Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 40 (2006) 5759–5767
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Off-line diagnostic analyses of a three-dimensional PM model using two matrix factorization methods

Jinyou Liang[a,*], Ajith Kaduwela[a,b], Bruce Jackson[a], Kemal Gürer[a], Paul Allen[a]

[a]Air Resources Board, California Environmental Protection Agency, 1001 I Street, Sacramento, CA 95814, USA
[b]Department of Land, Air, and Water Resources, University of California at Davis, Davis, CA, USA

Received 26 October 2005; received in revised form 28 April 2006; accepted 11 May 2006

Abstract

Error diagnosis of fine-grid photochemical transport models (CTM) has become a formidable task, which requires thorough understanding of complex microphysical and photochemical processes in the atmosphere as well as scientific computing. In an initial modeling exercise conducted for the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (CRPAQS), abnormally high, unrealistic, PM sulfate concentrations were simulated in central California. To aid the error diagnosis, two matrix factorization methods, namely absolute principal component analysis (APCA) and an efficient non-negative matrix factorization method (NMFROC), were used to analyze the relationships among the input and output parameters of a CTM for PM modeling and to apportion the relative importance of individual factors to an abnormal sample. The APCA method corroborated sciences implemented in the PM model, but failed to apportion the relative importance of individual factors to PM sulfate in an abnormal case. On the other hand, the NMFROC method performed well on the apportionment of an abnormally high PM sulfate. The factors produced from the NMFROC method shared common features with the APCA method, but significant differences remain between the two methods, which can be understood from their difference in methodology. Subsequent PM modeling results were shown to validate the results from the NMFROC method.
Published by Elsevier Ltd.

Keywords: Corroborative analysis tool; Air quality model; Sulfate; San Joaquin Valley

1. Introduction

Grid-based photochemical transport models, such as CMAQ (USEPA, 1999; CMAS, 2005) and CAMx (Environ, 2005), require inputs of three-dimensional meteorological parameters and emission rates of gas and particle species, to generate outputs for concentrations of chemical species and particle parameters. Owing to the large volume of input and output data, the error diagnosis of PM models has become a formidable task. Implementation of process analysis in PM models was shown to provide helpful information on certain processes (Tonnesen and Dennis, 2000), and sensitivity analysis tools may be also helpful (Morris et al., 2003; Zhang et al., 2005b). However, these techniques have to be run on-line, which further slows down 3D PM models that are already computationally demanding (Zhang et al., 2005a). A complementary off-line diagnostic tool is desirable

*Corresponding author. Tel.: +1 916 322 1223; fax: +1 916 327 8524.
E-mail address: jliang@arb.ca.gov (J. Liang).

1352-2310/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.atmosenv.2006.05.035

especially if the computer resource is an issue as for the 2000–2001 wintertime PM modeling in the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (Liang et al., 2006a; Magliano and McDade, 2005; Zhang et al., 2005a).

Receptor-oriented models have been previously applied to addressing source identification and apportionment issues of water and air pollution (Winchester and Nifong, 1971; Henry et al., 1984; Hopke, 1985; Watson et al., 1990, 2001; Chow and Watson, 2002; Lewis et al., 2003). For receptor models, the application problems have to be linear, and no significant change is allowed for source profiles between the emission and receptor points. Inert or slow-reacting primary pollutants, such as elements and CO, are about linear in terms of source apportionment between sources and receptors. Photochemical products, such as ozone and secondary PM, are non-linear in terms of source apportionment, since their responses at receptors to precursor reductions at sources are often not proportional. Meteorological parameters are also non-linear in nature, since they are non-additive and source apportionment is irrelevant for them. In sum, for non-linear species and parameters, the source apportionment function of receptor models is meaningless. Matrix factorization (MF) methods used in receptor-oriented models, however, could be used to analyze the relationship between model inputs and outputs, to be elaborated below, since the mathematical algorithms of MF methods, such as principal component analysis (PCA) (Thurston and Spengler, 1985; Jolliffe, 2002) and non-negative MF (NMF) methods (Paatero, 1997; Lee and Seung, 1999, 2001; Liang and Fairley, 2006), were designed for broader applications. Using NMF as an example, its mathematical goal is to extract a number of extreme rays (or called parts, components, vectors, etc.) in the positive orthant from sample matrix to account for major features of the sample matrix. The NMF method carries no assumption to or inference from the information before the data were acquired. Thus, it leaves the interpretation of results to users in specific fields according to the properties of the sample matrix and the nature of the NMF method.

To simulate an extended 2000–2001 winter PM episode captured in the Central Valley during the California Regional $PM_{10}/PM_{2.5}$ Air Quality Study (CRPAQS), we conducted a series of simulations using CMAQ with MM5 meteorological inputs. Details about the CMAQ simulations for the above CRPAQS episode (Liang et al., 2006a,b; Zhang et al., 2005a) are not the focus of this paper. In earlier simulations, abnormally high, unrealistic concentrations of PM sulfate were produced in the model. We applied two MF methods to aid in error diagnosis, as well as corroborate model performance. An efficient non-negative matrix factorization method (NMFROC) (Liang and Fairley, 2006) and the absolute PCA method (Thurston and Spengler, 1985; Cao et al., 2005) were coded in a statistical language (R Development Core Team, 2005). First, we will briefly introduce the two MF methods in Section 2. Then, we will describe the PM modeling problem and parameters in Section 3. After that, we will present the results from MF methods in Section 4. Finally, we will conclude with a summary.

2. The two matrix factorization methods

In this section, we will briefly describe the two MF methods used in this paper. For more detailed formulation, readers are referred to Thurston and Spengler (1985) for absolute PCA (APCA) and Liang and Fairley (2006) for NMFROC.

2.1. The APCA method

PCA has been widely used in many fields (Jolliffe, 2002). PCA makes use of eigenvectors of the correlation matrix of input data matrix A with v variables and s samples, to split normalized input matrix $\mathring{A}$ (2.1) into two matrices, namely, an eigenvector matrix $D[v, v]$ that is also termed PC coefficients, and a PC score matrix ($D^t\mathring{A}$). It is common practice to discard those eigenvectors with eigenvalues less than 1, so that only p ($<v$) factors are retained. The APCA method rotates the $D[v, p]$ matrix with a scheme called varimax to reach a final coefficient matrix D^*, and calibrates the corresponding PC score matrix ($S = D^{*t}\mathring{A}$) to reach the absolute PC score matrix X, as shown in Eq. (2.2). For factor identification purposes, the correlation between variables and PCs in the samples was calculated to form a PC loading matrix. X can be used in subsequent regression against variables of interest related to samples.

$$\mathring{A}[iv, is] = \frac{A[iv, is] - \bar{A}[iv]}{\sigma[iv]},$$
$$is = 1 : s, \quad iv = 1 : v, \tag{2.1}$$