Supplementary Material for

Urban influence on the concentration and composition of submicron particulate matter in central Amazonia

Suzane S. de Sá (1), Brett B. Palm (2), Pedro Campuzano-Jost (2), Douglas A. Day (2), Weiwei Hu (2), Gabriel Isaacman-VanWertz[a] (3), Lindsay D. Yee (3), Joel Brito[b] (4), Samara Carbone[c] (4), Igor O. Ribeiro (5), Glauber G. Cirino[d] (6), Yingjun J. Liu[e] (1), Ryan Thalman[f] (7), Arthur Sedlacek (7), Aaron Funk (8), Courtney Schumacher (8), John E. Shilling (9), Johannes Schneider (10), Paulo Artaxo (4), Allen H. Goldstein (3), Rodrigo A.F. Souza (5), Jian Wang (7), Karena A. McKinney[g] (1), Henrique Barbosa (4), M. Lizabeth Alexander (11), Jose L. Jimenez (2), Scot T. Martin* (1, 12)

(1) School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA
(2) Department of Chemistry and Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA
(3) Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA
(4) Institute of Physics, University of São Paulo, São Paulo, Brazil
(5) School of Technology, Amazonas State University, Manaus, Amazonas, Brazil
(6) National Institute for Amazonian Research, Manaus, Amazonas, Brazil
(7) Brookhaven National Laboratory, Upton, New York, USA
(8) Department of Atmospheric Sciences, Texas A&M University, College Station, Texas, USA
(9) Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, USA
(10) Particle Chemistry Department, Max Planck Institute for Chemistry, Mainz, Germany
(11) Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA
(12) Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, USA
[a] Now at Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA
[b] Now at Laboratory for Meteorological Physics (LaMP), University Blaise Pascal, Aubière, France
[c] Now at Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil
[d] Now at Department of Meteorology, Geosciences Institute, Federal University of Pará, Belém, Brazil
[e] Now at University of California, Berkeley, California, USA
[f] Now at Department of Chemistry, Snow College, Richfield, Utah, USA
[g] Now at Colby College, Waterville, Maine, USA

*To Whom Correspondence Should be Addressed
E-mail: scot_martin@harvard.edu
https://martin.seas.harvard.edu/

S1. Positive-matrix factorization

S1.1 Diagnostics of the six-factor solution

The time series of organic mass spectra measured by the AMS was analyzed by positive-matrix factorization (PMF) using a standard analysis toolkit Ulbrich et al., 2009. High-resolution "V-mode" data were used. The PMF solution was based on minimization of the "Q-value" (i.e., the sum of the weighed squared residuals for a chosen number of factors) and the physical meaningfulness of factors, as evaluated by profile characteristics and correlations with gas and particle phase measurements by other instruments.

Technical diagnostics of the six-factor solution are presented in Figure S3 in complement to the diagnostics presented in de Sá et al. (2017). The analysis was run for a number of factors from 1 to 10, and the rotational ambiguity parameter f_{peak} was varied from -1 to 1 in intervals of 0.2. Panel a shows the statistics of residuals for solutions with different number of factors. There was a large improvement in the solution when a sixth factor was introduced, as shown by a significant decrease in residuals, and only a marginal improvement when a seventh factor was added. Panel b shows, on the ordinate, the correlation between the time series of loadings for each pair of factors and, on the abscissa, the correlation between the profiles of each pair of factors. For the six-factor solution, the correlations among factor profiles are overall lower, also suggesting a better separation of factors and an improvement in the solution. Figure S4 corroborates this analysis by showing the factor profiles and loading time series of the 5- and 7-factor solutions. In the 5-factor solution, factors 4 and 5 seem to be a result of mixing of the three factors that are associated with secondary processing in the 6-factor solution (MO-OOA, LO-OOA, IEPOX-SOA). Conversely, in the 7-factor solution, some splitting seems to occur as factor 7 is physically meaningless, and a few pairs of factors have higher correlations between their

24 loading time series (cf. Figure S3). An f_{peak} of zero was chosen for the final 6-factor solution,

25 since it yielded the minimum quality of fit parameter $Q/Q_{expected}$ de Sá et al. (2017), and no

26 significant improvements in the external validation of factors were observed by varying f_{peak}.

27 **S1.2 Discussion of the ADOA PMF factor**

28   ADOA is interpreted as a primary anthropogenic factor due to the correlation of its

29 loadings with several tracers of anthropogenic activities (Figure 5), its spectral profile, and its

30 diel behavior (Figure 4). Even though factors containing a characteristic m/z 91 have been

31 reported in the literature as a biogenic factor (Robinson et al., 2011; Budisulistiorini et al., 2015;

32 Chen et al., 2015; Riva et al., 2016), the ADOA of this study showed similarity with primary

33 organic material from cooking activities. Figure S5 shows the high similarity of ADOA of this

34 study to a factor representing cooking emissions at an urban background site in Barcelona, Spain

35 (Mohr et al., 2012), and to a factor representing a cooking source tied to restaurants in an urban

36 background site in Zurich, Switzerland (Lanz et al., 2007). By contrast, a lower similarity is

37 found with the "91fac" factor found in the Borneo forest, a predominantly biogenic site. This

38 result emphasizes that a characteristic marker ion $C_7H_7^+$ at m/z 91 does not directly imply either

39 biogenic or anthropogenic origin, and the interpretation of a PMF factor with such marker should

40 also strongly rely on the atmospheric context of the measurements, including the correlations of

41 the factor loadings with external measurements and the diel behavior.

42 **S2. Estimates of organic and inorganic nitrates based on AMS analysis**

43   The typical AMS analysis reports total nitrate, meaning that nitrate fragments originating

44 from both organic and inorganic nitrates are reported indistinctively as nitrate. In the absence of

45 external measurements of inorganic nitrate, an estimation method using the ratio of NO_2^+ to NO^+

46 signal intensities measured by the AMS was employed (Figure S6; Fry et al., 2009; Farmer et al.,

47   2010; Fry et al., 2013). Calculations were done on a 60-min time base to increase signal over

48   noise. The obtained organic and inorganic nitrate time series were then interpolated into the

49   original AMS timestamp for ambient measurements (i.e., one point every 8-min interval). The

50   analysis excluded points that had total nitrate below the estimated detection limit, $DL_{Nitrate}$, which

51   was estimated as three times the standard deviation for "closed AMS spectra", i.e., when chopper

52   was in closed position and particles did not reach the vaporizer. Mathematically,

53    $DL_{Nitrate} = 3 \times \sqrt{E}$, where E is the "closed" error calculated by the standard *PIKA* software

54   (Ulbrich et al., 2009). The dark blue dashed line in Figure S6c that defines NO_2^+/NO^+ for

55   inorganic nitrate was determined by linear fit of ammonium nitrate calibrations performed

56   regularly, as shown by the grey triangles. The small drift over time can be attributed to a gradual

57   clean-up of the vaporizer. Worth noting, whether the linear fit or an average value was used for

58   the calculations, the overall results did not change considerably, as all calibration ratios lied

59   within ± 20% of the campaign-average ratio. The ratio NO_2^+/NO^+ for organic nitrates was

60   assumed to be a factor of 2.25 lower than that or inorganic nitrate based on previous field studies

61   (Farmer et al., 2010; Fry et al., 2013). The resulting IOP1-average for the fraction of organic

62   nitrate in total nitrate (Figure S6b) was 87%.

63   **S3. Fuzzy c-means clustering**

64   Fuzzy c-means (FCM) clustering was applied to the dataset consisting of concentrations

65   of particle number, NO_y, ozone, black carbon, and sulfate (Bezdek et al., 1984). The use of a

66   fuzzy clustering method stems from the understanding that any point in time may be affected by

67   a combination of different sources and processes and could therefore be anywhere on the scale

68   between pristine background and extreme polluted conditions, as opposed to a simpler binary

69   classification. Given the scope of the analysis as non-overcast afternoon times, data points were

70   restricted to (i) local 12:00-16:00 h, (ii) local solar radiation over the past 4 h not less than 200 W

71   m^{-2} (i.e., excluding the lower 20 percentile), and (iii) insignificant precipitation (< 0.1 mm) over

72   the previous 10 h along backward trajectory (a threshold was used as most rain radar grid cells

73   had non-zero yet negligible values). The data were normalized prior to the FCM analysis using

74   the z-score method, which transforms all variables into a common scale with an average of zero

75   and standard deviation of one.

76   The FCM algorithm minimizes the objective function represented in Eq. S1, which is a

77   weighted sum of squared errors where the error is the Euclidean distance between each data

78   point and a cluster centroid.

79                    $$J(U,v) = \sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}{}^{m} \|y_{k} - v_{i}\|^{2}$$ (Eq. S1)

80   The input data is given by the matrix $Y = [y_1, y_2, \ldots, y_N]$, where y_k is a vector of length X at the k-

81   th time point. X is the number of variables (i.e., measurements) used as input in the analysis. The

82   number of time points is represented by N, and the associated running index is k. N in this case

83   was 313. The number of clusters is represented by c, and the corresponding running index is i.

84   The coordinates of the centroid of each cluster i are represented by v_i, a vector of length X. The

85   exponent of the Fuzzy partition matrix is represented by m. The algorithm returns (1) the Fuzzy

86   partition matrix of Y, given by $U = [u_{ik}]$ where u_{ik} is the degree of membership of time point k to

87   cluster i, (2) the vectors of coordinates of cluster centers, given by $v = [v_i]$, as well as (3) the

88   value J of the objective function.

89   The analysis was performed in MATLAB® using the "fcm" function in the Fuzzy logic

90   toolbox^{TM}. The stop criterion of the algorithm is that either the maximum number of iterations is

91   reached or the improvement of the objective function between two consecutive iterations is less

92   than the minimum amount of improvement specified. The default value of 1×10^{-5} was used for

93   the minimum amount of improvement, and the maximum number of iterations was set to 1000 so

94   that convergence always happened before this maximum was reached. A default value of 2 was

95   used for the exponent m of the partition matrix. Fuzzy clustering algorithms are not sensitive to

96   small fluctuations in m (Chatzis, 2011), and a value in the range of 1.5 to 3 is recommended

97   (Bezdek et al., 1984; Hathaway and Bezdek, 2001).

98   The analysis was run for a number of clusters varying from two to eight, and the value of

99   the objective function for each run is shown in Figure S7. The choice of number of clusters

100   hinges on a balance between increased complexity and additional information provided by each

101   extra cluster. The improvement in the objective function was larger in the range of two to four

102   clusters, with marginal improvements above four clusters (Figure S7). The location of cluster

103   centroids was also examined for evaluation of cluster overlap (Figure S8). The addition of a fifth

104   cluster made two pairs of clusters very similar, as can be seen by the locations of cluster

105   centroids in Figure S8. The solution of four clusters was therefore a reasonable choice to

106   represent the studied system. The subsequent characterization of the PM chemical composition

107   associated with each cluster further confirmed the meaningfulness of the solution. Although the

108   three-cluster solution could also provide a reasonable representation of the system, the four-

109   cluster solution provided further insight by differentiating two background and two polluted

110   conditions.

111   Subsequently, the PM composition associated with each of the clusters was determined

112   by calculating the corresponding coordinates of the centroids for AMS species concentrations

113   and PMF factor loadings, which were not input to the FCM analysis. The calculation followed

114   the mathematical definition of the centroid (Eq. S2). The resulting characterization of clusters is

115   shown in Figure 8 and Table 2.

116
$$\nu_i = \frac{\sum_{k=1}^{N} (u_{ik})^m y_k}{\sum_{k=1}^{N} (u_{ik})^m}$$
(Eq. S2)

References

Bezdek, J. C., Ehrlich, R., and Full, W.: FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, 10, 191-203, 1984.

Budisulistiorini, S. H., Li, X., Bairai, S. T., Renfro, J., Liu, Y., Liu, Y. J., McKinney, K. A., Martin, S. T., McNeill, V. F., Pye, H. O. T., Nenes, A., Neff, M. E., Stone, E. A., Mueller, S., Knote, C., Shaw, S. L., Zhang, Z., Gold, A., and Surratt, J. D.: Examining the effects of anthropogenic emissions on isoprene-derived secondary organic aerosol formation during the 2013 Southern Oxidant and Aerosol Study (SOAS) at the Look Rock, Tennessee ground site, Atmos. Chem. Phys., 15, 8871-8888, 2015, 10.5194/acp-15-8871-2015.

Chatzis, S. P.: A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional, Expert Systems with Applications, 38, 8684-8689, 2011, https://doi.org/10.1016/j.eswa.2011.01.074.

Chen, Q., Farmer, D. K., Schneider, J., Zorn, S. R., Heald, C. L., Karl, T. G., Guenther, A., Allan, J. D., Robinson, N., Coe, H., Kimmel, J. R., Pauliquevis, T., Borrmann, S., Pöschl, U., Andreae, M. O., Artaxo, P., Jimenez, J. L., and Martin, S. T.: Mass spectral characterization of submicron biogenic organic particles in the Amazon Basin, Geophys. Res. Lett., 36, L20806, 2009, 10.1029/2009GL039880.

Chen, Q., Farmer, D. K., Rizzo, L. V., Pauliquevis, T., Kuwata, M., Karl, T. G., Guenther, A., Allan, J. D., Coe, H., Andreae, M. O., Pöschl, U., Jimenez, J. L., Artaxo, P., and Martin, S. T.: Submicron particle mass concentrations and sources in the Amazonian wet season (AMAZE-08), Atmos. Chem. Phys., 15, 3687-3701, 2015, 10.5194/acp-15-3687-2015.

de Sá, S. S., Palm, B. B., Campuzano-Jost, P., Day, D. A., Newburn, M. K., Hu, W., Isaacman-VanWertz, G., Yee, L. D., Thalman, R., Brito, J., Carbone, S., Artaxo, P., Goldstein, A. H., Manzi, A. O., Souza, R. A. F., Mei, F., Shilling, J. E., Springston, S. R., Wang, J., Surratt, J. D., Alexander, M. L., Jimenez, J. L., and Martin, S. T.: Influence of urban pollution on the production of organic particulate matter from isoprene epoxydiols in central Amazonia, Atmos. Chem. Phys., 17, 6611-6629, 2017, 10.5194/acp-17-6611-2017.

Farmer, D. K., Matsunaga, A., Docherty, K. S., Surratt, J. D., Seinfeld, J. H., Ziemann, P. J., and Jimenez, J. L.: Response of an aerosol mass spectrometer to organonitrates and organosulfates and implications for atmospheric chemistry, Proc. Natl. Acad. Sci. USA, 107, 6670-6675, 2010, 10.1073/pnas.0912340107.

Fry, J. L., Kiendler-Scharr, A., Rollins, A. W., Wooldridge, P. J., Brown, S. S., Fuchs, H., Dubé, W., Mensah, A., dal Maso, M., Tillmann, R., Dorn, H. P., Brauers, T., and Cohen, R. C.: Organic nitrate and secondary organic aerosol yield from NO_3 oxidation of β-pinene evaluated using a gas-phase kinetics/aerosol partitioning model, Atmos. Chem. Phys., 9, 1431-1449, 2009, 10.5194/acp-9-1431-2009.

Fry, J. L., Draper, D. C., Zarzana, K. J., Campuzano-Jost, P., Day, D. A., Jimenez, J. L., Brown, S. S., Cohen, R. C., Kaser, L., Hansel, A., Cappellin, L., Karl, T., Hodzic Roux, A., Turnipseed, A., Cantrell, C., Lefer, B. L., and Grossberg, N.: Observations of gas- and aerosol-phase organic nitrates at BEACHON-RoMBAS 2011, Atmos. Chem. Phys., 13, 8585-8605, 2013, 10.5194/acp-13-8585-2013.

Hathaway, R. J. and Bezdek, J. C.: Fuzzy c-means clustering of incomplete data, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31, 735-744, 2001, 10.1109/3477.956035.

IBGE: Malhas digitais, Setor Censitário, 2010, https://mapas.ibge.gov.br/bases-e-referenciais/bases-cartograficas/malhas-digitais.html, last access: January 2018.

Lanz, V. A., Alfarra, M. R., Baltensperger, U., Buchmann, B., Hueglin, C., and Prévôt, A. S. H.: Source apportionment of submicron organic aerosols at an urban site by factor analytical modelling of aerosol mass spectra, Atmos. Chem. Phys., 7, 1503-1522, 2007, 10.5194/acp-7-1503-2007.

Medeiros, A., Souza, R. A. F., and Martin, S. T., in preparation.

Mohr, C., DeCarlo, P. F., Heringa, M. F., Chirico, R., Slowik, J. G., Richter, R., Reche, C., Alastuey, A., Querol, X., Seco, R., Peñuelas, J., Jiménez, J. L., Crippa, M., Zimmermann, R., Baltensperger, U., and Prévôt, A. S. H.: Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data, Atmos. Chem. Phys., 12, 1649-1665, 2012, 10.5194/acp-12-1649-2012.

Riva, M., Budisulistiorini, S. H., Chen, Y., Zhang, Z., D'Ambro, E. L., Zhang, X., Gold, A., Turpin, B. J., Thornton, J. A., Canagaratna, M. R., and Surratt, J. D.: Chemical characterization of secondary organic aerosol from oxidation of isoprene hydroxyhydroperoxides, Environ. Sci. Technol., 2016, 10.1021/acs.est.6b02511.

Robinson, N. H., Hamilton, J. F., Allan, J. D., Langford, B., Oram, D. E., Chen, Q., Docherty, K., Farmer, D. K., Jimenez, J. L., Ward, M. W., Hewitt, C. N., Barley, M. H., Jenkin, M. E., Rickard, A. R., Martin, S. T., McFiggans, G., and Coe, H.: Evidence for a significant proportion of Secondary Organic Aerosol from isoprene above a maritime tropical forest, Atmos. Chem. Phys., 11, 1039-1050, 2011, 10.5194/acp-11-1039-2011.

Schneider, J., Freutel, F., Zorn, S., Chen, Q., Farmer, D., Jimenez, J., Martin, S., Artaxo, P., Wiedensohler, A., and Borrmann, S.: Mass-spectrometric identification of primary biological particle markers and application to pristine submicron aerosol measurements in Amazonia, Atmos Chem Phys, 11, 11415-11429, 2011, doi.org/10.5194/acp-11-11415-2011.

Ulbrich, I., Canagaratna, M., Zhang, Q., Worsnop, D., and Jimenez, J.: Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9, 2891-2918, 2009, 10.5194/acp-9-2891-2009.

List of Supplementary Figures

Figure S1. Location of the GoAmazon2014/5 sites relevant for this study. Image data: Google earth.

Figure S2. Scatter plot of the AMS signal fraction at m/z 44 (f_{44}) against that at m/z 43 (f_{43}). Green and yellow markers correspond to measurements made by two different AMS instruments at T0t in the wet season of 2008 during the AMAZE-08 campaign (Chen et al., 2009; Schneider et al., 2011). Red markers correspond to measurements made at the T0a (ATTO) by an ACSM during the wet season of 2015. A correction factor of 0.75 was applied to the f_{44} values of the ACSM based on calibrations with standards. Solid squares represent median values, and whiskers represent 10 and 90 percentiles. The plot shows a significant variability between the observations of 2008 and 2015 for the two background sites. An explanation of the differences is not attempted herein and warrants further investigation through longer-term continuous measurements.

Figure S3. Diagnostics of the PMF analysis. (a) Statistics of the sum of for solutions with different number of factors. Box plots show the interquartile ranges, including the medians as a horizontal line. Red markers show the means. Whiskers show the 5 and 95 percentiles. (b) Correlations expressed as between each pair of factors within each PMF solution, with number of factors varying from 2 to 7. The Pearson R value between factor loadings is shown on the coordinate and between factor profiles is shown on the abscissa. Numbers in red indicate the identity of the pair of factors.

Figure S4. Results of the PMF analysis for 5 factors (a and b) and 7 factors (c and d). Panels on the left (a and c) show the time series of factor loadings and panels on the right (b and

10

d) show the profiles of factors. The signals shown in panels b and d were summed to unit mass resolution.

Figure S5. Comparison of the ADOA factor profile from the present study to factors found in three other field studies. "COA" are factors representative of cooking activities, and the "91fac" from Robinson et al. (2011) was tied to biogenic sources.

Figure S6. Summary of the analysis for estimating organic and inorganic nitrates from AMS bulk measurements. (a) Resulting time series of organic and inorganic nitrates are shown together with the original nitrate AMS times series. (b) Time series of the fraction of organic nitrate in total nitrate. (b) Time series of the measured NO_2^+/NO^+ ratio is shown in red and values of NO_2^+/NO^+ from ammonium nitrate calibrations are shown in gray triangles. A linear fit to those calibration ratios is shown by the dashed dark blue line and constitutes the reference ratio for inorganic nitrate over time. The dashed light blue line is the reference ratio for organic nitrates over time. Calculations were done for data binned to one hour (as plotted), and the resulting time series were interpolated to the native time stamp for evaluation of correlations in the PMF analysis.

Figure S7. Value of the objective function of the FCM analysis (Eq. 1) in the last iteration plotted against the number of clusters.

Figure S8. Locations of cluster centroids from the FCM analysis as visualized by a 2-D projection on the plane defined by each pair of input variables. Results for two to five clusters are shown in panels a to d. Red circles are observational data and black squares are cluster centroids.

Figure S9. Map of Manaus city depicting population density as well as main avenues and representative locations of industry, restaurants, and other businesses. Population density data are from the 2010 census by the Brazilian Institute of Geography and Statistics (IBGE, 2010).

Figure S10. Measurements showing the geographical heterogeneity of emissions from Manaus. On the top row, concentrations of sulfate (red) and particle number (white) measured onboard the G-1 aircraft on (a) March 19 and (b) Mar 21. Image data: Google earth. On the bottom row, rose plots of mean (c) sulfate mass concentrations and (d) particle number concentrations observed at T2 during IOP1. The angles represent wind direction, the radial scale (0 to 5 m s^{-1}) represents wind speed, and the color scale represents the concentrations. The interactions of emissions from Manaus with the daily river breeze is complex, and the detailed interpretation of the data sets is not fully attempted herein. Of importance, the river breeze terminates well below 500 m based on the G-1 flights so that the complexities of the river breeze largely do not affect the measurements at T3 because most pollution is lofted above the river breeze before reaching T3 (Medeiros et al., in preparation). These surface-level plots, although complicated by the river breeze, demonstrate the heterogeneity of Manaus emissions.
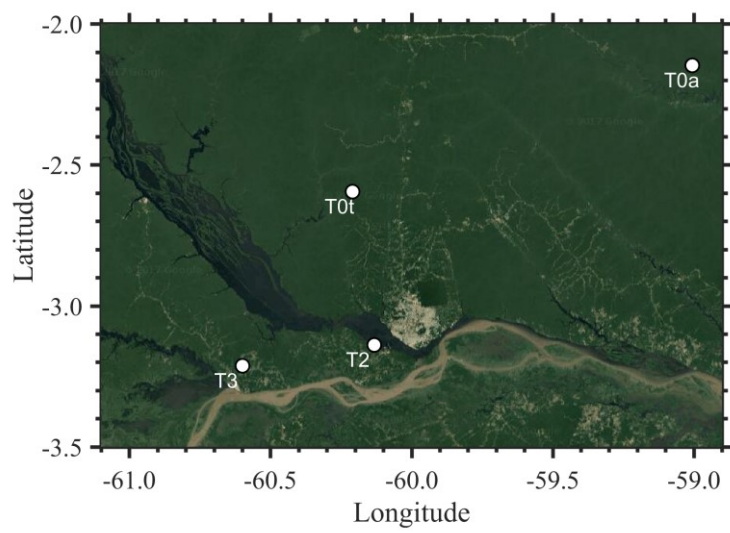
Figure S1
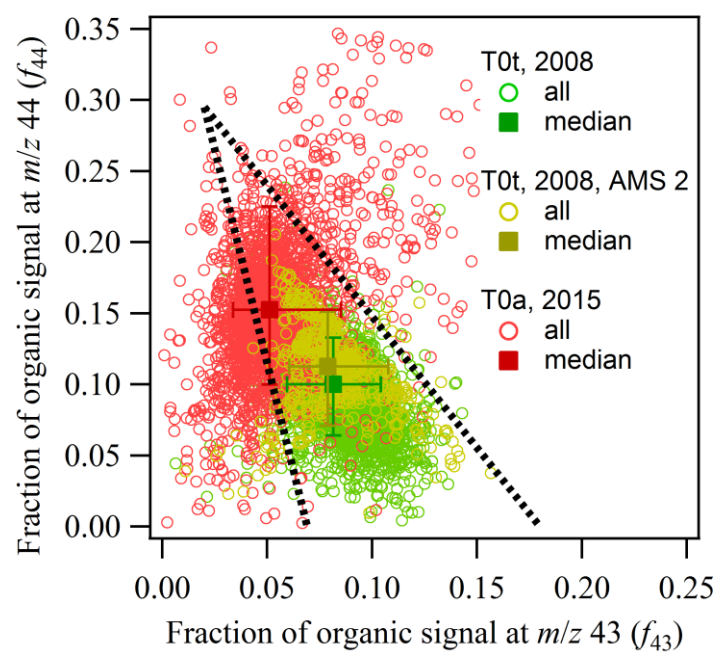
Figure S2
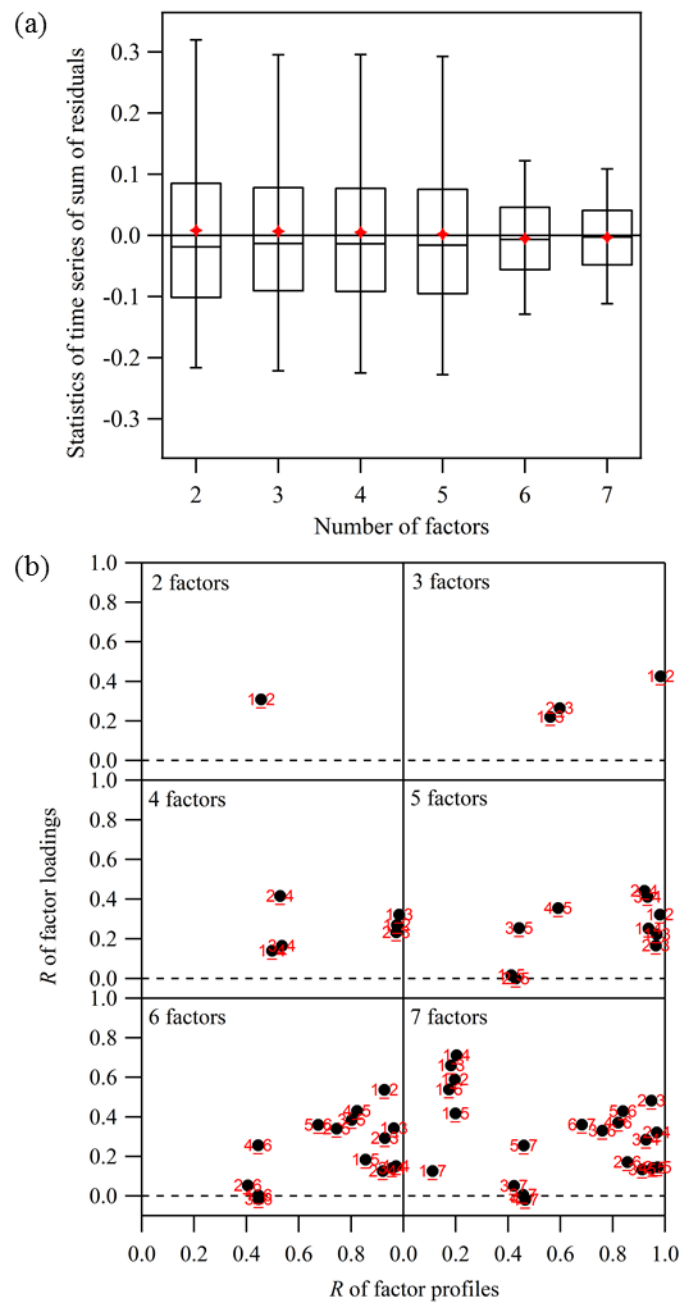
Figure S3
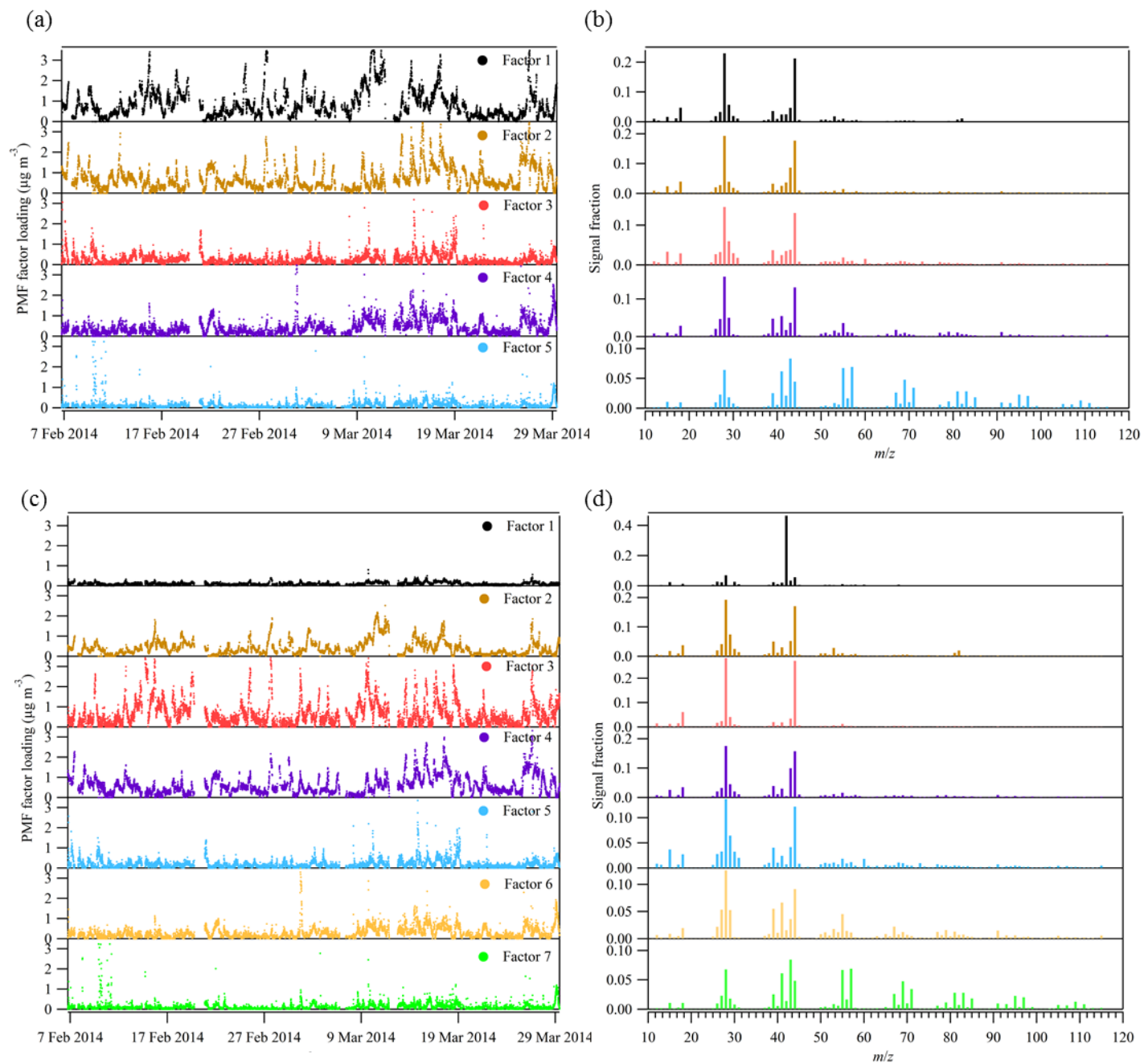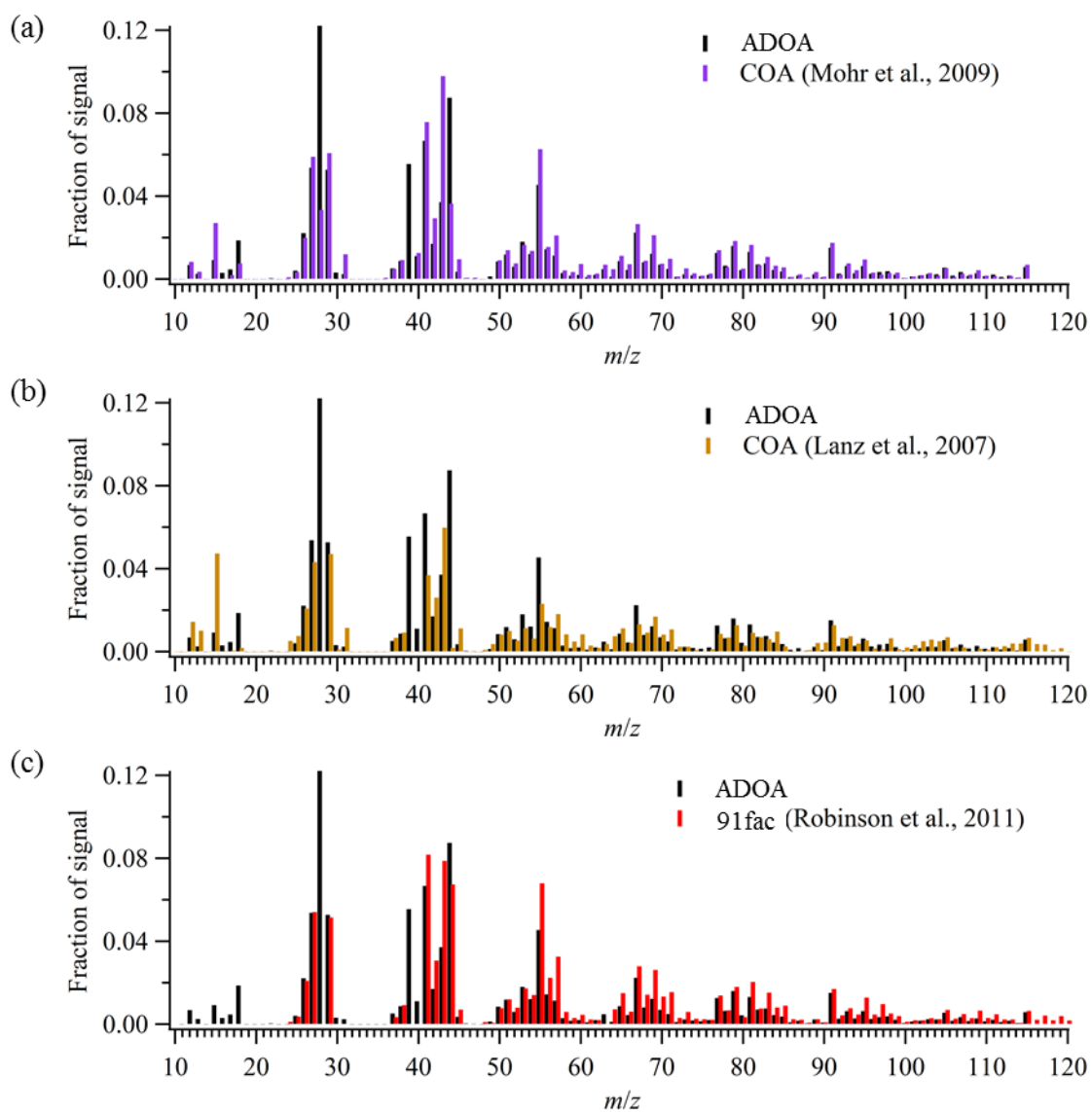
Figure S5
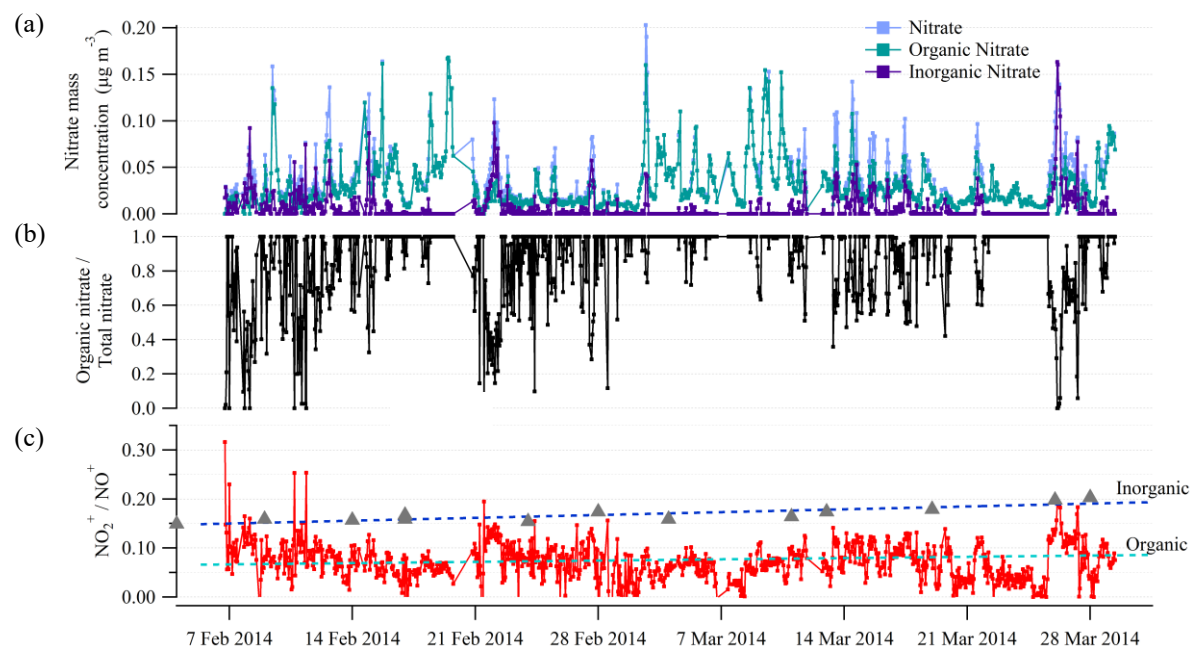
Figure S6
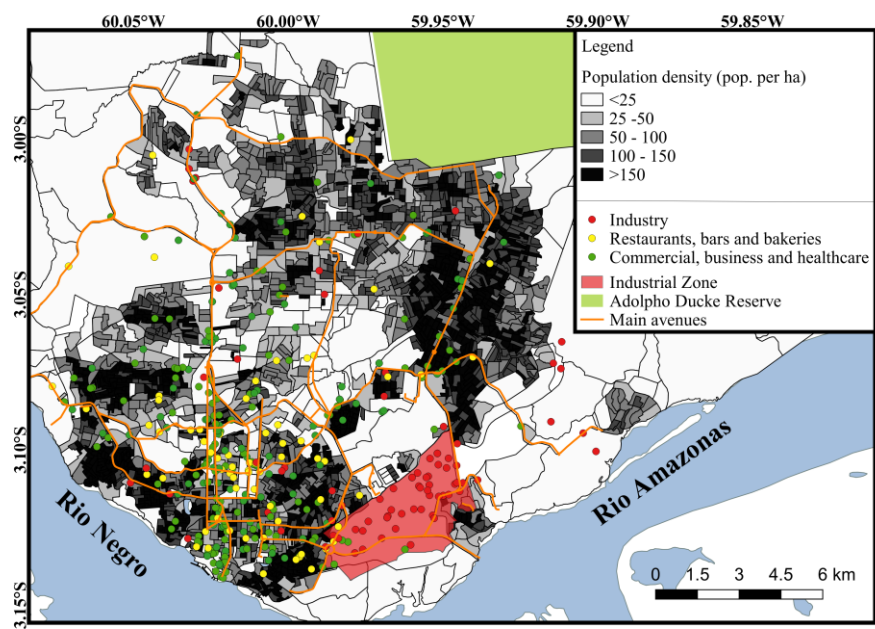
Figure S7

(a)

(b)

(c)

(d)
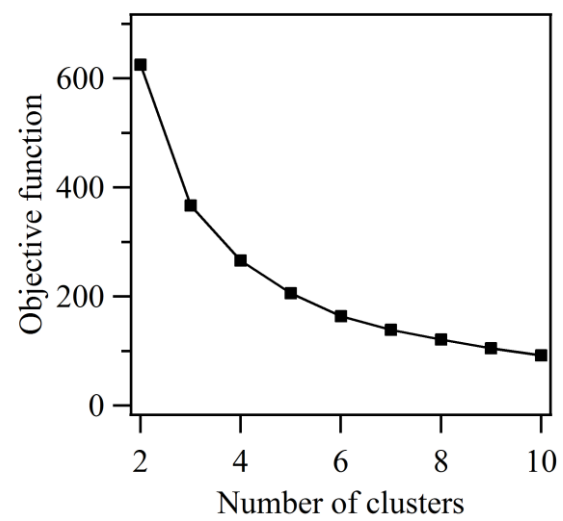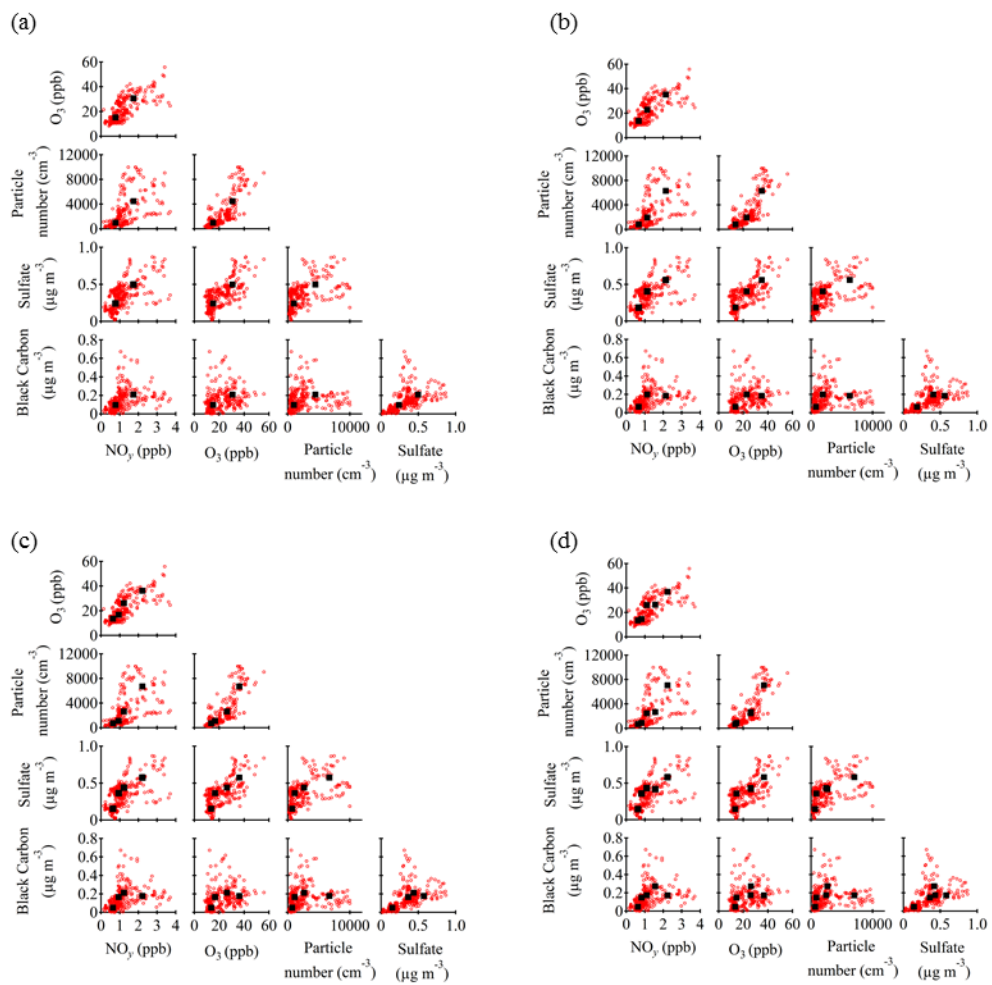
Figure S8

Figure S9
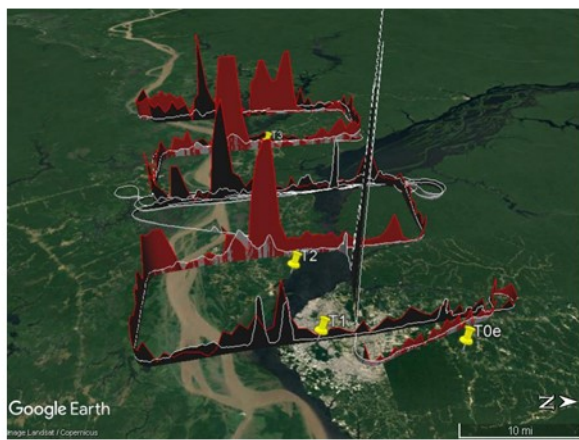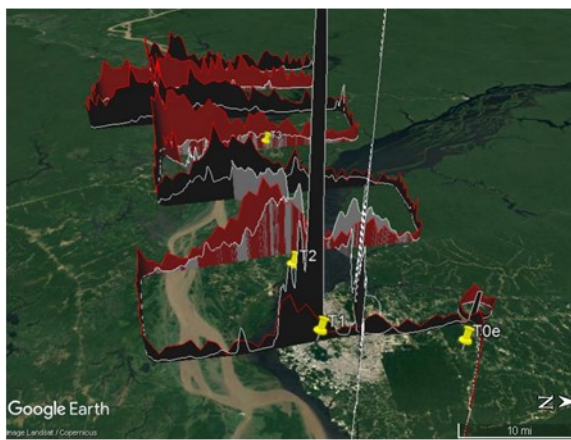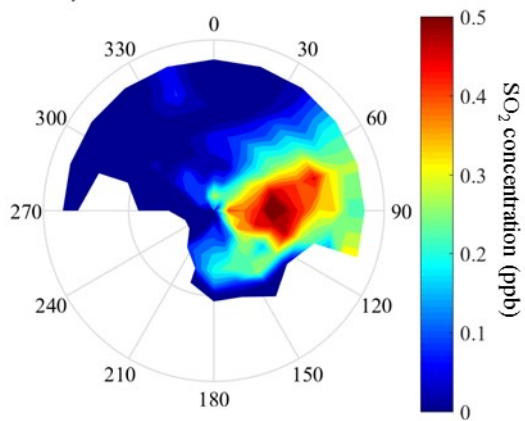
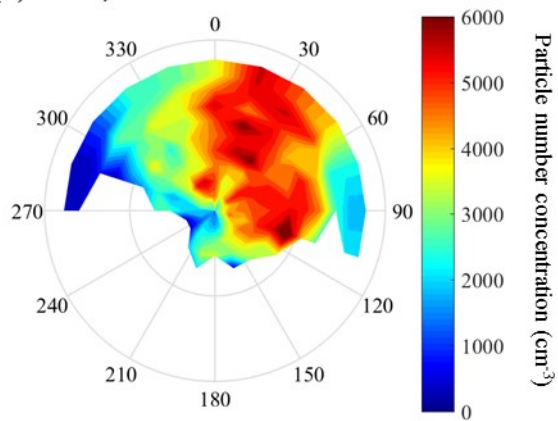(a) 19 Mar 2014

(b) 21 Mar 2014

(c) IOP1, T2

(d) IOP1, T2

Figure S10