or to the location of first use of each symbol. Use customary matrix element notation whenever possible. In this way, you could avoid using scalar "a" first as an index and then vector a_j as a vector of unknowns.
**RESPONSE:** This system has been adopted in the interests of clarity.

Description of the linear regression model (section 2.4) is strange. I have never seen that the coefficients are called "gradients". Also, correlations should not be mentioned when discussing linear least squares. It would be best to simply show the equation. I recommend that explanation of regression be omitted, except that the equation, using matrix element notation, should be shown.
**RESPONSE:** Changes made as recommended.

Figure 6 is unclear. What is illustrated by the bivariate plots? Figure caption only tells that they are bivariate plots, plotted using the Openair program. Instead of naming the plotting program, it would be more important to define what is plotted vs. what, and what are the dimensions in individual diagrams. After working with the ms for a long time, I tend to guess that the "bivariate plots" might represent NSD concentrations in polar plots of wind direction and wind speed. Why did you not say this? Saving one sentence from the ms may cost hours for your new readers.
**RESPONSE:** A description of the Bivariate plots is given as an added section and the details requested added in the figure legend.

# 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

**ABSTRACT**

Some air pollution datasets contain multiple variables with a range of measurement units, and combined analysis by Positive Matrix Factorization (PMF) ~~is~~ can be ~~-~~ 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 apportioned $PM_{10}$ 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

## 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~~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 NSD data (merged SMPS and 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 (13nm - 10µm) data combined with several auxiliary variables collected in Los Angeles. These included BC, EC/OC, PM mass, gaseous pollutants, meteorological, and traffic flow data. A six-factor solution was chosen comprising of: nucleation, 2 x traffic, an urban background aerosol, a secondary aerosol and a soil factor. The two traffic sources ~~made~~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, 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.

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~~ difficulties arise if the two datasets have different and/or ambiguous rotations (discussed in ~~s~~Section 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. ~~quantitative attribution of either particle mass or particle number to the source factors.~~ Experimental design can of course circumnavigate this problem, for instance, using chemical data which is 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 data~~–~~sets sufficiently overlap across the size bins. However, ~~impact~~ cascade~~rs~~ impactors do not ~~don't necessarily enjoy~~offer the high time

resolution of particle counting instruments, with individual measurements lasting hours or days. Even so, for the case where two or more instruments are available in a campaign to measure two or more different metrics, e.g. PM mass and particle number ~~and~~ (PN), then ~~thea prescribed method~~combined data analysis is useful. ~~Furthermore,~~ Emami and Hopke (2017) have~~s~~ shown~~ed~~ that the effect of adding variables as auxiliary data (with potentially different units) to a NSD data set is to decrease the rotational ambiguity of a solution from a 1-step PMF analysis.

~~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~~

~~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~~ a single 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, together with a~~. A further factor was the~~ Nucleation factor not seen in the $PM_{10}$ mass data analysis (see Figure 1). 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. As a continuation of this work, we present an alternative method ~~of~~ for analysing the combined dataset in a so called, two-step methodology. In the first step, we analyse the mass data ($PM_{10}$; units: $\mu g/m^3$) according to the methodology of Beddows et al. (2015). This results in a time series factor G which is carried forward into a second PMF analysis of a combined data~~-~~set consisting of the G time series and an auxillary data set (~~eg~~i.e. NSD; units: $1/cm^3$). The first step identifies sources and apportions the G factors to their contribution to mass and in the second step, an FKEY matrix is chosen such that G 'drives' the model and the NSD data 'follow'. This means that we have $PM_{10}$ factors each of which ~~are~~ is augmented by its number size distribution. Furthermore, we also consider linear regression as a second step in a PMF-LR analysis to show that although the initial analysis is biased toward mass by analysing $PM_{10}$~~,~~ factors only, ~~seen~~ unseen factors influencing the NSD data (e.g. nucleation) can be identified in the data.

## 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~~As alluded to, this work is a continuation of the study carried out by Beddows et al (2015), which analys~~is~~ed ~~chemical~~ NSD and $PM_{10}$ chemical composition data ~~were~~ collected at the London NK receptor site. ~~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~~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~~.55~~ nm to 604~~.3~~ nm and the 15 min distributions were aggregated up to hourly ~~samples~~averages (where there were at least 3 x 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 and. 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) the reader is also referred to Becceaceci et al. (2013–a,b) for an extensive studyaccount of how the NSD data was collected and quality assured.

Accompanying the NSD data from the study of Beddows et al. (2015) was the PMF output from the analysis of Analysis of this data resulted in PMF source profiles F and source time series G of the $PM_{10}$ chemical $PM_{10}$ composition data. $PM_{10}$ and NSD data sets which were carried forward into this work. The latter data consisted of 24h air samples taken daily over a 2-year period (2011 and 2012) using a Thermo Partisol 2025 sampler fitted with a $PM_{10}$ size selective inlet. These filters were analysed for total metals $PM_{metals}$ (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 HF acid digestion of GN-4 Metricel membrane filters. Similarly, wWater-soluble ions $PM_{ions}$ ($Ca^{2+}$, $Mg^{2+}$ , K, $NH_4^+$ , $Cl^-$ , $NO_3^-$ and $SO_4^{2-}$) 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. TheseData gaps were filled by measurements made on daily $PM_{10}$ 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 thea lower data capture for the cations. But aAgain, 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) utilising Aethalometer and EC/OC data, as described in Fuller et al. (2014). Samples were also collected using a Partisol 2025 with a $PM_{10}$ size selective inlet and concentrations of elemental carbon (EC) and organic carbon (OC) were measured by 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 $PM_{carbon}$. In addition, particle mass was determined on samples collected on Teflon-coated glass fibre filters (TX40HI20WW) with a Partisol sampler and $PM_{10}$ size-selective inlet.

This aforementioned $PM_{10}$ data was represented in this work as the PMF solution for $PM_{10}$-only data, derived in Beddows et al. (2015) and consisting of 6 sources, namely: Diffuse Urban; Marine; Secondary; Non-Exhaust~~;~~ Traffic/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, $Cl^-$ and $Mg^{2+}$, and the ~~s~~Secondary factor was identified from a strong association with $NH_4^+$, $NO_3^-$, $SO_4^{2-}$ and organic carbon. For the ~~t~~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 ~~n~~Non-~~e~~Exhaust ~~t~~Traffic/~~c~~Crustal factor explained a high proportion of the variation in the Al, $Ca^{2+}$ 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 ~~f~~Fuel ~~o~~Oil, characterised by a strong association with V and Ni together with significant $SO_4^{2-}$. This output comprised the first-step solution in the 2-step analysis of $PM_{10}$ 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 6 $PM_{10}$ factors.

## ~~2.2 Proxy Data~~

~~As alluded to in the introduction, experimental design can ensure that measurements cover the the same particle size range by using a single instrument such as a cascade impactor which gives size and mass information. However for our case, the NSD spanned anthe aerodynamic particle sizes 0.10 – 0.88 µm (i.e. $PN_{0.9}$) which is smaller than the range covered by the particulate mass, ≤10 µm, i.e. $PM_{10}$.~~

~~Besides the~~ ~~The value of~~ $PM_{10}$ ~~is given by the measured~~ $PM_{10}$ ~~mass, and t~~Thе $PM_{0.9}$ ~~mass is calculated by equation 3~~ 1 ~~from the NSD assuming estimates of PM mass can be derived using the NSD assuming spherical particles of a fixed density. For the SMPS settings, a particle size range between 16 and 604 nm is collected which can be~~and ~~was used to estimate a~~ $PM_{0.96}$ ~~value using equation 1~~.

$$PM_{0.96} = \rho_{eff} \times \frac{\pi}{6} \sum_{sizeBins} d^3 \quad (3\ 1)$$

~~w~~W~~here~~ $\rho_{eff}$ ~~is set to 2 g/cm³ (for the whole time series and across all size bins) for a Diffuse Urban background (based~~ ~~taken as the rounded midpoint of the rangeupon 1.8-2.5 g/cm³~~

~~for an urban background aerosol; Beddows et al., 2010). Due to the approximations used in this proxy variable, a large uncertainty was applied to it in the PMF analysis so not to affect the result.~~

~~Figure S1 plots the total apportioned $PM_{10}$ mass against the $PM_{0.96}$ 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 0.635, i.e. accounting for 635% 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 $PN_{0.6-10}$ to improve the match of the NSD matrix to the $PM_{10}$. This was appended as a column in to the input matrix of the PMF analysis with uncertainties taken as the 4 x $PN_{1.0-10_i}$ so not overly influence the analysis.~~

~~With the considertation of future measurements where this method wasis to be applied, an improved result might be possible if $PM_{1.0}$ was used in preference to $PM_{10}$ in order to give a better match between the size ranges of the mass and number data, if only, to close any uncertainty over the the issue. Another alternative is simply to include an APS in the measurements and merge the APS data to the SMPS data as in Beddows et al. (2010). This would provide a complete overlap between the NSD and $PM_{10}$ data.~~

## 2.~~3~~ 2 Methods

### 2.~~3~~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 *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 ~~4~~1:

$$x_{ij} = \sum_{k=1}^{p} g_{ik} \cdot f_{kj} + e_{ij} \qquad \text{i=1…n; j=1…m} \tag{~~4~~1}$$

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 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 *E* and the two matrices *G* and *F* are obtained by an iterative algorithm which minimises the object function Q (see equation ~~5~~2).

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

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{e_{ij}}{s_{ij}} \right)^2 \tag{~~5~~2}$$

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

~~It may be seen from equation (2) that~~ 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~~$s_{ij}$ , thus giving their variables a higher weighting in the analysis. Input variables with low weight ~~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, e.g. when ~~.~~ ~~When~~ apportioning total PM mass in a conventional one-step~~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. A similar approach can be followed in the PMF analysis of a combined data-set where higher weightings can be applied to the main data-set of interest such that it "drives" the analysis and the auxillary data set "follows", i.e. the uncertainties are chosen such that the balance of total weights from the two data sets is *tipped* towards the measurement of interest and highest reliabilit~~i~~y in regards ~~to~~of rotational unabiguity. ~~In this work, the primary aim was to define a particle size distribution associated with each factor derived from the PMF of PM10 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.~~

To assess the PMF mod~~a~~el, the ~~In this work, the~~ *Q* value is outputted by PMF~~2~~ 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 ~~$X_{scale}$~~ $b_{scale}$ until the ratio between Q and $Q_{theory}$ is

approximately one (rQ value = $Q/Q_{theory}$ =≈ 1 ± 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 scal~~in~~ing either G or F to unity such that the units from X are carried over to either F or G respectively to complete the apportionment. However, different routes have to be considered depending on whether X has homogen~~i~~eous or hetero~~gen~~i~~e~~ous units.

### 2.~~3~~2.2 *1-Step method using data ~~from the same instrument~~in the same units - ~~H~~homogeneous units*

Given a PMF input data matrix X, a solution GF + E can be computed where G represents the time series of the source profiles F, with a residual matrix E. Often X comprises ~~of~~ columns of $PM_{10}$ ~~composition values~~component concentrations (e.g. ICPMS values measured from ~~quartz filters~~acid-digested filters collected with a Partisol 2025) and it is common practice to also include a Total ~~V~~variable (e.g. column of $PM_{10}$, measured using a TEOM) in the data matrix. The resulting $PM_{10}$ profile element value can then be used to scale G and F such that G carries the units of X with F unit-less. Note that neither G or F is scaled to unity in thi~~es~~ approach. Instead, scaling is done after the analysis using a constant $a_k$, determined by the time series of a Total ~~V~~variable (e.g. $PM_{10}$), down weighted by ~~a factor of 4~~applying a high uncertainty, within the input data.

$$x_{ij} = \sum_{k=1}^{p} (a_k g_{ik}) \left(\frac{f_{kj}}{a_{\cancel{z}k}}\right) \qquad (\cancel{6}3)$$

The resulting value for the $PM_{10}$ contribution for each factor within the F matrix is then used as a scaling constant $a_k$ in equation ~~6~~3. Such scaling results in unit-less factors F which describe the characteristics of the sources and time series G with units of µg/m$^3$. Apportionment can then be carried out by averaging the G values for each source ~~profile~~factor, ~~and~~or a fully quantified time series of each factor can be presented, e.g. in Bivariate plots. Of course, the G and F can be normalized such that G is unitless and F carries ~~the~~ units; an approach necessary when X ~~contrain~~contains heterogeneous units. This approach however, requires each column of G to be scaled to unity, by using the PMF setting Mean lGl = 1.

**2.~~3~~2.3** *1-Step method using data ~~from different instruments~~with different units - ~~H~~heterogeneous units*

If the analysis of X was to be enhanced by the inclusion of data from a second instrument with different units, then a different approach to the *1-Step method with ~~H~~homogeneous units* would be required to analyse the joint data matrix [X,Z] = G[X,Z] F[X,Z] + E[X,Z]. If the previous method was applied where F was normalized, then it would not be clear what units to assign to G, whether the units from X or Z. To get around this problem, G is scaled to unity. This results in a unit-less time series G and a quantified F matrix. For each source profile the sum of the species associated with either data type gives the average total apportionment, e.g. of $PM_{10}$ or number concentration PN. Of course, this requires the complete mass or number closure of the elements making up either $PM_{10}$ or PN respectively, although inclusion of measurements of total $PM_{10}$ or PN can used instead, if available.

In the ideal case, if the individually computed factors for both data sets result in G(X) and G(Z) being identical, then a straight-forward joint model [X,Z] is successful and G[X,Z] =

G(X) = G(Z). However, if G(X) and G(Z) are significantly different then the joint model will fail, identified by a too large Q value. A solution to this problem is to set the total weights of the better dataset X significantly higher than the total weights of the auxiliary data set Z such that X will "drive the model" and G[X,Z] will be approximately equal to G(X) and a reasonable Q value is obtained for the Z. However, care is required to ensure that X or Z do not contain rotational ambiguity because such rotation for X may not be suitable for Z. For such cases, equal total weights for both X and Z are applied in the hope that the best rotation for both X and Z can be found.

### 2.~~3~~2.4 *2-Step method using data ~~from different instruments~~with different units - ~~H~~heterogeneous units*

The method proposed in this work~~,~~ separates the analysis of the two data sets X and Z into two different PMF analyses. Dataset X is first analysed and an unambiguous rotation is selected which gives computed factors G(X). These are then carried over into a second PMF step in which G(X) are combined with Z to form a joint matrix for analysis. By using FKEY (described below) factors, G(X,Z) are forced to be equal to G(X) from step 1. So for example, if in the first step we analyse $PM_{10}$ data and carry forward the output G($PM_{10}$) into a second step combined with the NSD data, i.e. [G($PM_{10}$),NSD] this results in profiles F[G($PM_{10}$),NSD]. In other words, we force out of the NSD data source profiles which have the same G factors as the $PM_{10}$ data and extend the list of components of the sources iden~~d~~tified in the first step and thus improve ~~identification~~characterisation of the source. Note that this is equivalent to non-negative weighted regression of matrix Z by columns of matrix G for which other tools exist. Furthermore, by using a two step method, we can continue to use the scaling method described in Section 2.~~3~~2.2 to apportion the sources using a quantified time series G(X) rather than normalising the G(X,Z) matrix sums to 1 and

relying on the summation of the elements in the rows of F(X,Z) to give the apportionment of X and Z.

### 2.~~3.2~~2.5. *Application of PMF*

Positive Matrix Factori~~s~~zation was carried out in this work using the DOS based ex~~e~~cutable file PMF2 v4.2 compiled by Pentti Paatero and released ~~in~~on Feb 11, 2010 (downloaded from www.helsinki.fi/~paatero/PMF/). This is used by the author ~~as a~~in preference to a GUI version of PMF (e.g. US EPA PMF 5.0, Norris et al., 2014)~~)~~ because of the ease with with it can be ~~encorported~~incorporated into a~~n~~ Cran R procedure script using shell commands, thus facilitating automation of the analysis and any optimisation. R-script can be written to manipulate and organise input data for PMF2, run PMF2, collect the output and produce the necessary output for consideration as text, table or plot. The main strength for this approach is to improve the repeatability and transfer~~a~~ence of a method between practitioners within our group.

The two step method is shown schematically in Figure 2. Matrix X yields fact~~r~~ors $^{1}G$ and $^{1}F$ in the first step. The timeseries $^{1}G$ matrix is carried through to the second step where it is combined with an auxiliar~~i~~y data set Z, to give the a step 2 input matrix [$^{1}G$ Z]. This in ~~tern~~turn is analysed to produce factors $^{2}G$ and $^{2}F$. In the current example, ~~it uses the PMF output of~~ the dataset of Beddows et al. (2015) is used as a starting ~~point~~ matrix X and comprises~~assumes that a PMF analysis~~ ~~of~~ the $PM_{10}$ chemical composition dataset. This yields timeseries $^{1}G$ and source profile $^{1}F$ 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 optimi~~m~~um solution after considering 3 to 8 factor solutions.

~~(Step One) has already been carried out and dealt with as in the previous study.~~ The normalised timeseries matrix $^{1}G$ from this analysis ~~were~~was combined with the NSD data - concurrently measured with the $PM_{10}$ data - to form the input matrix [$^{1}GZ$], for step 2. ~~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 $^{1}G1$ matrix, $^{1}\triangle G$ are transferred from the output of the first step and entered as input uncertainties for the second step. The hourly NSD data was aggregated into daily ~~samples~~values to match the daily $^{1}G$ factors outputted from the PMF analysis of the daily $PM_{10}$ data sampled. This reduced the data matrix down to 590 rows by 57 ~~52~~ columns ($^{1}G1$…$^{1}G6$, $NSD_{1}^{16nm}$…$NSD_{51}^{640nm}$) for which a we have a $Q_{theory}$ value of 29748 ~~30,916~~ for a 6 factor solution. For the NSD data, the uncertainties are taken as ~~X times~~ the NSD values multiplied by the value of an arbitrary parameter $b_{scale}$ ~~in order to be large and ensure that the PMF is driven by the G1 matrix~~ (see Figure 2). Initially, ~~this~~$b_{scale}$ was set to 4 to ~~inorder~~to ensure that the model was weighted such that it was driven by the $PM_{10}$ 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~~The~~ value of ~~X~~ the parameter $b_{scale}$ (typically, 0.2) was optimised in Cran R so that the ratio of $Q/Q_{theory}$ = 1 ± 0.0~~1~~2~~~ 1~~, indicating an relative percentage uncertainty in the region of 20%. In retrospect – by taking into account the decrease in reliability of the size bin~~s~~ 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 ~~over the lower and upper size bins from 5 % upto~~to a pre defined larger value, e.g. 50-%,~~-~~ over the lower and upper size bins. The uncertainties were entered directly into the model using PMF matrix T with U and V redundant.

~~Using equation 7 the uncertainies were entered directly into the model using via matrix T with U and V redundant.~~

$$s_{ij} = t_{ij} + u_{ij}\sqrt{|x_{ij}|} + v_{ij}|x_{ij}| \quad (7)$$

## 2.~~3.3~~2.6 ~~Fkey~~*Pulling down with GKEY and FKEY*

GKEY and FKEY~~key~~ are~~is a feature~~ matrices with the same dimensions ~~of~~as G and F respectively, ~~in~~ for incorporating *a priori* information into a PMF analysis. They are ~~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~~ GKEY and FKEY indicate~~s~~ the location of suspected zeros in source profiles $^{2}F$ or contributions $^{2}G$ (Figure S1~~2~~~~3~~). Since ~~here it is~~we are concerned with the profiles, this information is given in the form of integer values in an FKEY~~key~~. 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 [$^{1}G$ NSD], ~~-~~it is certain that only one~~-~~ unique contribution will be strong for each row of the profile $^{2}F$, outputted from the second PMF analysis, ~~PM G score from one of the sources will be strong,~~ e.g. only $^{1}G1$ and not $^{1}G2$.. $^{1}G6$ will contribute the to ($^{1}G1$, $^{2}F_{1}$) position in output factor $^{2}F_{1}$. ~~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 S1~~2~~~~3~~). All 'non-zero' elements within the output of $^{2}F$ take a FKEY value of zero whereas all elements of $^{2}F$ which are pulled to zero take an non-zero value of $fkey_{1}$. This leads to a FKEY matrix which can be understood in two parts. The first part is a square matrix of dimension equal to the number of columns of $^{1}G$ with all its entries equal to $fkey_{1}$ except for the leading diagonal; this part ensures that $^{1}G$ is the same as $^{2}G$. The second part of the matrix consist of all the elements ~~are~~as zero and represents the NSD

input data. An $fkey_1$ 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 $^1G \approx {}^2G$.

To extend the analysis from 6 factors to 7 factors an extra row was added to FKEY. This was done in order to investigate any factors missed in the NSD data ~~for~~ which the first analysis using $PM_{10}$ would not be sensitive to. For example, a nucleation mode would be detected in NSD data but not $PM_{10}$ data. In order to give the model freedom to factorise out a nucleation factor, the 7th row of of FKEY values consisted {$fkey_1$, $fkey_2$… $fkey_6$,$nsd_1$, $nsd_2$… $nsd_{51}$}. This ensured that all the $^2G$ contributions were allocted to the first 6 factors only leaving the 7th factor to account for the remaining unfactorised NSD data. There is no reason why more than 7 factors could not be used to investigate ~~to see if there are more~~possible ~~un-resolved~~unresolved NSD factors. However, we constrained the scope of our investigation to reidentifying those in Figure 1~~2~~.

## 2.~~4~~ 3 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 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 to using PMF in the second step, a regression was carried out. E~~e~~ach column of data for each of the 51 size bins *j* within the NSD was regressed against the six $^{1}G$~~1~~ time series using~~,~~ Equation ~~8~~34~~.~~

$$NSD_j = \alpha_{0,j} + \alpha_{1,j}\,{}^{1}G_1 + \alpha_{2,j}\,{}^{1}G_2 + \cdots + \alpha_{6,j}\,{}^{1}G_6 \qquad (8\ 4)$$

where $\alpha_0$ is the population intercept and $\alpha_{1\text{-}6}$ are the populations slope coefficents. 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 6 of the factors (giving an indication of how each size bin ~~correlates~~ scales with each of the 6 factors) and an intercept. When $\alpha_{1\text{-}6,j}$ ~~$grad_{a,j}$~~ is plotted against the size bin, 6 plots showing the dependence of each size bin *j* on each of the 6 $PM_{10}$ factors are produced. It is also assumed that these (~~we refer to~~ referred to here as NSD regression source profiles) will be comparable to the actual NSD PMF source profile. Similarly, the $\alpha_{0,j}$ ~~$int_j$~~ values are expected to give a background value due ~~to,~~ possibly to noise; however, it is more likely to yield a source (such nucleation) to which the $PM_{10}$ mass analysis is insensitive. ~~However, this method can extract information known as a remainder factor, shown later in this paper.~~

## 2.~~5~~ 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 ~~indeed~~ cause loss of information implying that minor factors (e.g. due to event episodes) might well be averaged out of the data. ~~Given the $i^{th}$ number size distribution, $NSD_i$, 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.~~ Given the $j^{th}$ size bin in the $i_{th}$ number size distribution $NSD_{i,j}$ (of dimensions M x N), the factors can be fitted using equation (~~9~~5).

$$D_i = \sum_{i=1}^{M} d_i \qquad (\cancel{9}5)$$

which is the $i_{th}$ sum $D_i$ of the difference ($d_i$ give by equation 6) across the size bins of the $i_{th}$ $NSD_i$ and the linear ~~superposition~~sum of the $p$ ~~factors~~NSD source profiles ($p$ = 7 in this case) scaled ~~by~~with respect to the scalar values $c_{ik}$, representing the timeseries of each fitted NSD source profile~~as shown by equation 10~~.

$$d_i = \sum_{j=1}^{N} \left\{ NSD_{ij} - \sum\nolimits_{k=0}^{p} c_{ik} \times f_{kj} \right\}, \qquad c_{ik} \geq 0$$
$$1 \times 10^{10}, \qquad c_{ik} < 0 \qquad (\cancel{10}6)$$

The Cran R package Non-Linear Minimization (nlm) (R Core Team, 2018) was used to minimise ~~equation~~ the value of $D_i$ with respect to the scalar values $c_{ik}$ ~~3.~~ ~~A~~ with a non-negative constraint ~~is~~ on $c_{ik}$ placed in the function. If a negative value is returned by any of the ~~$a_k$~~ $c_k$ values then $D$ 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~~ scalar $\beta_k$. ~~Six~~ Seven values were derived for $\beta_k$ by regressing the total particle number (total hourly SMPS) against each of the fitted values $c_k$ (~~using~~ equation ~~11~~9). ~~7~~ 7.

$$PN = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_7 c_7 \qquad (\cancel{11}7)$$

The resulting scaled-fitted values were then used to calculate the PN concentration for each of the regression source profiles (equation ~~12~~ 8) ~~and plot~~allowing subsequent plotting of the 7 diurnal cycles ~~in Figure 5~~.

$$PN_k = \beta_k c_k \qquad (\cancel{12}8)$$

## 2.~~6~~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 m/s,... and 0-10, 10-20 degrees 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 ~~background~~diffuse source will tend to have its highest concentration ~~yeild~~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 recognised by their ~~proportionality~~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 visualising how concentrations of pollutants vary by wind direction and hour of the day. For example, for the North Kensington site – positioned West of the ~~central~~ city centre – we might well expect most of the anthropogenic sources (traffic, diffuse urban, etc) to ~~have~~show an Easterly direction with the appropriate diurnal cycle (e.g. ~~yeilding~~ rush hour traffic patterns). Similarly, we might expect cleaner air (Marine, Nucleation, etc) to occur from a Westerly direction and ~~during~~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~~, given~~ a given PMF result~~, it~~ can be complemented with concurrently measured auxillary data. We exemplify this using $PM_{10}$ and NSD data collected from the North Kensington receptor site in London and start with the premise that we are completely satisfied with the $PM_{10}$ analysis and are using a rotation which gives quantified factors (quantified G and scaled F) which best represent~~ation of~~ the urban atmosphere sampled, i.e. the output from Beddows et al. (2015). ~~And, f~~For each $PM_{10}$ factor we wish to assign a NSD distribution. ~~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.~~ Rather than repeat the PMF analysis using a combined $PM_{10}$+NSD ~~analysis~~dataset which can be complicated if the rotations of the individual PMF analyses of $PM_{10}$ and NSD data are mis~~s~~-matched or ambiguous, we can carry out a ~~This can be done using~~ a second PMF analysis or a regression.

Furth~~t~~ermore, by the nature of any factor analysis, we also have to make the assumption that each source chemical profile and size distribution not only remain unchanged between

source and receptor but that it remains 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. ~~We might well~~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 ~~indeed~~ have different atmospheric transit ~~sizes~~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 ~~nature~~composition of ~~the~~an air mass measured over a long period (e.g. 2 years) dataset ~~(eg 2 year)~~ using PMF.

### 3.1 2-Step PMF-PMF Analysis

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

The ~~G1~~ time series $^{1}G_k$ and uncertainties $^{1}\triangle G_k$ from the first PMF analysis of $PM_{10}$ data we~~a~~re carried over into the second step where they are combined with the NSD data for PMF analysis (Figure 2). The uncertainties of the NSD data are taken as an optimised multiple of the NSD values themselves (~ 5 % uncertainty, y~~ei~~ielding a Q value of 30,333 in the robust mode; see Table S2 for PMF settings). Also in order to ~~maintain the solution from step 1 in step 2 the~~ encourage $^{2}G_k$ to be proportional to $^{1}G_k$ for $k$ = 1-6 (see Table S4), the

FKEY~~Fkey~~ matrix is applied to pull elements in the source matrix to zero as described in section 2.3.3. This ensured~~s~~ that the PMF analysis of the NSD data wa~~i~~s driven by the $^{1}G$~~G1~~ time series and ~~. This~~ resulted~~s~~ in a 6 factor output~~solution~~ in which there ~~-~~we~~a~~re unique contributions from the $k_{th}$ factor $^{1}G_{k}$ from the first analysis to the $k_{th}$ factor ~~from one of the G1~~ $^{2}F_{k}$ ~~scores and an~~ in the second analysis. This is mainly due to the aggressive pulling of the factor element in $^{2}F$ applied using FKEY.

~~associated NSD source profile, and it is~~ When inspecting Figure 3 it is notable that ~~they~~ the source profiles 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~~ same 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. The Fuel Oil factor ~~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 S23~~2~~ shows the results of the linear regression of the NSD data plotted against the $PM_{10}$ $^{1}G_{k}$~~1~~ scores and again what is remarkable is the similarity between these regression source profiles~~correlation plots~~ and both the factors derived in Beddows et al. (2015) and those from the 2-step PMF-PMF analysis.

This PMF-LR analysis was carried out using daily averaged data and ~~. T~~to obtain hourly information - and thus obtain the diurnal patterns (Figure S2~~3~~5) -~~,~~ the resulting regression source profiles ~~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 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 $PM_{10}$ 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~~ regression source profile, 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. 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 6 factors to 7 factors, ~~and adding~~ an extra row in the FKEY ~~key~~ matrix was added to ~~which~~ pull~~s~~ all of the ${}^{1}G_{7}$~~1~~ ~~scores~~ contributions to ${}^{2}F_{7}$ to zero in the solution (Figure S1~~2~~).~~,~~ The same FKEY matrix of *fkey*$_1$ and 0 values was used but this time it was augmented with a 7$^{th}$ row of *fkey*$_2$ and zero

values. In this case, the *fkey*$_2$ values were set to a value of 20.

T~~t~~he same 6 factor solution is obtained with the additional 7$^{th}$ factor (Figure 4 and Figure S3~~3~~~~4~~) and ~~. A~~as expected, this seventh factor was~~is~~ a Nucleation factor. It was~~is~~ suspected that in the 6 factor solution, the ~~n~~Nucleation factor was combined with the Fuel-Oil factor. This does not suggest any link between the ~~n~~Nucleation and Fuel-Oil factor other than there ~~were~~ was an insufficient number of factors within the model for the two to factorise out of the data giving the Fuel-Oil NSD profile a more reasonable modal peak between 50 and 60 nm rather than 20, 30~~10~~ and 60 nm.

~~In the results of~~ Beddows et al. (2015), applied a 1-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 homogen~~i~~eous units - produced quantitative timeseries G. This was unlike the analysis of the $PM_{10}$+NSD with heterogen~~i~~eous units which could not apportion its 5 factors using G but was able to factorise out a ~~the~~ Nucleation factor from the data, seen also in the 4 sources in the PMF solution for the NSD-only data. ~~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 $PM_{10}$-only seven factor solution ~~to PMF of the $PM_{10}$ chemical composition data~~ did not reveal this factor ~~either~~, presumably because the mass associated with nucleation mode particles is too small to affect composition significantly. Furthermore, Fuel Oil was not factorised out of the $PM_{10}$+NSD data and was more likely divided across all 5 factors.

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 $PM_{10}$ 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). Also note~~,~~ that there is no reason why any further investigation might not explore using more than 7 factors. In fact~~or~~ the ~~n~~Nucleation factor appears at first ~~glance~~sight to be multimodal. However, we restricted our analysis to 7 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 ~~required~~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 Section 2.~~5~~4). Figure 5 shows the resulting diurnal profiles. The diurnal trends of the parameter $c_k$ (equation ~~1~~17), required ~~fitted peaks show the values required in equation 3~~ to fit the 7 daily NSD factors to the hourly NSD data are shown. These have been scaled to PN (measured in $1/cm^3$)~~ in these plots according to~~ using the integral of the NSD (equation ~~1~~28)~~factor measured in $1/cm^3$~~. The ~~n~~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 all follow a trend which is synchronised 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. ~~also follows a similar anthropogenic cycle, however, although the polar plots are comparable to the and would be expected to be strongest at night those of Beddows et al 2015, the nighttime contributions is small. This results a contribution not being strong during the night in the diurnal trend plot of Figure 5 as would be expected when compared to the NSD diurnal trend of Beddows et al 2015.~~ Fuel Oil is highest during the evening and night and may correspond to home heating rather than ~~marine~~ shipping emissions~~activity~~. 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 μm in the Marine factor. As pointed out in Beddows et al. (2015), the ~~m~~Marine factor is ~~interesting by way of the fact that we would indeed expect to see the wing of a coarse particle mode, whose modal diameter is way above the upper size bin of the SMPS and in the coarse mode. Instead, the factor is~~ identified by its chemical profile of sodium and chlo~~i~~ride and is accompanied by an aged nucleation mode at around 30nm. This can be either viewed simply as clean marine air being 'polluted' by traffic emission and/or as the consequence of nucleation occuring over at city ~~for~~in ~~a~~ clean maritime air masses (Brines et al. 2015). The key point here is that the factors derived in this work are comparable to those factorised in Beddows et al. (2015) using the combined data-set and the advantage of the 2-step approach is that now we have quantified hourly timeseries G.

The hourly contributions are aggregated into daily values and plotted as bivariate plots in Figure 5~~7~~ to assist comparison with the daily plots in Beddows et al. (2015). In that work, the same PMF analysis of the NSD data yielded 4 factors which are ~~represented here again~~named identically to those in the bivariate plots. The similarity of both of the polar and annular plots for each of the 4 factors ~~justifies~~ supports our ~~aformentioned~~ previous factor~~-fitting assumption~~ identification. The Secondary and Diffuse Urban 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 expecting to be most dense. Nucleation is also seen to be strongest for those wind direction from the West which are expected to be cleaner, and have a lower condensation sink. NET & Crustal and Fuel Oil are similar to Diffuse Urban suggesting a similar predominant source location in the centre of London. Marine 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~~of~~ the Nucleation ~~Hidden~~ Factor

The Nucleation factor was extracted from the two-step PMF-PMF analysis ~~when~~ which included pulling the $^{1}G_{1}$-$^{1}G_{6}$ to zero of factor $^{2}F_{7}$. ~~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. For this, the time series of the 4 NSD factors ($^{1}G_{1}$-$^{1}G_{4}$) reported in Beddows et al. (2015) were combined with the $PM_{10}$ data. We again assume that the first PMF step has

been carried out and that we are satisified with how the final solution represents the urban environment of the receptor site and that there are no rotational ambi~~t~~guities. We then carry out the second step PMF analysis on the 34 x 591 input matrix ([$^1$G1…$^1$G4], $PM_{10}$[PM,$PM_{carbon}$,$PM_{ions}$,$PM_{metals}$]). The hourly output uncertainies from the first PMF analysis of the NSD data $^1\Delta$G1…$^1\Delta$G4 were carried forward into the second PMF analysis by adding them *in quadrature* to give daily uncertainties. As with the analysis of the auxillary data in the $PM_{10}$-NSD data, the measurement uncertainties for the $PM_{10}$ data (this time the auxillary data) was naively taken as 4 times the $PM_{10}$ matrix. Extra care could have been take in assigning the $PM_{10}$ uncertainties but since we force the output using FKEY a simpler approach was taken. In fact, the FKEY consisted of a 4 x 4 diagonal matrix of zer~~s~~o values with an $fkey_1$ of 20 for all the off-diagonal positions joined to a 4 x 30 matrix of zeros. Furthermore, the uncertainty values of the $PM_{10}$ were scaled until $Q/Q_{theory}$ = 0.99 using parameter $b_{scale}$ = 0.35 (see Table S3 for more details).

Ideally, ~~for this~~ the chemical data would be ~~more informed with regards~~limited to the composition of the particles ~~below 100 nm (eg using PM0.1 or PM1.0)~~in the same size range as the SMPS data. However, when since we are using the $PM_{10}$ composition data we can at best describe the composition of the aerosol which accompanied each factor (Figure S4~~5~~). For the NSD Secondary ~~NSD~~ factor with its strongest contribution (indicated by the ~~Expected~~Explained Variation) ~400 nm, we have a strong contribution to $PM_{10}$ and $PM_{2.5}$ ~~contribution~~ together with nitrate, sulphate and ammonium. Diffuse Urban, with its strongest contribution at 100 nm is accompanied by contributions ~~to~~from elemental carbon and wood smoke indicative of traffic and recreational wood burning. There ~~is~~are also contributions from barium, chromium, iron, ~~M~~molybdenum, ~~A~~antimony and ~~V~~vanadium, all indicative of non~~e~~-exhaust traffic emissions and the burning of fuel oil. Similarly, the ~~t~~Traffic factor has a modal

-diameter at roughly 30 nm which is indicative of exhaust emissions and this is accompanied by contributionsn to aluminum, barium, calcium, copper, iron, manganese, titanium and various other metals attributed to vehicles, albeit from tyre or brake wear or resuspension.

For tThe Nucleation factor with its peak ~20 nm, this was associated with marine air withas indicated by the strong contributions to Na, Cl and Mg (Figure S4S4). . There are also traces of V, Cr, Ni and a high contribution to $PM_{10}$ level 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 which is 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 Figure S34. There are also strong contributisons to vanadium which is most likely from an un-resolved Fuel Oil source being mixed into the mMarine and dDiffuse uUrban factors. 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, t~~

A two-step PMF analysis method is presented whereby existing PMF profiles can be extend to incorporate auxillary data con~~d~~currently measured and having different units. This is exemplified using $PM_{10}$ and NSD data.

When analysing $PM_{10}$ data, the inclusion of auxillary 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 ~~ambibuity~~ambiguity in either the $PM_{10}$ data or in the auxillary data. In the ideal case, the individually computed factors G(X), G(Z) and G(X,Z) need to be similar if the joint model is to be successful and not produce ~~to~~ large residuals and hence a too large ~~a~~ Q value. In the best case, the total weight of the $PM_{10}$ data can be set higher than the auxillary data so that the $PM_{10}$ data drives the analysis. In this work, we present an alternative method called the 2-step PMF method. ~~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 omitted. This problem is overcome in this work by using a novel Two-Step PMF approach.~~ In the first step the $PM_{10}$ data is 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 G (and errors) of the $PM_{10}$ 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 G time series and NSD dataset. In order to ensure that unique factors are obtained for the G scores, ~~Fkey~~ 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 $PM_{10}$ solution and the $PM_{10}$ analysis is not affected

by any rotational ambiguity of the NSD data. For our demonstration using the Beddows et al. (2015) analysis, ~~T~~this results in 6 $PM_{10}$ factors ~~which are not only~~whose time series are not only apportioned in mass but the source profiles are ~~augmented by~~identified for the NSD data. Comparisons of both 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~~d~~ solutions for $PM_{10}$ and NSD data datasets together. This generates confidence that the NSD and $PM_{10}$ 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 $PM_{10}$ mass can be identified as in the case of the Nucleation factor. This information can also be extracted using a linear regression PMF-LR where the size bins of the NSD data are regressed against the $PM_{10}$ 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 6 to 7.

## 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


Beccaceci, S., Mustoe, C., Butterfield, D., Tompkins, J., Sarantaridis, D., Quincey, D., Brown, R., Green, D., Grieve, A., Jones, A.: Airborne Particulate Concentrations and Numbers in the United Kingdom (phase 3), Annual Report 2011, NPL Report as 74, 2013a, https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1306241448_Particles_Network_Annual_Report_2011_(AS74).pdf.

Beccaceci, S., Mustoe, C., Butterfield, D., Tompkins, J., Sarantaridis, D., Quincey, D., Brown, R., Green, D., Fuller, G., Tremper, A., Priestman, M., Font, A. F., Jones, A.: Airborne Particulate Concentrations and Numbers in the United Kingdom (phase 3), Annual Report 2012, NPL Report as 74, 2013b, https://uk-air.defra.gov.uk/assets/documents/reports/cat05/1312100920_Particles_Network_Annual_report_2012_AS_83.pdf.

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.

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, 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, 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.

~~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.~~

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. Tot. 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 10 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 $PM_{2.5}$ 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 compositon data, Atmos. Environ., 40, S396-S410, 2006.

Paatero, P.: User's Guide to Positive Matrix Factorization Programs PMF2 and PMF3, Part 2, 2002.

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, 1998, doi:10.1029/98JD01212.

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/.

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, D.: Source apportionment of ambient particle number concentrations in central Los Angeles using positive matrix factorization (PMF), Atmos. Chem. Phys., 16, 4849-4866, 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 $PM_{2.5}$ 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:**

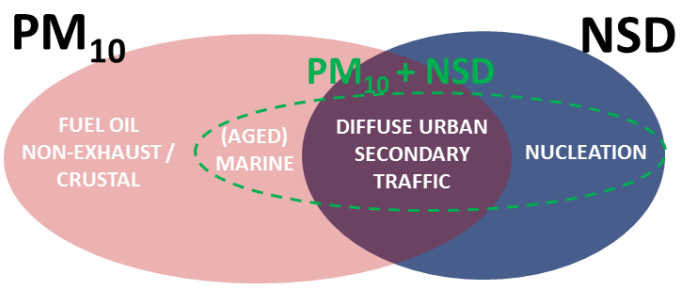
**Figure 1.** Venn Diagram showing the summary of the findings of Beddows et al. (2015); applying PMF to $PM_{10}$-only, NSD-only and $PM_{10}$+NSD datasets. Table shows the apportionment of $PM_{10}$ and NSD taken from Beddows et al. (2015).

**Figure 2.** Flow diagram showing the flow of data through the 2-step PMF-PMF analysis. The PMF analyses of single data set X are considered in step 1 and output indicated by factors/uncertainties $^1G$, $^1\triangle G$, $^1F$ and $^1\triangle F$. The second PMF analysis is carried out on the joint data set [$^1GZ$] and yie~~i~~lds factors/uncertainties $i^2G$, $^2\triangle G$, $^2F$ and $^2\triangle F$. In our analysis, X and $^1G$ are the $PM_{10}$ and resulting time series from the analysis of Beddows et al. (2015) and Z is the auxillary NSD data concurrently measured using a SMPS.

**Figure 3.** Source profiles $^1F$ and $^2F$ from both the first and second PMF step using 6 factors. [Grey bars and black line indicate~~s~~ the values of F; red lines and dots indicate~~d~~ the explained variations; and grey dotted line indicates the dV/dlogDp.].

**Figure 4.** Nucleation and Fuel Oil factors derived when extending the second PMF analysis from the 6 factors (shown in Figure 3) to ~~allow for a~~ 7~~th~~ factors. Source profiles $^2F_1$ to $^2F_6$ are given in Figure S3~~4~~. Each plot is divided into 2 showing the output $^1F_k$ and $^2F_k$. [Grey bars and black line indicate~~s~~ the values of F; red lines and dots indicate~~d~~ the explained variations; and grey dotted line indicates the dV/dlogDp.]

**Figure 5.** Diurnal cycles derived $PN_k$ calculated by the fitting of the daily PMF factor profiles to the hourly NSD data fitted (see equation ~~12~~ 8 and ~~s~~Section 2.~~5~~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 acording 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 represent wind direction at an hour of the day between 0 and 23 hrs.]. Note that the diurnal plots do not start at zero.




| | $PM_{10}$ [µg/m³] | NSD [1/cm³] |
|---|---|---|
| ~~Urban Background~~Diffuse Urban | 4.1 | 2370 |
| Marine | 2.6 | - |
| Secondary | 4.4 | 243 |
| NET / Crustal | 4.3 | - |
| Fuel Oil | 1.0 | - |
| Traffic | 0.8 | 2460 |
| Nucleation | - | 430 |
| Total | 17.2 | 5512 |

**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. Table shows the apportionment of $PM_{10}$ and NSD taken from Beddows et al. (2015).

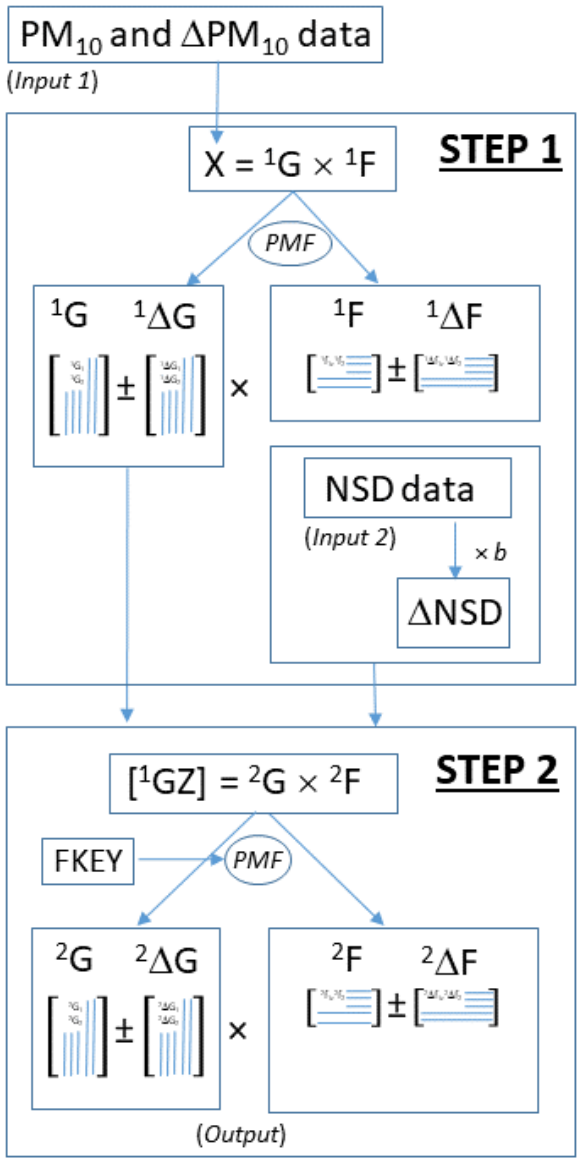
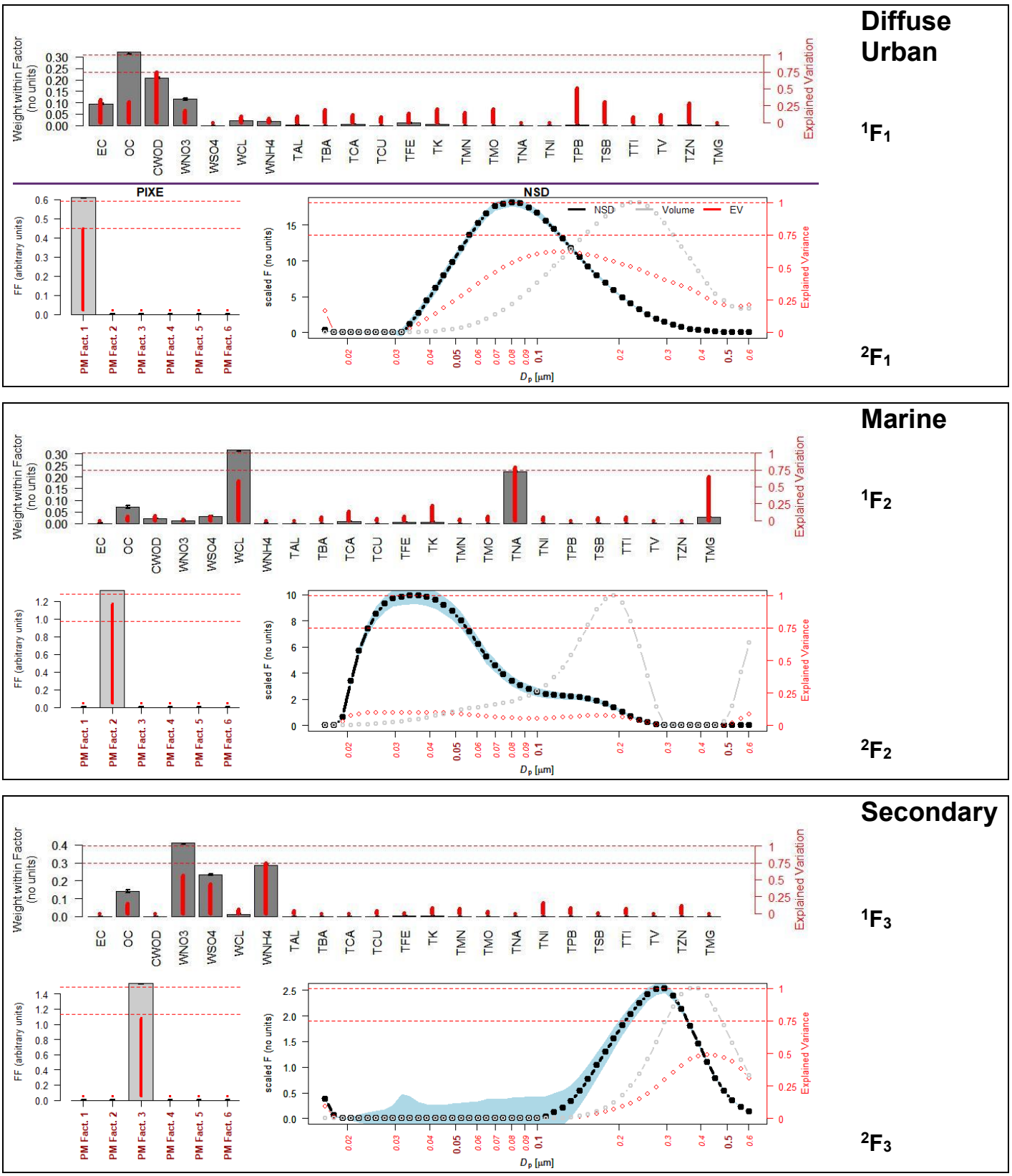



**Figure 2.** Flow diagram showing the flow of data through the 2-step PMF-PMF analysis. The PMF analyses of single data set X are considered in step 1 and output indicated by factors/uncertainties $^{1}G$, $^{1}\triangle G$, $^{1}F$ and $^{1}\triangle F$. The second PMF analysis is carried out on the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


Diffuse Urban
$^1F_1$
$^2F_1$
Marine
$^1F_2$
$^2F_2$
Secondary
$^1F_3$
$^2F_3$
Weight within Factor (no units)
Explained Variation
EC OC CWOD WNO3 WSO4 WCL WNH4 TAL TBA TCA TCU TFE TK TMN TMO TNA TNI TPB TSB TTI TV TZN TMG
PIXE
FF (arbitrary units)
PM Fact. 1 PM Fact. 2 PM Fact. 3 PM Fact. 4 PM Fact. 5 PM Fact. 6
NSD
scaled F (no units)
Volume
EV
Explained Variance
$D_p$ [µm]

1007

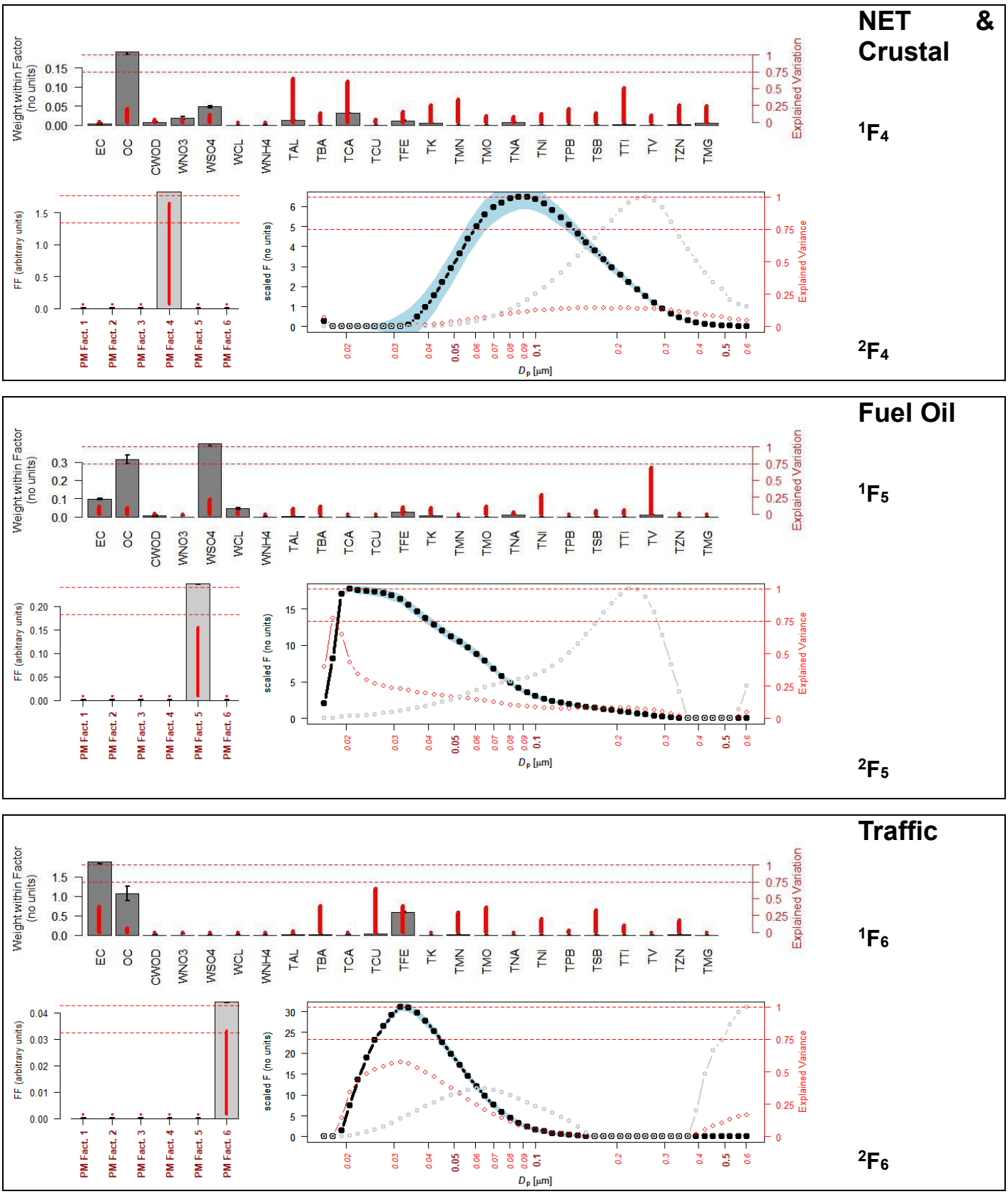



**Figure 3.** Source profiles $^1F$ and $^2F$ from both the first and second PMF step using 6 factors. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variations; and grey dotted line indicates the dV/dlogDp.]

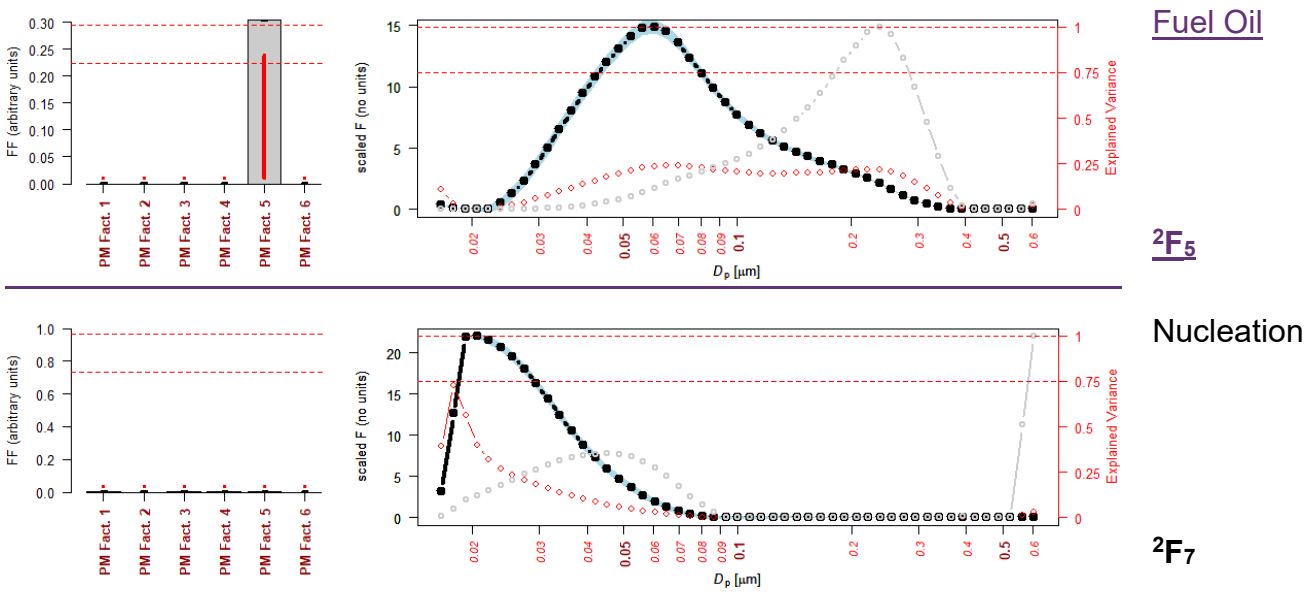



**Figure 4.** Nucleation and Fuel Oil factors derived when extending the second PMF analysis from the 6 factors (shown in Figure 3) to ~~allow for a~~ 7th factors. Source profiles $^2F_1$ to $^2F_6$ are given in Figure S~~3~~4. Each plot is divided into 2 showing the output $^1F_k$ and $^2F_k$. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variations; and grey dotted line indicates the dV/dlogDp.]

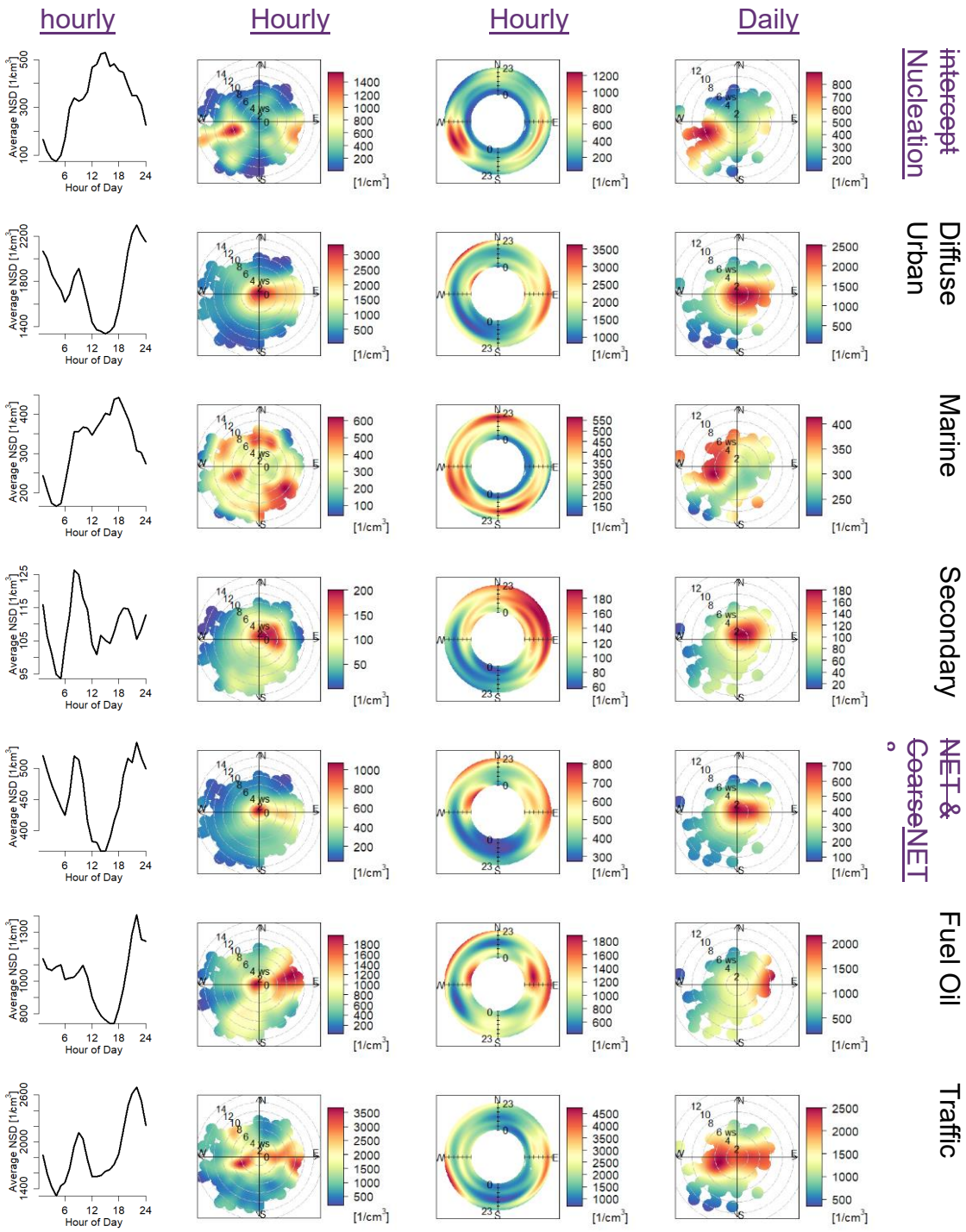



**Figure 5.** Diurnal cycles derived $PN_k$ calculated by the fitting of the daily PMF factor profiles to the hourly NSD data fitted (see equation ~~12~~ 8 and ~~s~~Section 2.~~5~~4). [Left-left column – diurnal trends of $PN_k$; left-middle column – bivariate plot of $PN_k$; middle-right – annular plot $PN_k$; right-right – bivariate plot of $PN_k$, plotted using the Openair program. Polar plots show a point coloured acording 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 represent wind direction at an hour of the day between 0 and 23 hrs.]. Note that the diurnal plots do not start at zero.

1019

# SUPPLEMENTARY INFORMATION

## 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**

**Table S1.** Setup of TSI SMPS.

| EC 3080 | |
|---|---|
| Neutraliser | Kr-85 radioactive source |
| Drier | ✓ EUSAAR/ACTRIS Drier |
| DMA | TSI 3081 long DMA |
| Aerosol Flow | 0.3 lpm |
| Sheath Flow | 3.0 lpm |
| Impactor Type | 0.0508 cm |
| HV Polarity | Neg |
| AIM version | 9.0 |
| Scans per Sample | 6 |
| Number of Samples | 1 |
| Total Sample Time | 14 min 0 sec |
| Multiple charge | ✓ |
| Diffusion loss correction | ✓ |
| Particle Density | 1.2 g/cc |
| Gas Density | 0.0012 g/cc |
| Nano Aggregate Mobility Analysis | × |
| CPC3775 | |
| ~~CPC3775 li~~nlet flow | 0.3 lpm |
| Data Coverage | 72.5 % over the 2 years 2011/2012 |
| Service and Calibration Date | February 2011 and February/March 2012 |

**Table S2.** Miscellaneous PMF-PMF details for the $PM_{10}$-NSD data set.

| INPUT DATA | ($^1$G1…$^1$G6, $NSD_1^{16nm}$…$NSD_{52}^{640nm}$) |
|---|---|
| *Input Settings* | |
| PMF2 version number | 4.2 |
| Number of Factors | 6 |
| FPEAK | 0.1 |
| Input dimensions: Row x Columns | 590 x 58 |
| Number of Repeats | 1 |
| Outlier Distance | 4 |
| Robust Analysis | ✓ |
| Error Model | -12 |
| Seed | 3 |
| Initially Skipped | 0 |
| Uncertainty Matrices T/U/V | ✓/×/× |
| Normalization of factor vectors before output | None |
| Optional parameter lines | missingneg 10 |
| | |
| *Output values* | |
| Q in the robust mode | 30333 |
| Q when not down weighting outliers | 32568 |
| POS-Outlier limit (4.0) exceeded by | 221 positive residuals |
| NEG-Outlier limit (4.0) exceeded by | 38 negative residuals |

**Table S3.** Miscellaneous PMF-PMF details for the NSD-$PM_{10}$ data set.

| INPUT DATA | ([$^1$G1…$^1$G4],$PM_{10}$[PM,$PM_{carbon}$,$PM_{ions}$,$PM_{metals}$]). |
|---|---|
| *Input Settings* | |
| PMF2 version number | 4.2 |
| Number of Factors | 4 |
| FPEAK | 0.1 |
| Input dimensions: Row x Columns | 591 x 34 |
| Number of Repeats | 1 |
| Outlier Distance | 4 |
| Robust Analysis | ✓ |
| Error Model | -12 |
| Seed | 3 |
| Initially Skipped | 0 |
| Uncertainty Matrices T/U/V | ✓/×/× |
| Normalization of factor vectors before output | None |
| Optional parameter lines | missingneg 10 |
| | |
| *Output values* | |
| Q in the robust mode | 17652 |
| Q when not down weighting outliers | 18089 |
| POS-Outlier limit ( 4.0) exceeded by | 19 positive residuals |
| NEG-Outlier limit ( 4.0) exceeded by | 3 negative residuals |

**Table S4.** Summary of the regression results, comparing $^1G$k with $^2G$k for *k* in 1 to 6.

| | | *const* in $^1G$k = *const* x $^2G$k | $R^2$ |
|---|---|---|---|
| Diffuse Urban | $^1G_1$ vs $^2G_1$ | 1.2 | 0.72 |
| Marine | $^1G_2$ vs $^2G_2$ | 0.73 | 0.94 |
| Secondary | $^1G_3$ vs $^2G_3$ | 0.56 | 0.71 |
| NET & Crustal | $^1G_4$ vs $^2G_4$ | 0.54 | 0.96 |
| Fuel Oil | $^1G_5$ vs $^2G_5$ | 2.9 | 0.41 |
| Traffic | $^1G_6$ vs $^2G_6$ | 15.5 | 0.40 |

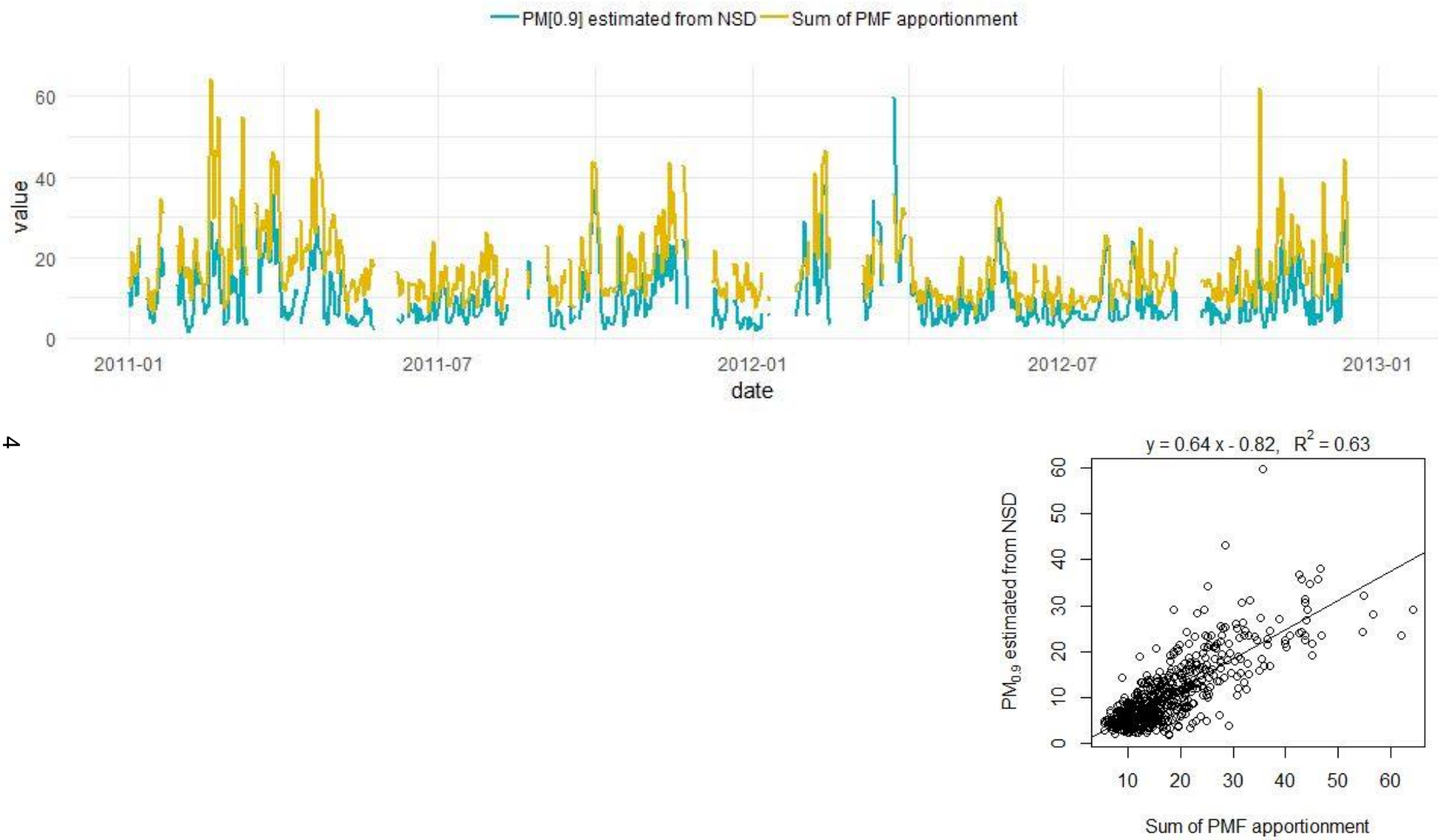



**Figure S1.** Plot of total apportioned $PM_{10}$ mass against $PM_{0.9}$ mass estimated using the NSD data.

| Factors from Step 2 | G Time Series from Step 1 | | | | | | Number Size Distribution (nm) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $^{1}G1$ | $^{1}G2$ | $^{1}G3$ | $^{1}G4$ | $^{1}G5$ | $^{1}G6$ | 16.6 | 17.8 | 19.2 | ... | 604 |
| $^{2}F_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_2$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_3$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_4$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_5$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_6$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | 0 | | 0 |

| Factors from Step 2 | G Time Series from Step 1 | | | | | | Number Size Distribution (nm) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $^{1}G1$ | $^{1}G2$ | $^{1}G3$ | $^{1}G4$ | $^{1}G5$ | $^{1}G6$ | 16.6 | 17.8 | 19.2 | ... | 604 |
| $^{2}F_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_2$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_3$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_4$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_5$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | $fkey_1$ | 0 | 0 | 0 | | 0 |
| $^{2}F_6$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | $fkey_1$ | 0 | 0 | 0 | 0 | | 0 |
| $^{2}F_7$ | $fkey_2$ | $fkey_2$ | $fkey_2$ | $fkey_2$ | $fkey_2$ | $fkey_2$ | 0 | 0 | 0 | | 0 |

**Figure ~~S2~~S1.** Entries in the FKEY matrix used in step 2 of the PMF-PMF analysis using (a) 6 factors and (b) 7 factors. An extremely strong value of 24 was chosen for $fkey_1$ and 20 for $fkey_2$.

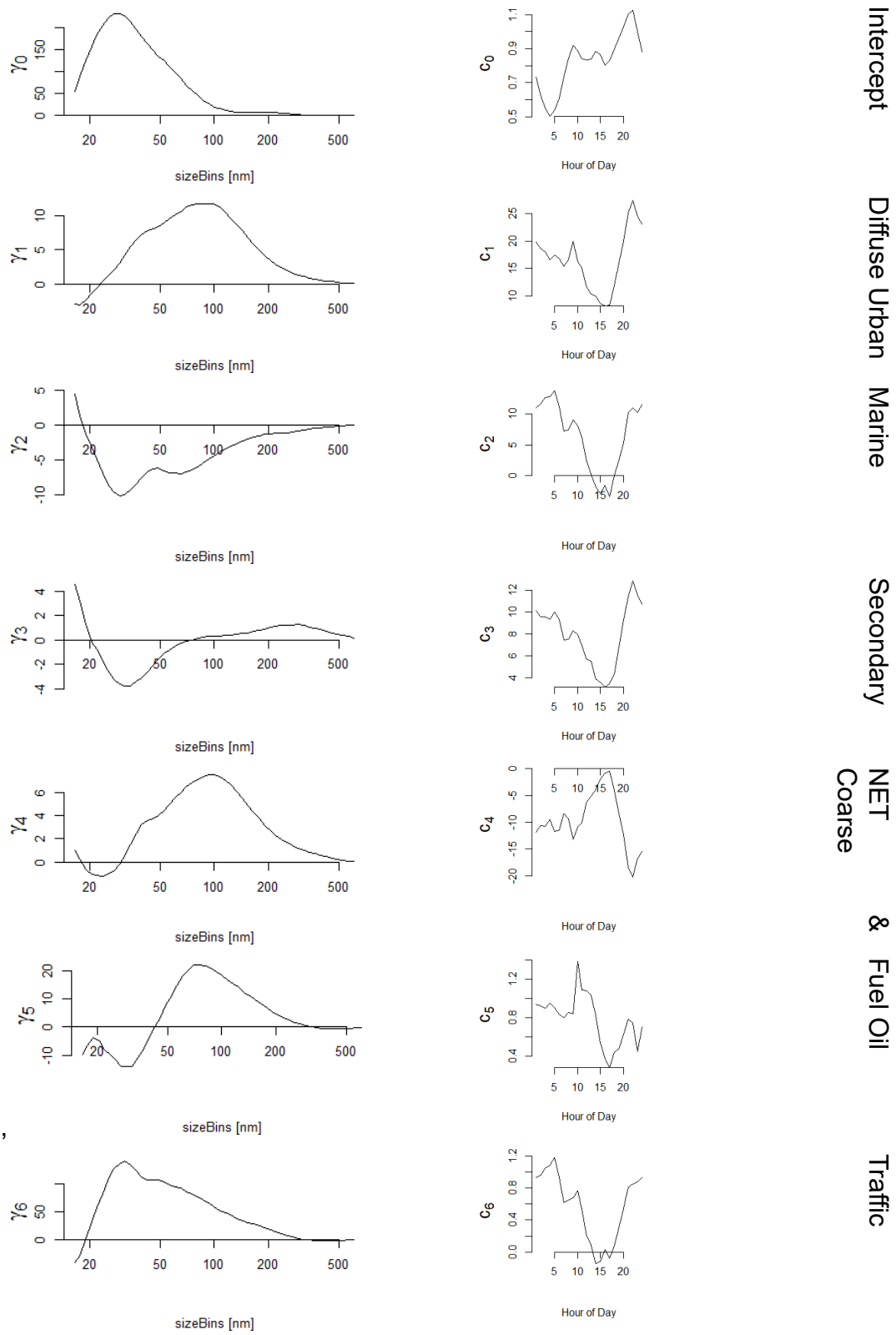
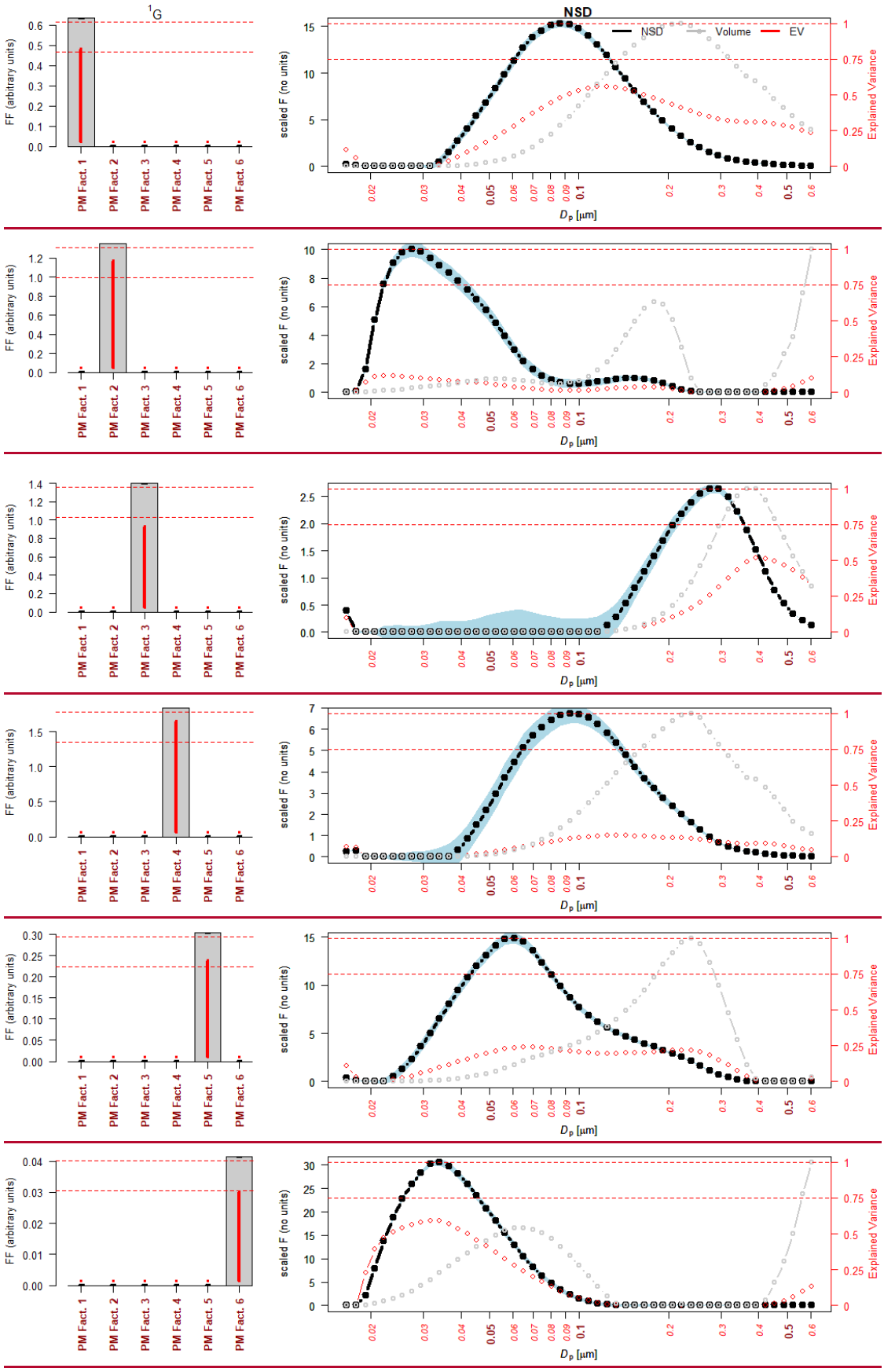



**Figure ~~S3~~S2.** Daily regression source profiles ($\gamma_k$ vs $d_p$ ~~in equation 8)~~ obtained from regressing the NSD data against ${}^{1}G_k$ (left hand panels) as in equation 4 and diurnal trends of the fit parameter $c_k$ resulting from the fit of the daily regression source profiles to the hourly NSD data ~~(equations 10 & 11)~~.


1G
NSD
NSD
Volume
EV
FF (arbitrary units)
scaled F (no units)
Explained Variance
PM Fact. 1
PM Fact. 2
PM Fact. 3
PM Fact. 4
PM Fact. 5
PM Fact. 6
Dp [µm]
Diffuse Urban
2F1
Marine
2F2
Secondary
2F3
NET & Coarse
2F4
Fuel Oil
2F5
Traffic
2F6

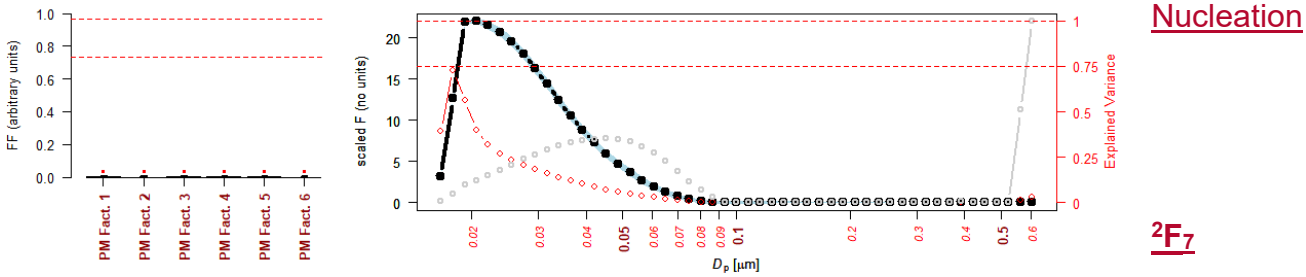



**Figure S3.** Source profiles $^1F$ and $^2F$ from both the first and second PMF step using 7 factors. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variation; and grey dotted line indicates the dV/dlogDp.]

.

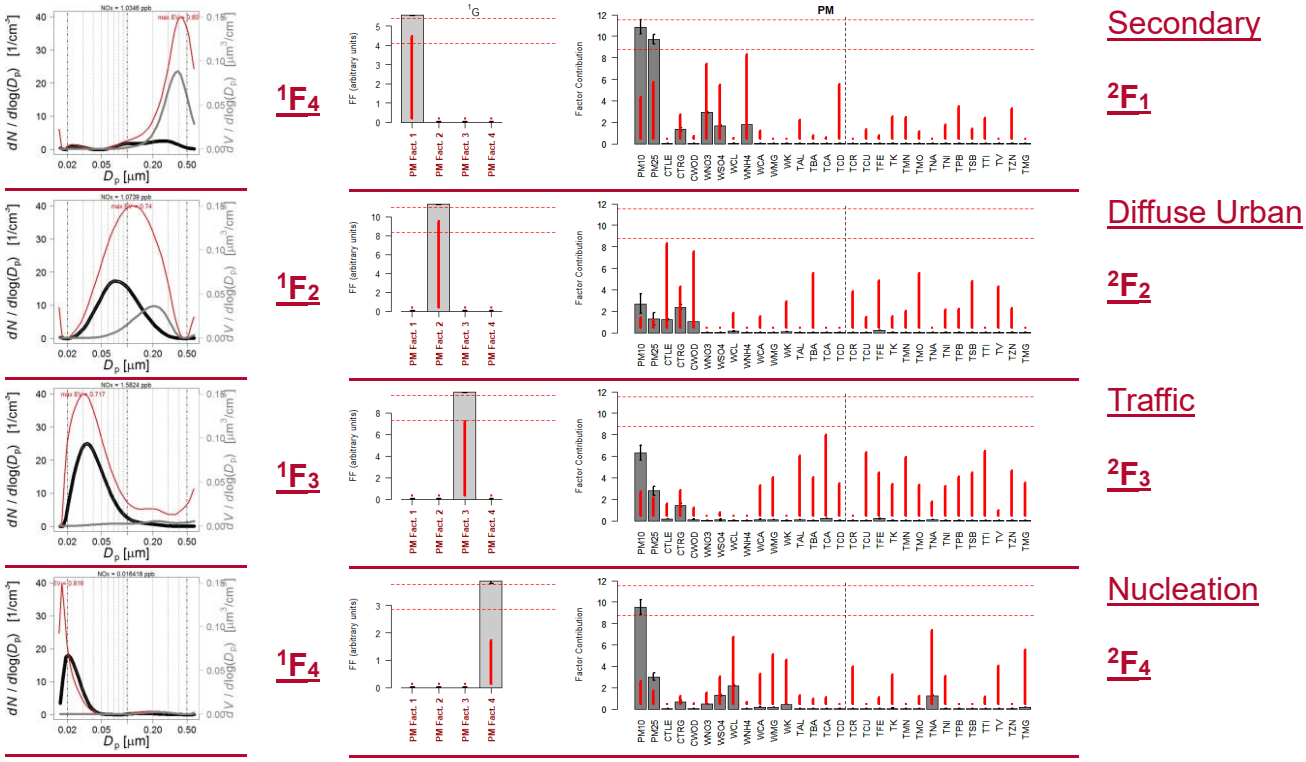
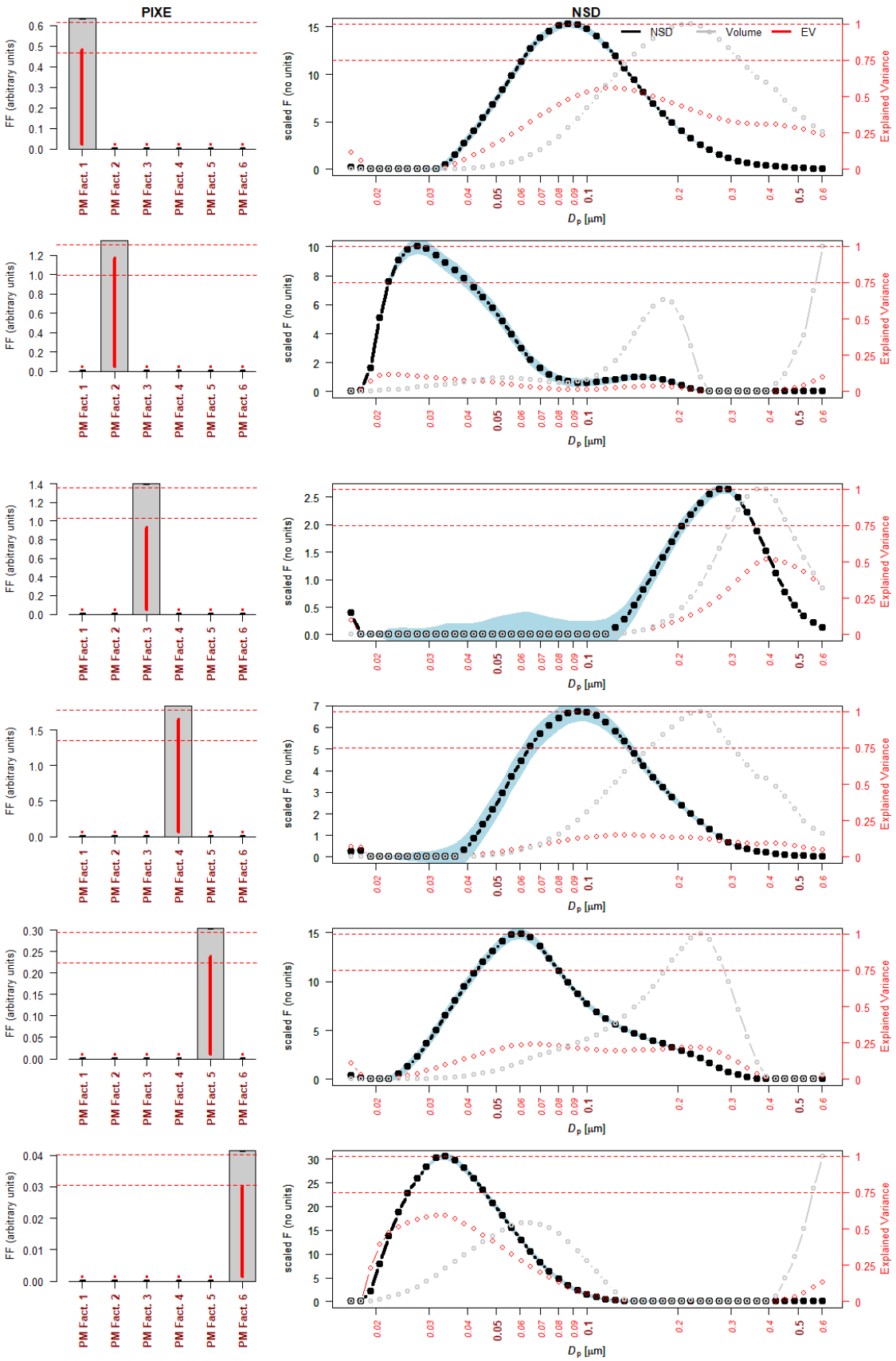



**Figure S4.** PMF-PMF 4 factor analysis of NSD data followed by $PM_{10}$. Each plot is divided into 2 showing the output $^1F_k$ and $^2F_k$. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variation; and grey dotted line indicates the dV/dlogDp.]




~~Diffuse Urban~~

$^2$~~F$_1$~~

~~Marine~~

$^2$~~F$_2$~~

~~Secondary~~

$^2$~~F$_3$~~

~~NET & Coarse~~

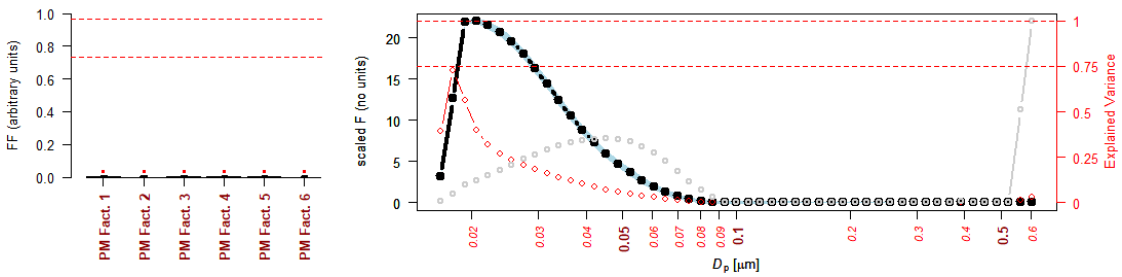
$^2$~~F$_4$~~

~~Fuel Oil~~

$^2$~~F$_5$~~

~~Traffic~~

$^2$~~F$_6$~~

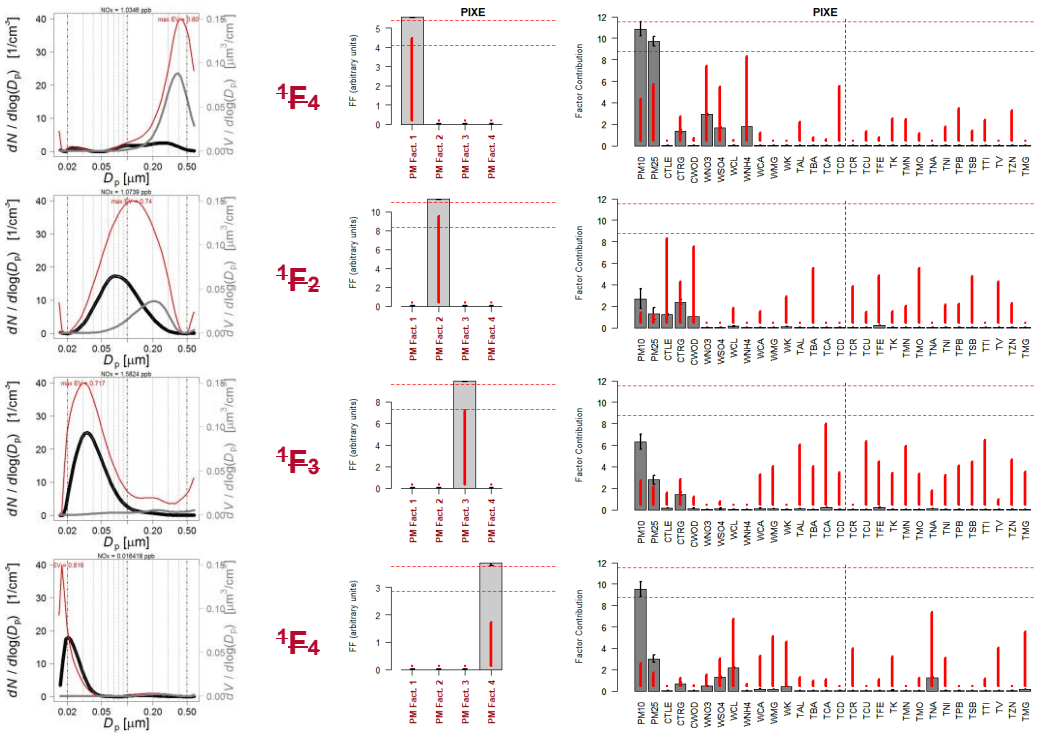



~~Nucleation~~

~~$^2F_7$~~

~~**Figure S4.** Source profiles $^1F$ and $^2F$ from both the first and second PMF step using 7 factors. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variation; and grey dotted line indicates the dV/dlogDp.]~~

~~.~~




~~$^1F_4$~~ ~~Secondary~~ ~~$^2F_1$~~

~~$^1F_2$~~ ~~Diffuse Urban~~ ~~$^2F_2$~~

~~$^1F_3$~~ ~~Traffic~~ ~~$^2F_3$~~

~~$^1F_4$~~ ~~Nucleation~~ ~~$^2F_4$~~

~~**Figure S5.** PMF-PMF 4 factor analysis of NSD data followed by $PM_{10}$. Each plot is divided into 2 showing the output $^1F_k$ and $^2F_k$. [Grey bars and black line indicates the values of F; red lines and dots indicated the explained variation; and grey dotted line indicates the dV/dlogDp.]~~