Chapter 2. SMOKE concepts

Contents

Introduction

The paradigm for atmospheric emissions models prior to SMOKE was a network-of-pipes-and-filters. This means that at any given stage in the processing, an emissions file must be a self-contained record describing each source and all of the attributes acquired from previous processing stages. Each processing stage is a filter that inputs a stream of these self-defined records, combines it with data from one or more auxiliary files, and produces a new stream of these records. Redundant data are passed down the pipe at the cost of extra input/output (I/O), storage, data processing, and program complexity. Using this method, all processing is performed one record at a time, without structure or order to the records.

This old paradigm came about as a way to avoid repeatedly searching through unindexed Fortran data files for needed information, which would be very inefficient. It is admirably suited to older computer architectures with very small available memories and tape-only storage, but is not suitable for current desktop machines or high performance computers. SMOKE developers demonstrated this when EPS 2.0 was run on a Cray Y-MP. It ran four times slower on the Cray machine (a much faster computer) than on a desktop 150MHZ DEC Alphastation 3000/300.

The paradigm implemented in SMOKE came about from analysis indicating that emissions computations should be admirably adaptable to high performance computing if the paradigm is appropriately changed. For each category of emissions (i.e., area, biogenic, mobile, and point sources), the following tasks are performed by an emissions processor:

- read emissions inventory data files;
- optionally project emissions from the inventory year to the (future or past) modeled year (except biogenic sources);
- transform inventory species into chemical mechanism species defined by an air quality model;
- optionally apply controls (except for biogenic sources);
- model the temporal distribution of the emissions, including any meteorology effects;
- model the spatial distribution of the emissions; and
- merge the various source categories of emissions to form input files for the air quality model.
- perform quality assurance on the input data and results;

Each type of emissions has its particular complexities. For all of these types, however, most of the tasks are *factor-based* -- that is, they are linear operations that can be represented as multiplication by matrices. Further examination revealed that some of the matrices are *sparse* matrices (i.e., most of their entries are zeros). SMOKE is designed to take advantage of these facts by formulating emissions modeling in terms of sparse matrix operations, which can be performed by optimized sparse matrix libraries.

Because matrix multiplication is associative (i.e., can be regrouped), SMOKE developers further rearranged the order of multiplication operations to avoid redundant computations. For example, given a stable inventory, temporal modeling is performed only once per inventory and episode. Gridding matrices are calculated only once per model grid definition. Speciation matrices are calculated only once for each chemical mechanism. These facts contribute to computational efficiency, as well as to the integrity and consistency of the model inputs generated by the emission processing system.

Sparse matrix formulation of emissions modeling

Any part of emissions processing that is based on applying factors to a series of records can be restructure as a sparse matrix computation. For anthropogenic sources, the following emissions data processing steps are factor-based operations:

- projection of emissions from inventory year to model year,
- application of emission controls,
- conversion of chemical species from inventory species to chemical mechanism species (i.e., speciation),
- spatial allocation of emissions to grid cells in the modeling domain (gridding), and
- conversion of inventory emissions to hourly emissions (temporal allocation).

Consequently, emissions modeling has been reorganized by SMOKE into the following phases:

- **Data structuring:** order the data into vectors of emissions data, so that matrix operations are appropriate;
- **Future-past projection:** compute and apply projection factors to emissions inventory;
- **Temporal processing:** allocate emissions to hours of modeling episode;
- **Chemical speciation processing:** form a sparse matrix of chemical speciation factors;
- **Spatial processing:** form a sparse matrix of spatial allocation factors;
- **Control processing:** form a sparse matrix of control factors;
- **Emissions transform:** form the product of all matrices and apply the product matrix to the emissions inventory vector.

The organization described above has a number of advantages. First of all, the processes are made much more independent and modular, thereby promoting more reuse of data. For example, the same time-stepped, source level emissions file is used for the base case and for control strategies, as well as for each grid in a nest of grids. This design also avoids many expensive recalculations. The implementation of a new control strategy involving non-mobile sources, for example, typically requires only the construction of a new control matrix and a new transform matrix, which is then used by the emissions transform module. The time required to process a new control strategy is thereby reduced to minutes, as compared to several days using traditional emissions processing that requires re-running the entire emissions preprocessor.

The computational performance of SMOKE is even greater using modern HPC architectures, since the parallelism in the vector-matrix operations is explicitly available and can take advantage of vendor-optimized libraries. The use of a specific source ordering eliminates the

searches that degrade conventional emissions processor performance. Redundancy is eliminated in the emissions records, thereby greatly reducing storage requirements and improving data integrity. However, the user must be aware of the source ordering so that matrices intended to be used with one inventory are not accidentally used with other inventory of the same size.

Ordinarily, a new control strategy does not require any of the other components of the emissions processor to be executed. However, special cases require execution of additional components. For example, if a process is changed for a source category (e.g., organic solvents are replaced by aqueous solvents) Spcmat must be executed. Similarly, a change in operating schedule for a particular source category requires Temporal to be executed.

The following sections provide additional detail on each of the major processing stages.

Data structuring

The data structuring operation reads an inventory and puts its sources into a specific sorted order, stored in SMOKE inventory files. These files store both emissions and source characteristics, such as location, stack parameters, source category codes (SCCs), etc. Matrix operations require a consistent order between related files. The ordering enables assignment of factors to be made through subscripting instead of searching. In other words, the index to the SMOKE inventory file will match the index to the speciation matrix, and assignment of speciation factors is a simple matter of selecting the same source index from both files.

The specific order used is not important for the matrix multiplication; therefore, the sorting order may be chosen to optimize some other task. For SMOKE, the sorting order was optimized for efficient use of the cross-reference and profiles approach to emissions processing. This approach is described below.

The SMOKE program Smkinven is responsible for the data structuring operation, but it is not used for biogenic sources. The Smkinven program is used to import all emissions and activity data for all source categories but biogenic sources. The program is designed to import EPS2.0, EMS-95, and inventory data analyzer (IDA) formats. The latter is the format used by the SMOKE Tool to supply the SMOKE programs with inventory data.

The structure of the SMOKE inventory files is an extremely important concept in SMOKE. As long as no sources are added or source characteristics changed, that order will be the same. This means, for example, that a speciation matrix created for an inventory need not be ever recreated, unless the inventory or speciation information changes. This speciation matrix could be used for dozens of grids intersecting the same inventory.

If , however, the inventory source list and/or source characteristics do change, then the user must rerun all subsequent processing steps. If only the emission values change in the inventory, then the Temporal step and merge steps must be rerun, because these steps actually output emissions data values. The main point is that the user should think about what data are being changed, and what processing steps must subsequently be rerun. Of course, one can always rerun all steps just

to be safe, but this strategy will ultimately cause more time to be spent on emission processing than is necessary.

The SMOKE programs Rawbio (BEIS2) and Normbeis3 (BEIS3) are responsible for the data-structuring operation for biogenic sources. More information on the data import for biogenic emissions is available in Chapter 3.

Future/past projection

The future/past projection operation constructs emission data sets for years other than a year for which an emissions inventory is available. For example, if an inventory is available for 1996, but the modeling effort involves predicting ozone levels in 2007, then the emissions inventory must be grown to the year 2007. The first step to do this is to create a projection matrix. In SMOKE, this can be done with the Cntlmat program.

If no new sources are being added from the inventory year to the future year, then the emissions values in the SMOKE inventory files can be multiplied by the projection matrices. In SMOKE, the Grwinven program performs these operations. If new sources must be added, then the Data Structuring (Smkinven) step must be rerun for the new number of sources, then the future/past Projection (Cntlmat) can be run to create the projection matrix, and the matrix can be applied to the new inventory file (Grwinven).

Temporal processing

The temporal processing operation applies factors based on the source characteristics to the emissions data from the SMOKE inventory files. The resulting emissions data vectors (not a matrix) contain hourly emissions for the inventory species. An hourly time step is assumed in the current system. Most of the calculations are implemented as sparse matrix algebra based upon temporal cross-references and profiles, augmented by the substitution of values from day- and hour-specific emissions data sets. For biogenic and mobile sources, hourly emissions values also depend on meteorology (e.g., temperature dependence of evaporative emissions). For the more advanced emissions modeling needed for biogenic emissions, the temporal model is a true simulation model driven by ambient meteorology and other data. The SMOKE Temporal program performs temporal processing for anthropogenic sources and Tmpbio processes biogenic sources.

SMOKE treats the temporal profiles used in the Temporal as *local* profiles. In other words, the profile applied to the source is adjusted based on the difference between the time zone of the source (determined by the country, state, and county) and the output time zone (selected by the user with the OUTZONE environment variable). SMOKE automatically considers daylight time when converting from a region's standard time zone to the output time zone, and SMOKE can exclude regions that never use daylight time based on the country, state, and county. While

temporal has been tested for time zones in the Western Hemisphere, we have not tested it (and it is likely that it will fail) for time zones in the Eastern Hemisphere.

For more information on time zones, see http://time.greenwich2000.com/info/timezone.htm. Note that SMOKE expects OUTZONE to be set as a positive number for time zones in the Western Hemisphere, although standard notation would list these as negative values. For example, Eastern Standard Time is listed on this site as "-5", but OUTZONE for EST in SMOKE is "5".

Chemical speciation processing

An emission inventory is built and reported for a variety of compounds or chemical classes such as CO, NOx, VOC, PM10, and SO2. However, photochemical mechanisms (e.g., Carbon Bond 4, RADM) contain a simplified set of equations that use representative "model species" to represent atmospheric chemistry. Therefore, source-specific factors are required to convert the emissions from the chemical classes in the emission inventory to the species in the photochemical mechanism. The purpose of the chemical speciation processing operation program is to produce matrices that contain the factors for converting the input emissions to the species used in the photochemical mechanism of the air quality model.

The resulting mole- and mass-based speciation matrices are sparse matrices that transform column vectors of inventory-pollutant emissions into column vectors of model-species emissions. Note that the speciation matrices depend only upon the chemical mechanism and the inventory, and they are therefore independent of other factor-based operations for emissions processing. The SMOKE Spcmat program performs chemical speciation processing.

To support the flexibility inherent in Models-3 and future needs for reactivity assessments, SMOKE additionally supports run-time, user-selected inventory pollutants and chemical mechanisms. Although the user can define any names for pollutants and species, and indeed process any data, there are limits to how many pollutants and species can be processed at one time. In the current version of SMOKE, 120 species can be processed at once. The number of pollutants allowed in the inventory varies by sources category: area sources, 19; mobile sources 54; and point sources, 15.

Spatial processing

The spatial processing operation or *gridding* combines the grid specification for the air-quality modeling domain with source locations from the SMOKE inventory file. The resulting gridding matrix is a sparse matrix that describes where the emissions occur within the modeling domain. The gridding matrix can be applied to the inventory emissions to transform source-based inventory emissions to gridded emissions.

The SMOKE Grdmat program performs spatial processing. The gridding step is different depending on the type of source being processed. For area sources, county-total emissions are spread among the cells intersecting the county through the use of gridding surrogates. These surrogates are alternative data sources that can be obtained as gridded data sets and can reasonably be expected to represent emissions apportionment among the grid cells in a county. For example, an agriculture surrogate could be used to apportion farm-based emissions. For mobile sources, the data can be provided by county as area sources are, or the data can be provided as line sources ("links"). County-based mobile emissions are apportioned with gridding surrogates as well, preferably with surrogates based on the different road types for which the mobile emissions are provided. The line source emissions will be apportioned depending on the length of the link in each cell. Finally, for point sources, emissions are apportioned to the grid cell intersecting the point.

Note that the gridding matrix depends only upon the source locations, the grid definition, and in some cases gridding surrogates and cross-references, and it is therefore independent of other steps of emissions processing.

Control processing

The control processing operation applies control factors from a control definition file based on source characteristics in the inventory. The control scenario definition contains instructions on how to change the values of emissions based on regulations on corporate activities or personal behaviors. The resulting control matrix is a diagonal (hence, sparse) matrix of source-level control factors for each inventory pollutant. Note that the control matrix depends only upon the source characteristics in the SMOKE inventory and the set of controls chosen. The control step can therefore be decoupled from the rest of the processing steps. The SMOKE Cntlmat program performs control processing for area, mobile, and point sources. However, mobile sources controls are typically implemented by using different input files in generating mobile source emission factors.

In SMOKE v1, a new type of control has been added that requires more than simple factors for changing the emissions values. Reactivity controls, explained in more detail below, also depend on a future year SCC, which means that the process causing the emissions is changing in the future year. This type of control can impact speciation profiles, as well as the base year emissions data values.

Emissions transform

Emissions transformation (also called merging) occurs when one or more of the matrices are applied to emissions data. This is typically done to prepare emissions for an air-quality model (AQM), when one multiplies the control matrix, the gridding matrix, and the speciation matrix to produce a composite transform matrix, and then applies this combined matrix to the hourly

emission vectors. This vector-matrix multiplication is implemented by the SMOKE Smkmerge program. In addition, the Smkmerge program can multiply only the gridding matrix, or only the gridding and speciation matrices to the emissions, and it can multiply these with the annual emissions or the hourly emissions, permitting the user to evaluate the emissions with various stages of processing. The Smkmerge program also combines the emissions data from two or more source categories.

Because the processing steps in SMOKE are typically independent, a change in one of those steps does not usually require redoing the other steps. In order to generate model-input data files, however, the merge step always needs to be rerun when changes are made in other steps.

Processing exceptions

A few exceptions to the idea of independent emissions processing steps must be explained. First, biogenic processing is different than processing for other source categories. The landuse data provided as input to biogenic source processing are either already gridded or are converted from county data to a gridded dataset at the start of processing. This approach is in contrast to processing for other source categories, in which gridded data files are not created until the merge step.

Second, the temporal allocation of biogenic processing is completely meteorology dependant, because no anthropogenic activities are considered for modeling these emissions. So, the temporal factors used for anthropogenic emissions simply do not apply for biogenic emissions processing.

Third, the biogenic "inventory pollutants" are fixed in the current system. Although additional data are available, the BEIS2 model groups all biogenic emissions into four categories: monoterpenes, isoprene, nitrogen oxide, and other VOCs. All speciation profiles that are applied for biogenic emissions are limited to these compounds or groups of compounds as a starting point in the current system.

Fourth, temporal allocation for mobile source emissions depends partly on gridded temperatures. This causes an interdependence between gridding and temporal allocation for mobile sources. This relationship is handled in SMOKE through the concept of "ungridding." Ungridding is a term we use to describe the concept of converting gridded temperatures to source-based temperatures. In this process, the values of the gridded temperatures are combined using a weighted average of temperatures in grid cells that intersect an area or line mobile source. These temperatures are weighted based on the gridding factors, and the weighting factors are stored in an ungridding matrix.

The fifth exception is also for mobile-source temporal allocation. SMOKE currently can use MOBILE5b to generate emission factors. For some types of evaporative emissions, MOBILE5b depends on the minimum and maximum daily temperatures. Consequently, SMOKE has an additional mobile processing step: the Premobl program. The program computes the minimum and maximum daily temperatures by source and by day. These temperature combinations are

supplied to MOBILE5b to generate the appropriate daily emission factors.

Reactivity controls

Several issues are important when addressing emissions processing requirements for reactivity assessments. First, reactivity assessment involves the replacement of one compound in the inventory by another compound. This replacement can impact emission projections, the total magnitude of the inventory pollutants, and the associated SCCs. The market penetration of the replacement compound may vary in time and space, which affects the future-year emissions. Also, the replacement compound may be needed in much greater or much smaller amounts, thereby affecting the total inventory emissions. Finally, if a different process is required to use the different compound, the SCC for that source may change.

Second, the scale of the reactivity assessment is important, and it could be local, statewide, or national. The local case could involve investigating reactivity for one source. The statewide case could be implementing a change in compound based on reactivity considerations for a State Implementation Plan (SIP), and this would affect sources across the state. The national case could involve an EPA investigation of the formulation of nationally distributed consumer products.

Third, exemptions from controls for certain sources must be permitted as part of an emissions control strategy. These exemptions can occur when a reactivity assessment determines that certain compounds and/or processes do not significantly affect pollution formation.

To address these issues, SMOKE v1 is able to target changes in a VOC for specific classes of VOC emissions, and address the spatial and temporal considerations implied by market penetration issues. Furthermore, when replacement options are being investigated, the correct replacement operations are facilitated by SMOKE. These operations include selecting sources, changing underlying pollutant emissions, changing SCCs, correctly projecting future-year emissions based in part on market penetration issues, and appropriately speciating emissions for the new compound.

Cross-references and profiles

SMOKE uses profile tables and cross-reference tables to convert the emissions resolution for gridding, temporal allocation, speciation, and for mobile source emission factor assignment. The profile tables contain the factors for converting emissions from one resolution to another. For example, for converting from daily emissions to hourly emissions, SMOKE uses 24 factors (one for each hour). Each profile entry is assigned a profile number. The cross-reference tables assign the profiles to each source. They contain source characteristics and the profile numbers to use when SMOKE matches the source characteristics in this file to a source. The cross-reference tables are applied to the sources in a stepwise manner, such that the most specific entry is always

applied. For example, if a cross-reference entry were available for a specific plant, stack, and source category code (SCC), SMOKE would apply it instead of the cross-reference entry that matched that source only by SCC. These profiles and cross-references are prepared as part of the inventory primarily from data collected and assimilated by EPA.

SMOKE handles cross-references and profile application in a very efficient manner. In reading a cross-reference file, SMOKE first sorts the cross-reference records using the same sort criteria as is used for the inventory records. Then, these records are grouped according to the "level" of matching of each of the entries. For example, all entries that match to the inventory using only state and county codes would be in one group, while entries that match to the inventory using only SCCs would be in another group. Once the cross-reference entries are grouped, SMOKE loops through all records in the inventory, and can attempt to find a matching entry in one of the cross-reference groups. The most specific groups are searched first, such that when a match is found, the other groups are not searched. Because the cross-reference entries are sorted within each group, an efficient searching algorithm can be used. When a match to one of the cross-reference groups has been found, SMOKE continues to the next source in the inventory. This approach is much more efficient than other assignment methods.

File Formats

SMOKE uses two types of file formats: ASCII files and Input/Output Applications Programming Interface (I/O API) files.

ASCII files are simply text files that most computer users are familiar with. The ASCII files input by SMOKE come in two structures: column-specific and list-directed. In column-specific files, the fields in the files must appear in certain columns in the file. Each character on a line represents a single column. For example, a left-justified entry such as:

```
TEST
```

is in the columns 1 through 4.

In list-directed files, the exact positioning of the fields on a line is not important, but the order of the fields on that line is important. The fields must be delimited, and in SMOKE, they can be delimited by spaces, commas, or semi-colons. If a field contains any of these delimiters, then it must be provided in single or double quotes.I/O API files are read and written by a library that SMOKE uses. More information on this library is available at http://envpro.ncsc.org/products/ioapi. These files cannot be viewed with a text editor because they are binary files. The binary files use less disk space than ASCII files containing the same data. They also provide for much more efficient input and output of the data, and the I/O API library provides many quality assurance features useful for all I/O, including I/O for emissions processing.

Environment Variables

Environment variables are how SMOKE communicates with its operating environment. As is described in more detail in Chapter 5, SMOKE can be used from scripts, the Models-3 Study Planner, or with the Environmental Decision Support System (EDSS) Study Planner. The are several different uses for environment variables including assigning file names, setting options, and setting operating parameters.

For example, an environment variable in the Unix environment C-shell is defined using the *setenv* command. For example, to define the FILENAME environment variable to the file *testfile.txt* in the */home/user/* directory:

```
setenv FILENAME /home/user/testfile.txt
```

Much more will be described about environment variables and the variables that SMOKE uses in subsequent chapters.

Logical File Names

Environment variables that programs use to access files are called *logical file names*. In the example above, FILENAME is the logical file name for the physical file /home/user/testfile.txt. The I/O API is based on logical file names, and since SMOKE uses the I/O API for accessing all of its files, it too uses logical file names. The benefit of logical file names is that the programs do not require that the physical files always have the same name. Instead, the logical file names can be defined each time a program is run to use whatever physical file names the user would like. In Chapters 7 through 9 the logical file names are used to reference the files, their associations with programs, and their formats.
