EPA's New Emissions Modeling Framework

Marc R. Houyoux*
U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

Madeleine Strum
U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

Richard Mason, Bill Benjey, George Pouliot
Atmospheric Sciences Modeling Division, Air Resources Laboratory
NOAA, Research Triangle Park, NC
In partnership with the U.S. Environmental Protection Agency,
National Exposure Research Laboratory

Alison Eyth and Catherine Seppanen
Carolina Environmental Program, UNC-Chapel Hill, Chapel Hill, NC

1 INTRODUCTION

Preparation of emission inventories for use in air quality modeling is a complex, difficult, and error prone process, primarily because of the large number of disparate data sets that must be combined. Our primary goal in creating Emissions Modeling Framework (EMF) is to provide a tool to manage assoicated challenges for those who prepare emission inventories (called here "emissions modelers"). The EMF software is based on the following major principles:

- User-defined quality protocols
- Implementation of protocols for efficient application by emissions modelers
- A multi-user work paradigm
- Improved timeliness and quality of emissions modeling end product
- Transparency and issue tracking of data and emissions modeling steps
- Reduced learning curve and mistakes for novice emissions modelers
- Integrated processing of criteria and toxics emission inventories
- Tools created for both EPA and the wider modeling community

2 APPROACH

The EMF is a software application that is intended to be simultaneously accessed by multiple users within an organization through a central software server, though it can also be run on a single computer as well. The EMF is not a tool for sharing data publicly on the World Wide Web. As shown in Figure 1 below, the EMF coordinates the use of several existing and new tools that emissions modelers use to accomplish their work.

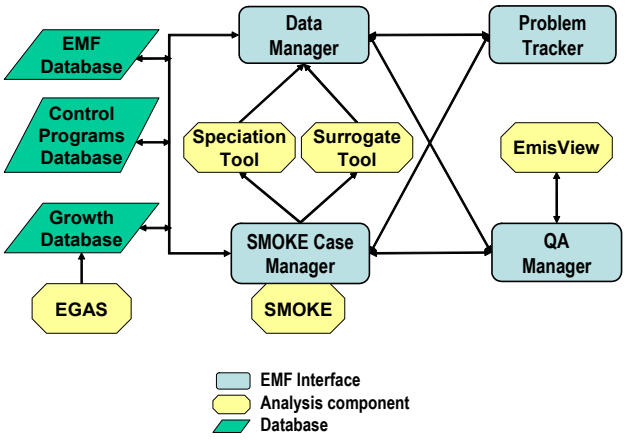

Fig. 1. Major EMF components and interactions.

The blue rectangles represent the major EMF functions of data management, case management of emissions modeling using the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system, problem tracking, and quality assurance (QA) management. The green rhombuses represent three databases on which the EMF will rely. The EMF database contains most emissions modeling data and the SMOKE case details, the control programs database contains details about control programs, estimates of emissions

*Corresponding author: Marc Houyoux, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards (D205-01), Research Triangle Park, NC 27711; e-mail: houyoux.marc@epa.gov.

reductions, and actual emissions reductions (if known). The growth database stores user-defined scenarios for emissions growth. The yellow octagons represent stand-alone software components that are also integrated with the EMF and have enhanced usability when accessed from the EMF.

The following subsections clarify the diagram above and describe how the EMF works.

2.1 EMF Database

The EMF database is a central repository of nearly all emissions data needed by the Data Manager and SMOKE Case Manager, including emission inventories, chemical speciation, spatial allocation, temporal allocation, and other SMOKE ancillary files. The only SMOKE inputs that will not be stored in the database directory are large binary files such as the meteorology data and Geographic Information System (GIS) shapefiles; however, the database will know about the existence of these files and their location on disk.

The EMF Database has features that allow emissions modelers to manage their ever-changing data needs. All datasets can have multiple versions, and emissions modelers can access any version. The EMF Database tracks when new versions are created and by whom. If a dataset has been used in a SMOKE Case and then a new version is created, the old version is still available for the old SMOKE Case.

Additionally, all datasets have extensive user-defined metadata (data about data). Some information is defined by the EMF (such as creator, creation date), but a generic metadata type can be defined by the emissions modeler for including whatever information is needed to appropriately document the origin and history of the dataset.

2.2 Data Management

The Data Management component of the EMF manages input and output of data to/from the EMF database, Control Programs Database, and Growth Database. This includes allowing the emissions modelers to perform the following:

- Import data from various formats
- Export data to SMOKE format
- Search for and find data in the database
- Edit data and create versions
- Provide metadata about the data

The Data Manager also provides data to the Case Manager, QA Manager, and Problem Tracker.

The Data Manager is closely coupled with data handling protocols that are both built into the software and customizable by users. The protocol covers how to investigate new data to make sure that it meets the users' expectations. Examples of activities included in the protocol are checking that data summaries match externally created summaries, checking that cross-references map correctly to inventories, and checking that an inventory dataset covers all of the states and counties it is purported to cover.

The Data Manager assumes that multiple users want to use the same (or similar) data at the same time. It therefore facilitates multiple emissions modelers on a team accessing the same information. Since all datasets have versions and descriptions of those versions and data changes, the Data Manager provides emissions modelers with a tool to keep organized and clear about which data are being used for any given task. This helps prevent using the wrong data for a given purpose and when used properly, provides additional documentation and transparency about data origins. The Data Manager also has concepts of public and private data, so that data can be kept private during development until the data are ready to be shared.

2.3 SMOKE Case Management

The SMOKE Case Manager is a flexible user interface for the SMOKE modeling system. The Case Manager provides the following benefits:

- Greatly decreases effort required to learn how to use SMOKE to create future-year emissions and inputs for air quality models
- Identifies when a datasets that is used in a case has a newer version available
- Organizes SMOKE cases using a searchable database of cases
- Fosters reuse and recording of SMOKE Cases, including the datasets and settings used in each
- Allows sharing and/or viewing of the same SMOKE case among multiple emissions modelers

The Case Manager provides a resource for novice emissions modelers to more easily perform emissions modeling without causing errors or delays. The Case Manager has been designed for shortening the learning curve for these new users. At the same time, the Case Manager allows advanced emissions modelers to create

specialized cases for specific purposes, and then share those with others on a project team for their use.

The Case Manager is also integrated with the quality protocols so that data created by the Case Manager can be analyzed and reviewed using standard and user-defined reports and quality checks. Since the Case Manager enables multiple users to access the same Case, emissions modelers can check each others' work as part of the quality protocol, and contractors' work can be reviewed by work managers at a lead organization.

2.4 Problem Tracking

The purpose of the Problem Tracking component is to both (1) provide a mechanism for logging, tracking, and documenting resolution of problems with data and cases, and (2) to fulfill the notification function needed by the Data Manager and Case Manager. Using the Problem Tracker, users can request to be notified of data changes about which they otherwise might not become aware. This request is based on a user's choice to "subscribe" to a particular dataset, so that they are only notified when subscribes datasets are changed by other users. The Problem Tracking component creates additional metadata to record the reasons for data changes.

2.5 QA Management

The QA Management component coordinates the implementation of both standard and user-defined QA protocols. While the EMF has default protocols included, these are editable and customizable so that QA steps can be incorporated when new types of problems are discovered or for new types of data that had not originally been part of the framework.

The Data Manager relies on the QA Manager for protocols needed for importing data, exporting data, and ensuring the multiple datasets can work together (such as inventories and cross-reference files).

The Case Manager relies on the QA Manager for protocols needed for setting up new cases, reusing or copying existing case (or templates), and for checking the results created by each case.

2.6 Control Programs Database

The Control Programs Database contains three types of control information: (1) controls and reductions that have been applied to a specific plant or source category (2) planned specific emissions reductions that are well defined but not yet installed, and (3) control programs and their general expected reductions. The first type of control information applies because the base-year for modeling (e.g., 2002) is usually 2-4 years earlier than the year that modeling work is done; therefore, information can be available on actual controls and their reductions. The second type can come from specific plans such as state and local State Implementation Plans (SIPs), and the third type comes from known Federal or State plans (e.g., Maximum Achievable Control Technology (MACT) program) for which the total reductions are set but the source of those reductions is flexible.

The Control Programs Database feeds the emissions reduction information to the Case Manager for SMOKE to use to create future-year emissions. The database contains details of the controls so that SMOKE can apply the information to emission inventories. This information includes implementation dates of the emissions reductions (actual or expected), percent reductions (and side-effect increases) by pollutant, and the facilities and/or processes affected.

2.7 EGAS and the Growth Database

The Economic Growth and Analysis System (EGAS) is used to help estimate growth or decrease in emissions based on projected economic activity changes. Version 5 of EGAS is stand-alone software that supports creating emissions growth scenarios from a variety of data sources. Users can use the data that comes with EGAS or input custom data. Future versions of EGAS will be more closely integrated with the EMF and will supply growth scenarios to the Growth Database.

The function of the Growth Database is to store emissions projection scenarios and their metadata. The EGAS is not designed to store multiple versions of the same data; the Growth Database will serve that function as well for growth information. The Growth Database may also store non-standard EGAS inputs if this seems necessary as the EMF development progresses. EGAS will remain stand-alone software, but its use will be enhanced when it is used with the EMF.

2.8 EmisView

EmisView is stand-alone software for emissions inventory and emissions modeling tabular and graphical analysis and QA. EmisView will also serve as the tool that the QA Manager will use to analyses SMOKE input and output data. EmisView will be used by the framework to

generate and display summary information, create charts and maps, and perform comparisons among datasets. These functions are needed for implementation of the QA protocols. EmisView supports both interactive analysis as well as scriptable analysis that can be run in the background by the EMF to help automate and facilitate data analysis steps of the QA protocols.

2.9 SMOKE

The SMOKE modeling system provides the core data processing function of the EMF. SMOKE is used within the EMF to compute future-year emissions and prepare emissions for input to air quality models. The Case Manager uses SMOKE programs to create these emissions modeling outputs.

As part of the EMF project, SMOKE is being updated for enhanced function and for better interaction with the EMF. The major functional enhancements are supporting additional air quality models. Specifically, SMOKE will be able to output for the AERMOD model, which stands for the AERMIC Dispersion Model, and AERMIC stands for American Meteorological Society/Environmental Protection Agency Regulatory Model Improvement Committee. AERMOD supports dispersion modeling of near-field concentrations from emissions sources, which will be used by EPA through the EMF to compliment CMAQ modeling. SMOKE output is also planned for the Assessment System for Population Exposure Nationwide (ASPEN), which has traditionally been used by EPA for the National Air Toxics Assessment (NATA) studies. SMOKE is also being updated to improve toxics support (e.g., reporting) to work better for EPA SMOKE applications. These SMOKE updates will also be supported in a stand-alone version of SMOKE. In Section 3, we provide examples of how SMOKE and the other components interact.

2.10 Speciation Tool

The Speciation Tool creates SMOKE's speciation profiles file (GSPRO) from available raw data. Creating, updating and editing the speciation cross-reference (GSREF) data is handled through the more generic Data Manager. The Speciation Tool pulls together several datasets stored in the EMF Database to create the speciation profiles, including:

- VOC and PM2.5 profiles from the SPECIATE4 database
- VOC-to-TOG factors from the SPECIATE4 project
- Definition of chemical compound lumping for model species for a chemical mechanism
- Mercury, chromium, and toxics-specific speciation profiles
- Biogenic speciation profiles
- Compound-to-species relationships for CO, NOx, SO2, NH3, and PMC.

The Speciation Tool allows users to define customized chemical groupings to support the various standard and toxics chemical mechanisms available in the Community Multiscale Air Quality (CMAQ) model. These chemical groupings can be specific regarding which toxics are included in the chemical mechanism and which are treated (essentially) as tracer species, depending on the design of the chemical mechanism within CMAQ. A critical distinction when modeling both criteria VOC (for ozone/PM purposes) and toxic VOCs, is for modelers to identify which toxics should be extracted from the criteria VOC to avoid double counting. The Speciation Tool supports these distinctions and creates SMOKE speciation profiles accordingly.

2.11 Surrogate Tool

The Surrogate Tool creates SMOKE's spatial surrogate files (AGPRO and MGPRO files) for allocating county total emissions to grid cells (for Eulerian grid models like CMAQ) or polygons (for dispersion models like AERMOD). Creating, updating and editing the spatial cross-reference data is handled through the Data Manager. The Surrogate Tool relies on the free MIMS Spatial Allocator (www.epa.gov/ttn/chief/conference/ei14/session11/eyth.pdf and www.cep.unc.edu/empd/projects/mims/spatial/) for GIS functions needed compute the surrogates. The Surrogate Tool enhances the Spatial Allocator's function by creating an entire set of spatial surrogates in a single run. The inputs to the Surrogate Tool are GIS shape files and user-defined surrogate configurations, which can include combinations of multiple spatial data in a single surrogate. Additionally, the surrogate tool will be able to create the gridded land use data needed for the biogenic emission processing.

3 EXAMPLE USES OF THE EMF

The following examples describe how an emissions modeler could use the EMF to (1) store and analyze base-year inventory data, (2) generate future-year emission inventories, and

(3) generate inputs to an air quality model. These examples are given from the perspective of an emissions modeler who has previously been using SMOKE and already has SMOKE-formatted inputs; however, the use of the EMF is not limited to that situation.

3.1 Store and analyze base-year inventory data

In this example, an emissions modeler (called below the "user") wants to get his data into the EMF and get it ready for use. The user can first import his inventory data using the Data Manager from existing ASCII file formats: SMOKE's Inventory Data Analyzer (IDA), SMOKE's One-record-per-line (ORL) format, or the NEI Input Format (NIF). The data values are then stored in the appropriate structure within the EMF Database. The data is labeled as new data by the EMF, which is not yet ready for use in SMOKE.

Using the built-in QA protocol steps for new inventory data, the user can create emissions summaries and check whether the inventory data meets the QA protocol checks expected for the particular dataset (different checks are required for point versus nonpoint emissions data, for example). If problems are found, the user can make the needed changes to the data through the Data Manager and save them along with descriptions about the changes. When the changes are completed, the EMF database assigns a new version number to the updated inventory data and stores only the changes to the inventory (for efficient storage) – the data that are the same between versions is not stored twice. The version is associated with both the data and the metadata, including who made the changes, when they were made, and the user-provided description about the changes. The user still has access to the original data as well as the updated data.

The QA protocol steps rely on the QA Manager working with EmisView to help create emissions summaries and graphics to use during the QA checks. Additionally, the emissions modeler can create additional summaries (e.g. county-Tier summaries) and graphics (e.g., bar plots, histograms, or maps) of the final emissions data for distribution to other team members or to fulfill requests for information about the data.

3.2 Generate future-year emission inventories

In this example, the emissions modeler (the "user") wants to use the EMF and the data imported in the last example to create a future-year inventory. The Case Manager is the point of access for inventory projections. Novice users can start with a projection "template" case, while advanced users can create a new template for approaches not covered by existing available templates. In this example, we assume a novice user will utilize an existing template.

The user will open the projection template and select the inventory data imported in the previous example. The user must also select growth data and control data. The user can either (1) select existing growth data from the growth database, or (2) create a new growth scenario using EGAS and/or other data, which can then be selected from the growth database. In both cases, QA protocol steps are used to ensure that the data in the growth scenario are properly "associated with" the inventory. This means that the growth information is available for the base year(s) of the inventory data and covers all of the emissions sources (by some combination of source classification code (SCC), Standard Industrial Code (SIC), MACT code, etc.).

The user must also select control data from the Control Programs Database. As the EMF evolves, EPA is seeking to provide a Control Programs Database that includes "on-the-books" reductions and their timing from known controls at the national, state, and local levels. The user could select the latest EPA-supplied information or could combine this information with user-supplied control data. The user-supplied data could be both "on-the-books" controls not yet in EPA's database or control programs for which the user wants to evaluate the impact. QA protocol steps help the user evaluate whether the control scenario selected works as expected with the inventory data.

During future-year projection with the EMF, the Case Manager obtains SMOKE input files for base-year inventory data from the EMF Database, growth data from the Growth Database, and control data from the Control Programs Database. Case Manager then runs SMOKE programs to create the future-year emissions data, which are then automatically stored in the EMF Database.

3.3 Generate inputs to an air quality model

In this example, the emissions modelers (again, the "user") wants to use the EMF to create inputs for an air quality model such as CMAQ, AERMOD, or any of the models supported by SMOKE. The Case Manager is the point of access

for creating inputs for an air quality model. As is the case for emissions projections, a novice user can choose from case templates and an advanced user can create or use templates. Through the Case Manager, the user specifies high-level settings such as the AQ model, grid settings, chemical mechanism of interest, and episode start and end. A SMOKE "case" includes all data and settings for all emissions sectors being processed for use in modeling.

The Case Manager then assists the user in ensuring that appropriate input data are included for use by SMOKE. This includes inventory data and SMOKE ancillary files. For example, if the spatial surrogates are unavailable for the grid requested, the user would be able to use the Spatial Surrogate Tool to create the surrogates needed. The Case Manager ensures that the spatial surrogates are selected that match the grid requested by the user. A similar functionality is available for speciation: the Speciation Tool can generate the speciation data needed for the particular CMAQ chemical mechanism (including mechanisms with explicit treatments of toxics).

Once the settings and input data are specified, the Case Manager starts the SMOKE run. The run is submitted to a "Queue Manager," which can run on the local computer or a remote computer. The Data Manager exports all of the data needed by SMOKE from the EMF Database that are consistent with the user's data/version selections and settings. The Case Manager then runs SMOKE in the "background" for all sectors specified in the case. The user is free to exit the EMF application while SMOKE is running and the EMF will still track the progress of the SMOKE run. Any automatic reports and graphics specified for the case are also generated during runtime relying on the EmisView component. The user will be notified when the run is complete. If the user has exited the EMF and returns, the status of the Case will be available upon re-entry to the application.

4 SCHEDULE

The EMF development is ongoing at EPA. The major milestone target dates have been set for all components except the Growth Database and Control Programs Database, and are as follows:

- Sep 2005 – Initial EmisView
- Nov 2005 – Complete Protocols
- Oct 2005 – Spatial Surrogate Tool
- Dec 2005 – Data Manager at EPA
- Dec 2005 – Speciation Tool
- Dec 2005 – SMOKE updates
- Feb 2006 – EMF - EmisView integration
- Mar 2006 – Data Manager public release
- Jun 2006 – Case Manager at EPA
- Sep 2006 – Full EMF Public Release

5 DISCLAIMER

The work presented here was performed in part under a Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.