ELSEVIER

Atmospheric Environment 37 (2003) 4861–4867

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

COPREM—A multivariate receptor model with a physical approach

Peter Wåhlin*

Department of Atmospheric Environment, National Environmental Research Institute, Roskilde DK-4000, Denmark

Received 12 March 2003; received in revised form 26 June 2003; accepted 13 August 2003

Abstract

A hybrid receptor model, combining the features of chemical mass balance models and non-negative factor analysis, is presented both in theoretical terms and from a more practical viewpoint. A section describes the mathematical details, and another section presents an uncomplicated set of aerosol data from Northeast Greenland and a step-to-step demonstration of how the data was analysed using the model. Since 1992 the model has proven useful in several air pollution studies of very different character.
© 2003 Elsevier Ltd. All rights reserved.

Keywords: Receptor models; Non-negative factor analysis; Chemical mass balance models; Source apportionment

1. Introduction

The two main branches of receptor models are Chemical Mass Balance (CMB) models and multivariate factor analysis models. CMB models give the most objective source apportionment and need only one sample. They assume knowledge of the number of required sources and their composition, while factor analysis, by artful mathematics, attempts to apportion the sources and determine their composition on the basis alone of a series of observations at the receptor site (Henry et al., 1984). The most recent approach in factor analysis is Positive Matrix Factorization (PMF), which excludes negative solutions (Paatero, 1997).

Factor analysis is a powerful tool, because the choice of the model dimension and the search for more realistic solutions by axis rotations can be based entirely on mathematical criteria. Nevertheless it is a problem, as has been emphasized by Henry (1987), that factor analysis attempts to get more information out of the data than is really there. Thus, it is a serious methodical limitation that the results always need subjective interpretation in relation to the physical world. As an improvement confirmatory factor analysis offers some control of the solutions by "fixing" or "freeing" specific parameters such as the factor correlation coefficients. These parameters are set according to the theoretical expectation of the researcher (Gleser, 1997; Christensen and Sain, 2002). In an analogous way, the PMF method offers some control of the solution. If specific values in the solution are known to be zero, then it is possible to force the solution toward zero for those values through appropriate settings, but it is not possible to use constant source profiles as in CBM. Recently a more general and complex program, The Multilinear Engine (Paatero, 1999), has been introduced, which can solve multi-linear problems with the possibility of implementing all kinds of additional constraints using a script language.

The Constrained Physical Receptor Model (COPREM) is a simple hybrid model that unifies qualities from factor analytic models and CMB. COPREM allows more insight and offers better control over the final solution than factor analysis, but also requires more input information. Since the introduction (Wåhlin, 1993a) COPREM has proven useful in air pollution studies at very different places and under very different

*Tel.: +45-4630-1841; fax: +45-4630-1214.
E-mail address: pew@dmu.dk (P. Wåhlin).

1352-2310/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2003.08.032

conditions, e.g. in Central Greenland (Wåhlin, 1996), in Northeast Greenland (Wåhlin, 1993b; Heidam et al., 1999), in Malaysia (Wåhlin et al., 1998), in Denmark (Swietlicki et al., 1999; Wåhlin et al., 2001), and in Thailand (Upadhyay et al., 2003).

2. The constrained physical receptor model

The fundamental hypothesis in mathematical receptor models is that the measured concentrations x_{ij} of n different chemical components in N samples can be satisfactorily expressed as a sum of contributions from a small number $p(p<n)$ of sources with a constant composition,

$$x_{ij} \cong \sum_k a_{ik} f_{kj}. \tag{1}$$

In Eq. (1) a_{ik} are constants representing the relative composition of the sources (the source profiles), and f_{kj} are the source strengths found in the different samples. Index i refers to the n chemical variables, index j refers to the N samples, and index k refers to the p sources.

The process of solving Eq. (1) implies the search for source profiles a_{ik} and source strengths f_{kj}, which, by insertion on the right-hand side of Eq. (1), give results close to the measured x_{ij}. From a physical point of view, the best solution minimizes χ^2, given by

$$\chi^2 = \sum_j \sum_i \frac{(x_{ij} - \sum_k a_{ik} f_{kj})^2}{\sigma_{ij}^2}, \tag{2}$$

where σ_{ij} is the expected scatter (or uncertainty) of x_{ij}. This is of particular importance for data near or below the detection limit, and guarantees that high values with high uncertainties will have a smaller weight in the fitting process. In general, infinitely many solutions are equally good in a purely mathematical sense, but the arbitrariness (the size of the solution space) can be reduced if physical restrictions are considered. For a simple linear model like Eq. (1) a fundamental physical demand is that both a_{ik} and f_{kj} must be non-negative. This demand can be complied with if constraints are introduced in the minimizing process every time a_{ik} and f_{kj} hit the non-physical barrier. In addition, as the most important feature, other general constraints can be set up for the source profiles using knowledge about real source composition. Thus, in some cases a careful choice of the constraints can yield a uniquely identifiable model.

An iterative process is used, in which the minimizing of χ^2 is performed solving the systems of linear equations derived from

$$\frac{\partial \chi^2}{\partial f_{\ell' j}} = 0 \tag{3a}$$

and

$$\frac{\partial \chi^2}{\partial a_{i\ell'}} = 0 \tag{4a}$$

alternately (alternating least squares, introduced by Yates, 1933). Solving the equations derived from Eq. (3a) corresponds to the approach in the CMB model, which presupposes known source profiles, while Eq. (4a) leads to a multiple regression analysis searching for better fitting source profiles. The apostrophe on the source index label (ℓ') indicates that some sources may be omitted in order to prevent violation of constraints. An initial profile matrix a_{ik} is set up, in which the column source vectors are chosen freely. The research worker can use background knowledge to direct the iteration to a rational result by choosing, for example, vectors that are proportional to known source profiles, or vectors with a single non-zero component used as tracer, and by setting up constraints that keep parts of source profiles or whole profiles constant and hereby prevents unwanted mixing of the source vectors. As more constraints are added to the hybrid model, it can be gradually changed from a pure multivariate model to a pure CMB model.

Each iteration loop thus involves two steps:

1. The source strengths are found for each sample j by solving the system of linear Eq. (3a), which after differentiation can be written:

$$\sum_{k'} \left(\sum_i \frac{a_{ik'} a_{i\ell'}}{\sigma_{ij}^2} \right) f_{k'j} = \sum_i \frac{x_{ij} a_{i\ell'}}{\sigma_{ij}^2}. \tag{3b}$$

   First all p sources are used, but if negative elements are found in the resulting source strength matrix f_{kj}, they are replaced by zeros and the affected columns are recalculated omitting the corresponding sources. The apostrophes of the indices k' and ℓ' indicate that the number of sources in the individual samples in this way may be reduced, $p_j \leqslant p$.
2. The finally achieved matrix f_{kj}, with non-negative elements, is used to calculate the residuals,

$$\Delta x_{ij} = x_{ij} - \sum_k a_{ik} f_{kj} \tag{5}$$

   and a correction Δa_{ik} of the source profile matrix is calculated by solving the system of linear equations derived from Eq. (4a),

$$\sum_{k'} \left(\sum_j \frac{f_{k'j} f_{\ell' j}}{\sigma_{ij}^2} \right) \Delta a_{ik'} = \sum_j \frac{\Delta x_{ij} f_{\ell' j}}{\sigma_{ij}^2}. \tag{4b}$$

   The k' and ℓ' apostrophes indicate that the imposed constraints may reduce the number of sources for the individual chemical component i: If an element a_{ik} is not allowed to change, the corresponding source is omitted, $p_i \leqslant p$. The information about which elements should be held constant is given by a form

matrix of the same size as a_{ik}, where zero values indicate that the corresponding element in a_{ik} is constant.

3. The calculation of the matrix Δa_{ik} is completed by inserting zeros for the elements that were skipped in the solution Δa_{ik}, of Eq. (4b), and the resulting matrix is added to the original matrix a_{ik} to get a better overall fit. If negative elements are found in the new a_{ik} matrix, zeros replace them, and the affected rows are recalculated without the corresponding sources taking the new emerging constraints into account. The method is a modification and extension of the NNLS method (non-negativity constrained least-squares regression, see Lawson and Hanson, 1974).

The iteration is continued until χ^2 reaches a stagnant minimum value. The degrees of freedom, approximately given by

$$v = Nn - \sum_j p_j - \sum_i p_i, \tag{6}$$

depend upon the number of negative values of a_{ik} and f_{kj} found in the last iteration loop, therefore both χ^2 and v are determined by the iterative process.

As a special option, the strengths of the first source ($k = 0$) can be constrained to constant unity values ($f_{0j} = 1$) enabling the use of constant offset values of the variables according to the equation

$$x_{ij} \cong x_{i0} + \sum_{k>0} a_{ik} f_{kj}. \tag{7}$$

A code has been written in C-language and compiled as a DOS executable program to perform the iteration. As inputs are used three matrices: An $n \times p$ profile matrix containing the initial source vectors, an $n \times p$ form matrix containing information about the constraints on the profile matrix elements in 0 = fixed/1 = free format, and a $2 \times n \times N$ data matrix containing the measured values and their absolute uncertainties. A negative uncertainty is interpreted by the code as an infinite uncertainty and can be used to exclude selected data from the fitting process. Output is a $p \times N$ non-negative source strength matrix, an $n \times p$ non-negative profile matrix, plus χ^2 and v. In the final calculation a gauge rule is needed for the two factors in the matrix product in Eq. (1). Here, the strength values of the individual sources in the source strength matrix are normalized to unity in average, and, in consequence, the profile matrix elements will be the average source contributions. Furthermore, using multiple linear regression, a "one-factor" analysis is performed on the residues to reveal a possibly ignored source, and the result is expressed as an extra row in the source strength matrix and an extra column in the source profile matrix.

3. A simple example

As an illustration of the nature and the use of the physical receptor model some aerosol data are analysed here. For the sake of the demonstration a very simple problem was selected, in which our understanding of the number of sources and the qualitative nature of the source profiles is well established. The data are weekly average concentrations of Si (silicon), Mn (manganese) and Pb (lead) measured at Station Nord in Northeast Greenland in the period January–December 2000. The two dominant sources in the Arctic of these elements are long-range and local soil erosion in spring and summer (Si, Mn) and long-range anthropogenic combustion emissions from Russia and Europe in winter and spring (Mn, Pb). Si can (to some extent) be considered a tracer for the soil source, and Pb a tracer for the combustion source, while Mn has significant contributions from

Table 1
Weekly mean concentrations with absolute uncertainties (both in ng/m^3) of Si, Mn and Pb at Station Nord in Northeastern Greenland in the year 2000

Start date	Si	σ(Si)	Mn	σ(Mn)	Pb	σ(Pb)
03-01-00	7	6	0.09	0.03	0.05	0.06
10-01-00	38	8	0.49	0.05	1.54	0.09
17-01-00	29	8	0.29	0.04	1.02	0.06
24-01-00	39	8	0.78	0.06	2.32	0.13
31-01-00	31	8	0.73	0.06	2.32	0.13
⋮						
27-11-00	13	3	0.04	0.03	0.16	0.05
04-12-00	21	4	0.26	0.04	0.73	0.07
11-12-00	46	7	0.42	0.05	1.16	0.09
18-12-00	18	3	0.11	0.04	0.22	0.06
25-12-00	24	4	0.13	0.04	0.41	0.06

The table shows the input format of the data in COPREM. Only a part of the table is shown. All the data are plotted in Fig. 4.

Table 2
The initial source profile matrix and the corresponding form matrix as set up in the illustrative example

	Soil	Combustion
Si	277200	40
Mn	1000	0
Pb	16	1
Si	0	1
Mn	1	1
Pb	0	1

The approximate composition of igneous rocks (in ppm) is used in the soil profile, in which the zeros in the soil column of the form matrix fix the Si/Pb ratio.

Table 3
Output data

Date	Soil	Combustion
03-01-00	0.16	0.11
10-01-00	0.36	2.33
17-01-00	0.24	1.52
24-01-00	0.35	3.62
31-01-00	0.23	3.60
⋮		
27-11-00	0.15	0.21
04-12-00	0.22	1.12
11-12-00	0.52	1.74
18-12-00	0.23	0.33
25-12-00	0.27	0.58
Average	1.00	1.00

Source strengths calculated by means of COPREM on basis of the data summarized in Tables 1 and 2. Only a part of the table is shown. In the final calculation a gauge rule is needed for the separation of the two factors in the matrix product in Eq. (1). Here, the strength values of the individual sources in the source strength matrix are normalized to unity in average, and, in consequence, the profile matrix elements (Table 4) will be normalized to the average source contributions.

both sources. A more complicated example of the use of COPREM in the Arctic, involving 22 elements and 4 sources, has been presented in Heidam et al. (1999).

The format of the input data is shown in Table 1, and the initial profile matrix and the form matrix defining the imposed constraints are shown in Table 2. Mixing of

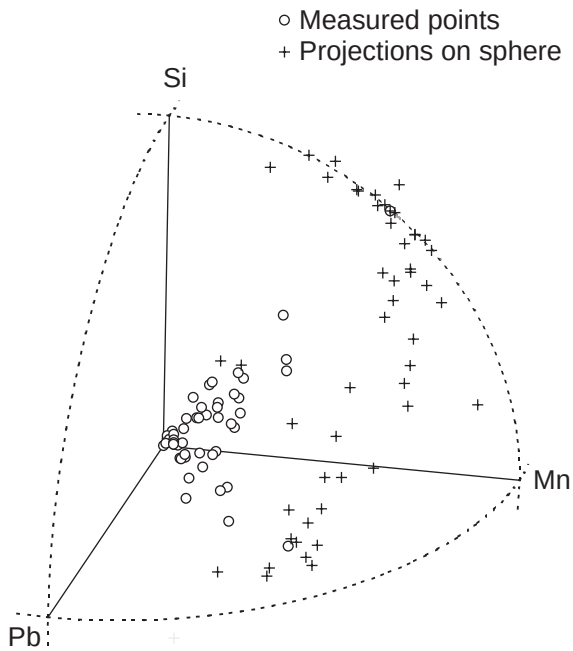

Fig. 1. 3D-view of the data in Table 1. The plotted data were normalized by division with 7.5 × the standard deviation. The dotted lines indicate intersections with a unit sphere on which radial projections of the measured points are drawn to improve the visualization. The three elements Si, Mn and Pb originate from two sources, a crustal source (soil) and an anthropogenic source (combustion). All measurements are positioned close to a 2D subspace (a plane) in the non-negative corner of the measurement space. The main objective is to find two source vectors that not only lie in the subspace, but also comply with physical restrictions, i.e., they must be non-negative, and the measurements must lie between the vectors.

Table 4
Output data

	Soil	Combustion	$\chi 2$	v
Si	70.224	5.20	16.4	
Mn	0.266	0.17	52.1	
Pb	0.004	0.65	9.0	
Total			77.5	55

The resulting source profiles are the average source contributions (in ng/m^3) due to the normalization of the source strengths (average = 1, see Table 3). χ^2 = chi-square and v = degrees of freedom. The χ^2 is a little too large to be explained by the statistics alone, $P(\chi^2 > 77.5, v = 55) = 2.4\%$.

the soil profile with contributions from the combustion source is prevented by fixing the Pb/Si-ratio in the soil profile to the small (essentially zero) ratio found in igneous rocks. The opposite mixing is not prevented, because Si cannot be considered to be a true tracer for the soil source, as Si is also emitted by coal combustion. The output format of the source strength matrix is shown in Table 3, and the resulting profile matrix, χ^2 and v in Table 4. A 3D-plot of the input data is shown in Fig. 1, where the existence of the two main sources is readily observed. The first step in the iteration process is shown in Fig. 2, where the projections of the measured points on the subspace in the angular interval between the two initial source profiles are shown. The route of iteration and the resulting source profiles are shown in Fig. 3. The fitted values were calculated by Eq. (1) and plotted together with the measured data in Fig. 4. The large Si/Pb ratio in the initial combustion profile was chosen to ensure minimum soil source strength values in the wintertime when Northeastern Greenland is covered by snow. A mathematically equivalent solution could be obtained without any Si in the combustion profile, but by doing it this way we retrieve valuable information about the upper possible limit of the Si content in the combustion source profile.

By inspection of Fig. 4 it is evident that the two sources are sufficient to make a good fit. The χ^2 is a little too large to be explained by the statistics alone, but this is probably due to source profile fluctuations in the sampling period. A closer inspection will reveal that the Si/Mn ratio is slightly higher in the summer than in the spring. This indicates that two different soil sources are at work: long-range soil dust with a lower Si content in the spring, and local soil with a higher Si content in the summer.

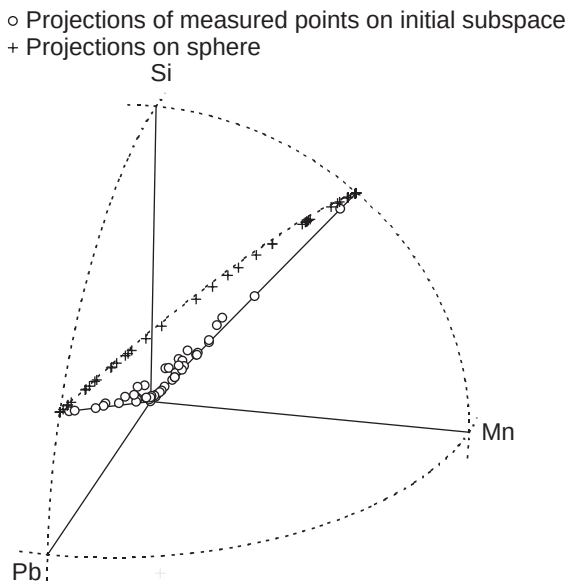

Fig. 2. First step in the iteration process. Model values are found by projection of the measured points on the subspace defined by the initial source vectors (in this case a plane). Projections outside the field between the vectors are moved to the nearest point on a vector line.

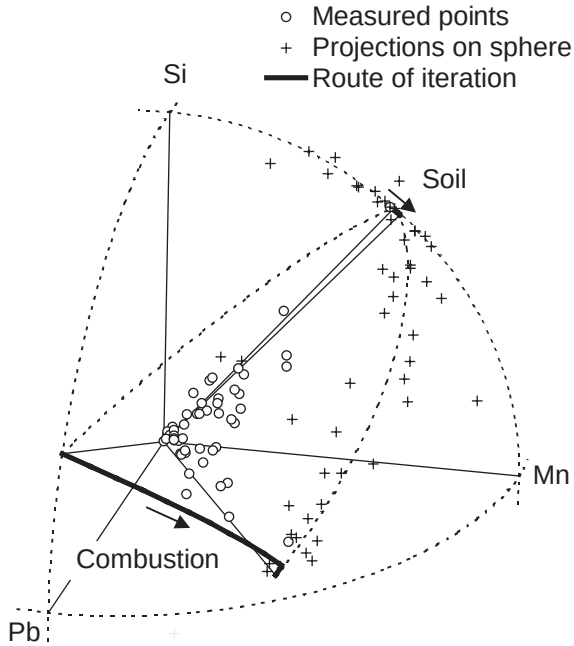

Fig. 3. The course and the final result of the iteration are guided by the choice of the initial source vectors and the constraints defined in the form matrix. In the illustrative example a surplus of Si in the initial combustion vector is chosen to ensure maximum crustal content in the resulting combustion vector. The route turns sharply at the end, governed by the non-negativity constraints imposed on the source strengths. The soil vector follows a route with constant Pb/Si ratio, defined by the Pb and Si values in the initial soil profile and the corresponding zeros in the form matrix (see Table 2).

Acknowledgements

I want to thank my colleague Kåre Kemp for valuable inspiration. The Danish Environmental Protection Agency has since 1994 supported the sampling and analysis of aerosols and gases at Station Nord in Northeast Greenland, which constitutes a part of the Danish contribution to the Arctic Monitoring and Assessment Program (AMAP).

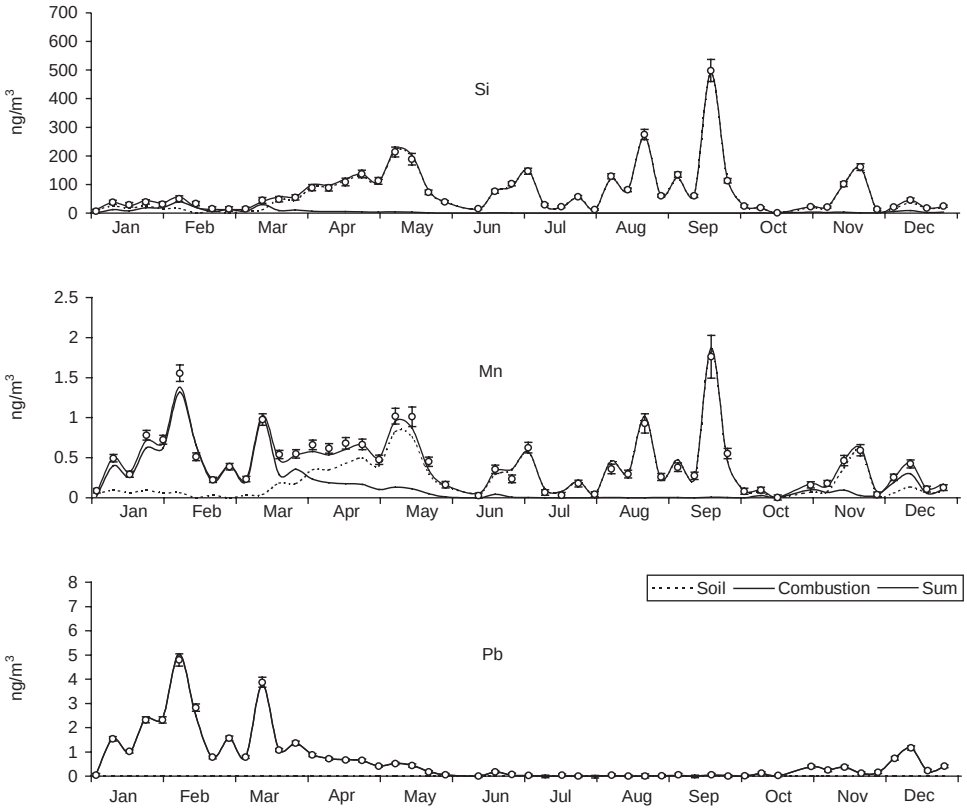

Fig. 4. Fits to the data in Table 1. The source contributions ('Soil' and 'Combustion') were obtained by multiplication of the numbers in the resulting profile matrix (Table 3) by the numbers in the source strength matrix (Table 4). The fit is the calculated sum according to Eq. (1).

References

Christensen, W.F., Sain, S.R., 2002. Accounting for dependence in a flexible multivariate receptor model. Technometrics 44, 328–337.

Gleser, L.J., 1997. Some thoughts on chemical mass balance models. Chemometrics and Intelligent Laboratory Systems 37, 15–22.

Heidam, N.Z., Wåhlin, P., Christensen, J.H., 1999. Tropospheric gases and aerosols in Northeast Greenland. Journal of the Atmospheric Sciences 56, 261–278.

Henry, R.C., 1987. Current factor analysis receptor models are ill-posed. Atmospheric Environment 21, 1815–1820.

Henry, R.C., Lewis, C.W., Hopke, P.K., Williamson, H.J., 1984. Review of receptor model fundamentals. Atmospheric Environment 18, 1507–1515.

Lawson, C.L., Hanson, R.J., 1974. Solving Least Squares Problems. Prentice-Hall Series in Automatic Computation. Prentice-Hall Inc, Englewood Cliffs, NJ.

Paatero, P., 1997. Least square formulation of robust non-negative factor analysis. Chemometrics and Intelligent Laboratory Systems 37, 23–35.

Paatero, P., 1999. The multilinear engine—a table-driven least squares program for solving multilinear problems, including the *n*-way parallel factor analysis model. Journal of Computational and Graphical Statistics 8, 854–888.

Swietlicki, E., Kemp, K., Wåhlin, P., Bartnicki, J., Jalkanen, L., Krejci, R., 1999. Source-receptor relationships for heavy metals in the European atmosphere. Nuclear Instruments and Methods B 150, 322–331.

Upadhyay, N., Oanh, N.T.K., Wåhlin, P., 2003. Source apportionment of $PM_{2.5}$ and $PM_{10-2.5}$ particles in Bangkok using receptor models, in preparation.

Wåhlin, P., 1993a. A multivariate receptor model with a physical approach. Proceedings of the Fifth International Symposium on Arctic Air Chemistry, Copenhagen, 8–10 September 1992; NERI Technical Report No. 70.

Wåhlin, P., 1993b. A receptor model applied to aerosol data from Northeastern Greenland. Proceedings of the Fifth International Symposium on Arctic Air Chemistry, Copenhagen, 8–10 September 1992; NERI Technical Report No. 70.

Wåhlin, P., 1996. One year's continuous aerosol sampling at Summit in Central Greenland. Chemical Exchange Between the Atmosphere and Polar Snow. NATO ASI Series, Vol. I 43. Springer, Berlin, Heidelberg.

Wåhlin, P., Leong, C.P., Lim, S.F., 1998. Source apportionment in Malaysia using aerosol samples collected during

haze and non-haze periods. Proceedings of Environmental Strategies for the 21st Century, Singapore, 8–10 April 1998.

Wåhlin, P., Palmgren, F., Van Dingenen, R., 2001. Experimental studies of ultrafine particles in streets and the relationship to traffic. Atmospheric Environment 35, S63–S69.

Yates, F., 1933. The analysis of replicated experiments when the field results are incomplete. The Empire Journal of Experimental Agriculture 1, 129.