EPA'S NEW EMISSIONS MODELING FRAMEWORK

Marc R. Houyoux
US EPA OAQPS (D205-01), Research Triangle Park, NC
houyoux.marc@epa.gov

Madeleine Strum, Norm Possiel
US EPA OAQPS, Research Triangle Park, NC

William G. Benjey*, Rich Mason*, and George Pouliot*
Atmospheric Sciences Modeling Division, Air Resources Laboratory
NOAA, Research Triangle Park, NC

Dan Loughlin
US EPA ORD, Research Triangle Park, NC

Alison Eyth and Catherine Seppanen
University of North Carolina at Chapel Hill, Carolina Environment Program, Chapel Hill, NC

ABSTRACT

The Environmental Protection Agency (EPA)'s Office of Air Quality Planning and Standards (OAQPS) is building a new Emissions Modeling Framework (EMF) that will solve many of the long-standing difficulties of emissions modeling. The goals of the Framework are to (1) prevent bottlenecks and errors caused by emissions modeling activities, (2) develop software infrastructure for performing emissions modeling in a consistent way across multiple projects, sharing emissions data in a multi-user environment, and enhancing transparency of emissions modeling, and (3) document and implement best-practice approaches for emissions modeling in support of criteria, particulate, toxics, and one-atmosphere air quality modeling. The EMF will link inventory databases such as EPA's National Emissions Inventory (NEI) with databases for emissions modeling inputs, growth data and control data. These databases will include metadata capabilities and will be connected to data tracking tools, advanced quality control (QC) tools with systematic QC procedures, computational modules based on the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system, and user interfaces for setup, application, quality assurance, and tracking of emissions modeling activities and data. In addition to supporting preparation of inventories for air quality modeling, the EMF will give non-modelers access to inventory processing functions such as speciation, growth, and control, which are often needed for policy development and other data analysis. This paper presents the full scope of the EMF and the plans and timetable for development and releases of the software.

INTRODUCTION

EPA's OAQPS is creating an Emissions Modeling Framework (EMF) that solves many of the long standing difficulties of emissions modeling at EPA. These problems stem from the complexity of emissions modeling, which requires emissions modelers to piece together many disparate data sources including emission inventories, chemical speciation factors, temporal allocation factors, spatial allocation factors, growth factors, and control factors. The inventories are usually compiled from many different data sources; this process can lead to inventory inconsistencies that are sometimes undesirable for modeling needs. Additionally, the inventories are usually developed separately from the speciation factors, temporal allocation factors, etc.; this leads to mismatches among the datasets that emissions

---

*In partnership with the U.S. Environmental Protection Agency, National Exposure Research Laboratory

modelers must ensure are resolved. With current practices and available tools, emissions modelers must manually keep track of multiple versions of data, the modeling cases for which the datasets were used, and when/why the data were replaced by newer information. Because of the volume of data and the increasingly large number of modeling cases, this manual tracking is excessively time consuming and can be prone to mistakes, sometimes leading to problems in air quality modeling and time delays. The need for systematic and integrated tools to help manage data for emissions modeling is therefore pressing.

Managing the large quantity of data and reducing the changes for errors can be accomplished through an advanced data management system that guides emissions modelers through time-proven methods of properly checking, correcting, and using data. The EMF ties such a data management system to an emissions modeling system that runs the Sparse Matrix Operator Kernel Emission (SMOKE) model. The data management and case management approaches allow modelers to retrieve previous versions of data and its descriptions to help address questions that often arise about ongoing or past work. Differences among versions can be quickly identified and associated with the reasons for changes among versions. Information about the data is stored with an expandable metadata approach. Data handling and emissions modeling quality improve with such a system, and emissions modelers have time freed for focusing on analysis, data improvements, and quality assurance. We believe that this system also will increase performance of new emissions modelers because the data handling practices will be enforced by the software instead of relying on trial and error of novice modelers.

In addition to benefiting EPA's emission modeling activities, the data management and emissions modeling functions will also be beneficial to external organizations that use emissions modeling, such as the Regional Planning Organizations (RPOs) and States. The benefits of this tool are as follows:

- Integrated quality control software with the purpose of fostering very high quality in not only emissions results, but also in data handling, organization of data, output from multi-user environments, tracking of emissions modeling efforts, and use of metadata.
- A user interface to allow non-experts to access emissions modeling capabilities such as projections, speciation, temporal allocation, growth and controls. This interface will be able to handle various separate activities related to emissions modeling in addition to the computation of emission inputs for air quality models. These activities include:
    1. Acceptance of new data (inventories, ancillary files) into the system with a quality control and sign-off process to document the data, label it for testing or operational use, availability to other system modules, and availability to other system users.
    2. Set-up and execution of emissions modeling for many air quality models for all durations and grid resolutions.
    3. Quality assurance (QA) of emissions modeling results, integrated with the EmisView tool being created for the Emissions Inventory Improvement Program (EIIP), and extensions to that tool needed for inclusion in the EMF.
    4. Reporting and archiving of reports of emission inventory data, including speciated, temporally allocated, and/or gridded emissions, with links to analysis tools for graphical and other analyses.
    5. Accessing a projection and control information database and allowing emissions modelers, economists and other non-emissions experts to create growth and control scenarios, the resulting emission inventories, and perform emissions modeling or analysis on these inventories.

- Updates to SMOKE scripting and other features that improve processing performance and support upcoming new approaches for the National Emission Inventory (NEI) development in 2005, fire modeling, and other innovations in emissions and air quality modeling.
- Create a one-model framework for criteria/PM, toxics, and one-atmosphere modeling for all air quality models used by EPA, including the Community Multiscale Air Quality (CMAQ) model, the Comprehsive Air Quality Model with Extensions (CAMx), Regional Modeling System for Aerosols and Deposition (REMSAD), Assessment System for Population Exposure Nationwide (ASPEN), and the American Meteorological Society/Environmental Protection Agency Regulatory Model Improvement Committee Dispersion Model (AERMOD).

The software and databases comprising the EMF will be released for use by external organizations on their computers.

APPROACH AND DESIGN

Overview

The EMF is a shared software application that is intended to be run by individual organizations for access by multiple users. The approach allows sharing of data within an organization through a central software server. The EMF is not a tool for sharing data publicly on the World Wide Web. The EMF has several major components that work together, as illustrated in Figure 1. As will be explained in more detail later in this section, these components are (1) **emissions modeling protocols** that describe the operational and quality control "rules" that will result in successful use of emissions modeling tools, (2) **user interfaces and middleware** that provide an expert system for guiding all types of users to the information and/or tools that they need, (3) an **emissions modeling database** that provides the master location for storage of data needed for using emissions models, (4) **SMOKE enhancements** that provide new emissions modeling capabilities, and (5) a **computational environment** of computers that can be simple (a single computer) or complex (multiple computing clusters) based as determined by the user needs and resources. This section provides details about each of these major components and explains how they work together.

Emissions modeling protocols

EPA's OAQPS and EPAs's Office of Research and Development (ORD) are working together to develop protocols to guide criteria/particulate and toxics emissions modeling activities at EPA[1]. These protocols cover all aspects of emissions modeling activities, building on past work by the Western Regional Air Partnership[2] and past experiences of all coauthors of this paper. The list below provides a summary of the items included in the protocol.

- Quality assurance of new emissions inventory data
  - Filtering out unneeded data
  - Internal consistency and data quality
  - Comparison of new inventory to previous inventories
  - Resolving multiple temporal resolutions
- Quality assurance of updated (new versions of) emissions inventory data
- Quality assurance of ancillary files
  - Accepting new data
  - Associating ancillary data with specific inventories and other ancillary data
- Running SMOKE to create inputs to air quality (AQ) models
  - Setup and associated QA
  - Populating modeling case with data files and associated QA

  - Running modeling case and associated QA
  - Quality assurance of intermediate and output data
- Running SMOKE to create future-year inventories
  - Setup and associated QA
  - Populating case with data files and associated QA
  - Running case and associated QA
  - Quality assurance of intermediate and output data
- Running SMOKE to create special analyses or summary reports
- Evaluation of new or revised software components for inclusion in EMF

The EMF integrates the protocols' methods into the user interface as an expert system, without adversely constraining the user's modeling capabilities. Deviations from the protocols are permitted to give flexibility, but these are recorded to provide transparency and future review and understanding of the quality of resulting data. This approach allows flexibility often needed in special circumstances while maintaining a known structure and standards for more routine emissions modeling runs, quality assurance, and data handling. Users can customize the protocol definitions as new emissions modeling needs arise, new data types are incorporated, or new types of modeling errors are identified and require methods to account for these.

Expert system user interface and middleware

The EMF is an open-source Java-based application for use on both Windows® and UNIX/Linux computers. The user interface is the part of the EMF that users see, and guides users to manage their data and modeling needs in ways that conform to customizable protocols for best practices in emissions modeling. The interface relies in part of the Multimedia Integrated Modeling System (MIMS) framework [http://www.epa.gov/asmdnerl/Multimedia/MIMS], which has also been used to develop user interfaces for other air quality modeling applications such as the Community Multiscale Air Quality (CMAQ) model. The middleware is software that provides behind-the-scenes functions that tie together the user interface, the database, and the emissions modeling capabilities of SMOKE.

A major goal of the EMF is to make it accessible by users with different levels of emissions modeling expertise. Table 1 provides examples of users that we expect to interact with the software.

Table 1. Examples of target user types for Emissions Modeling Framework

Emissions Modeling Expertise	User examples
Low	Wants to see the assumptions about SMOKE applications, the input data that were used, and/or output summaries
Low	Needs to access output emissions data (e.g., summaries or model-ready emissions) created by a more advanced user
Moderate	Needs to obtain SMOKE input data from the Emissions Modeling Database
Moderate	Wants to run SMOKE for a simplified modeling application, such as reusing a case run by a more advanced user to create new modeling inputs
Moderate	Wants to run SMOKE to create a new future-year inventory
High	Is responsible for QA of input and output data from SMOKE
High	Runs SMOKE from the beginning to generate inputs to air quality models

The EMF interface is designed to be a tractable tool for both novices and emissions modeling experts. This allows the EMF to be attuned to individual user needs, so that complexities do not get in the way of inexperienced users, but the details are close enough to the surface for advanced users to easily access.

The user interface and middleware provide many advantages for emissions modelers and other users. These include:

- A case manager that serves as a graphical user interface (GUI) for SMOKE, used for:
  - Creating grown and controlled inventories
  - Creating AQ modeling inputs
  - Performing additional (non-standard) quality assurance analysis
  - Providing specialized summaries that include chemical speciation, temporal allocation, and-or spatial allocation
- Management of simultaneous SMOKE runs from multiple users on one or many computers
- A database that tracks multiple versions of the same data set and can provide current and older versions, while using an efficient storage approach.
- A database manager to access to the central database of all SMOKE input files used in current and past SMOKE runs
- User notification of new versions of datasets or modeling cases, for only those datasets and cases about which a user has requested to be notified
- The ability to create and access metadata about all datasets included in the system
- Minimal administrative functions, by allowing users to control most of their own settings, such as data tracking subscriptions, and default run settings.
- Integration of the EmisView graphical QA tool[3] into a multi-user environment and including its functions in emissions modeling runs through its script interface

The software architecture is an example of a three-tier architecture with an application server. The three tiers are (1) the user interface, (2) a database management layer, and (3) a middleware layer between the other two tiers that is used to implement the modeling protocols. Since the 1990s, this architecture has demonstrated a high degree of flexibility, maintainability, reusability, and scalability over other architectures. We describe more details about how the user interface and the other components work together in the System Architecture section of this paper.

Emissions Modeling Database

As shown in Figures 2 and 3, the EMF uses three databases: the Growth Database, the Control Database, and the Emissions Modeling Database (EMD). The growth database will be a part of a future version of the Econonic Growth and Analysis System (EGAS), which is currently being revised [see http://www.epa.gov/ttn/ecas/egas5.htm]. It will store the outputs of runs of EGAS, that could use different input data (e.g., input from different economic models such as the Economic Model for Environmental Policy Analysis-Computable General Equilibrium (EMPAX-CGE) [http://www.epa.gov/ttnecas1/EMPAXCGE.htm] or the Regional Economic Models, Inc.(REMI) Policy Insight® model or different settings. It will contain enough information for SMOKE to create the inputs SMOKE needs for growth.

The control database will be a central repository of EPA's about existing and possible control programs. Although funding for this database has not yet been identified, as conceived it will adapt information from AirControlNET (http://www.epa.gov/ttnecas1/AirControlNET.htm) for criteria and particulate matter (PM) controls as well as information from the toxics control programs. AirControlNET is an existing tool that provides control strategy approaches and associated costs for reducing emissions in

future-year inventories to help develop emissions control strategies. Eventually, it will be an integrated criteria/PM and toxics database of control programs that will be able to be used to generate future-year base emissions and future-year emissions with proposed control programs included.

Unlike the growth and control databases, which are being developed through other efforts within EPA, the EMD is being developed along with the EMF and is therefore described in detail in this section. This database contains data for use as inputs to SMOKE, though a few datasets are not included because of their format and size. Specifically, the database does not contain meteorology files or the geographic information system (GIS) shapefiles that are used by the system to compute spatial surrogates. The database does store references (i.e., a server name and a full path) to these "external" datasets. It also does not contain the AQ model-ready emissions model outputs or intermediate files, but rather uses references for these files as well.

To load data into the EMD, the EMF uses importers for the types of data it will store, along with corresponding exporters that allow the database to provide SMOKE-formatted inputs. We anticipate that future EMF versions will allow SMOKE to read its own input data needs directly from the database. The list below contains the types of data that are contained in the EMD:

- Emission inventories (National emission inventory Input Format (NIF3) and/or Output Format (NOF), Inventory Data Analyzer (IDA), and One Record per Line (ORL) formats)
- Day- and hour-specific point-source inventories (NIF3 and SMOKE CEM formats)
- Spatial allocation surrogates
- Spatial allocation cross-references
- Speciation profiles and underlying raw (SPECIATE) speciation data and compound mapping information required to create the profiles
- Speciation cross-references
- Temporal allocation profiles
- Temporal allocation cross-references
- List identifying which volatile organic compounds are not are not hazardous air pollutants for use in toxics processing.
- Other ancillary data required to run SMOKE, which include country, state, and county codes; code descriptions for Standard Industry Codes (SIC), North American Industry Classification System (NAICS) codes, source category codes (SCC), maximum achievable control technology (MACT) codes, surrogate codes, and Office of Regulatory Information Systems (ORIS) codes; pollutant mappings (INVTABLE, conversion table); area-to-point mappings; holidays; grid descriptions; vehicle type and road class codes, on-road mobile process codes; and point-source stack replacements.
- Metadata about all input files
- Some specified report files output from Smkreport and EmisView

Additionally, the database stores other data used to run the EMF, but not for input to SMOKE. An example of this type of information is user settings and preferences.

SMOKE enhancements

We are making a number of SMOKE enhancements for this work. These are primarily for the further integration of toxics and criteria processing, but also for the integration of SMOKE with the EMF. These updates will continue to permit a stand-alone version of SMOKE to be run outside of the

EMF, which will help facilitate SMOKE users to continue to use SMOKE as they are now, whether or not they decide to transition to using the EMF. Two lists are provided below for (1) updates that are currently underway for the first version of the EMF and (2) updates that are planned for near-term future work expected in late 2005 and early 2006:

Ongoing SMOKEupdates:

- Read and process NIF3- and NOF-formatted data
- Update the Temporal module to process noncontinuous time periods
- Simplify SMOKE settings
- Improve reporting for growth and controls
- Update Smkmerge module to apply a growth matrix
- Update Smkreport to handle MACT and NAICS and include latitudes and longitudes of point sources.
- Updates for improved integration with EmisView for source tracking

Near-term future SMOKE updates:

- Read and process NIF4- and NOF-formatted data
- Access inventory and other ancillary data directly from the database
- Support ASPEN and AERMOD models
- Encapsulate SMOKE functions such as chemical speciation, gridding, and growth and control as library functions that can be included in other EPA software applications.

System Architecture

As mentioned in the Design and Approach section of this paper, the EMF software architecture is an example of a three-tier architecture with an application server. The three tiers are (1) the user interface, (2) a database management layer, and (3) a middleware layer between the other two tiers that is used to implement the modeling protocols.

Figure 2 shows a simplified diagram of the software as planned for EPA. The user interface runs on the desktops of individual users and interacts with the central **application server** to access and share data and modeling cases among the users. Users include EPA and NOAA staff in OAQPS and ORD. The data reside on a **database server**, which shares data among the many computational computers (shown at the bottom) via the application server. In the EPA environment, different offices have their own computers. In the figure, we show the **OAQPS compute servers** as a single entity though several such servers are included in our configuration. Additional, stand-along Linux workstations are available for use, as shown. Lastly, other compute servers such as the ORD compute servers also can access the same data as does OAQPS, but have their own cases and approaches for emissions modeling.

The EPA configuration has a large number of computers working together, but this large amount of computing power is beyond the resources for some expected users. The system has been designed to support the entire framework being housed on a single computer (most simple case), or with a central server and multiple user workstations, as shown in Figure 3**.** In this figure, the application server, database server, and compute servers are all hosted by a single computer. While the performance will not be as high as when separate servers are used (e.g., database access will be slower when the computer is running SMOKE), this will allow the system to be installed in a variety of configurations and grow as more computer resources become available.

SCHEDULE

The EMF is currently in development at EPA. Table 2 below provides a rough estimate of our expected major milestones and releases.

Table 2: EMF major milestones.

Milestone	**Expected Date**
Release updated SMOKE	August, 2005
Complete data management system and operational at EPA	September, 2005
Release beta public data management system	December, 2005
Complete EMF including case management system and operational at EPA	March, 2006
Release beta public complete EMF	June, 2006

DISCLAIMER

The research 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.

REFERENCES

1 Houyoux, M.; Strum, M.; Benjey, B.; Mason, R.; and Pouliot, G.; "EPA Operational and QA Protocols for Emissions Modeling", EPA internal documentation, *in progress*.

2 Adelman, Z.; "Quality Assurance Protocol: WRAP RMC Emissions Modeling with SMOKE", January 7, 2004.

3 Eyth, A.; Houyoux, M.; "EmisView: New Software for Visualizing and Quality Assuring Emission Modeling Data", 14th Annual International Emission Inventory Conference, Las Vegas, NV, April 11-14, 2005.

FIGURES

Figure 1: The major components of the Emissions Modeling Framework.

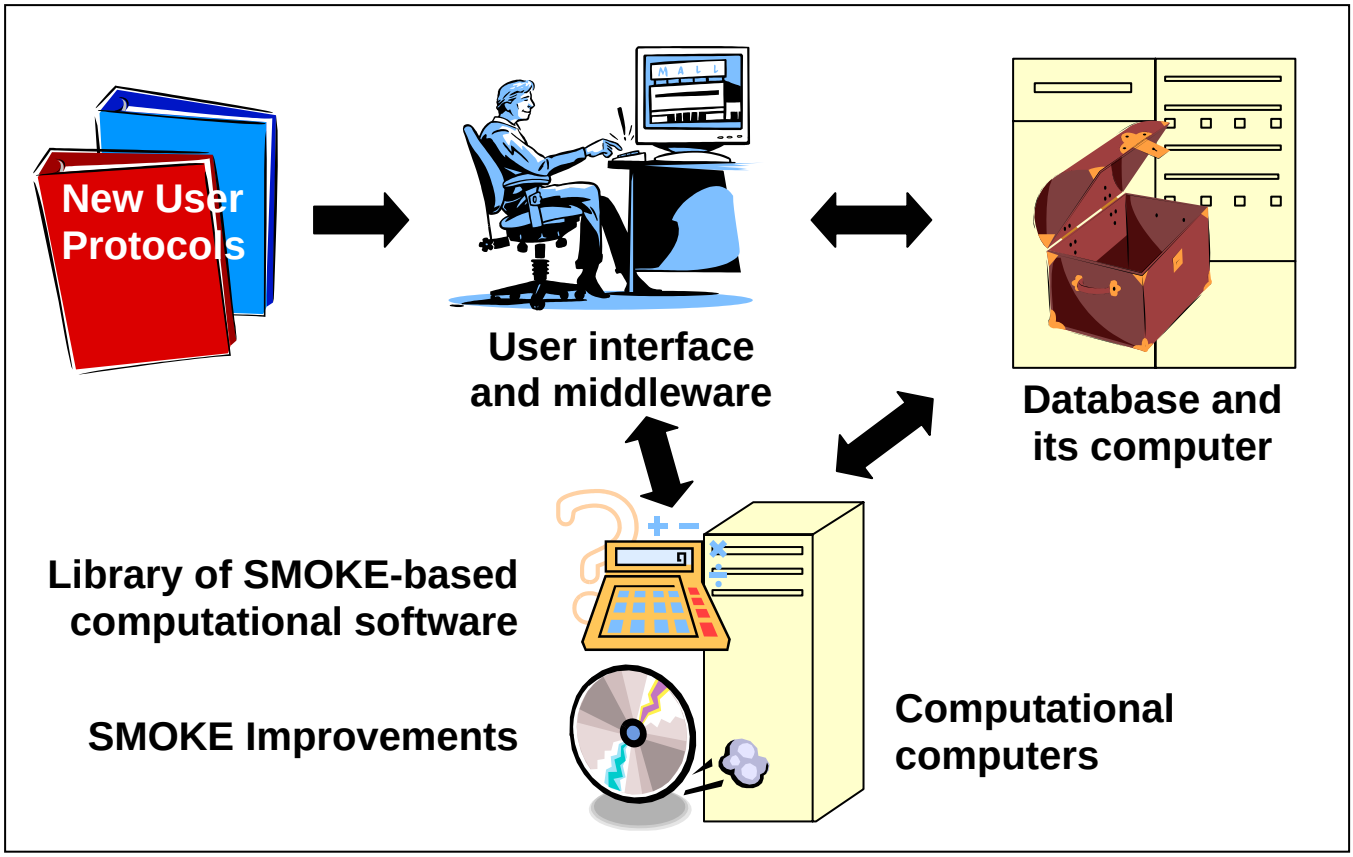


Figure 2: Example of EMF architecture used by EPA

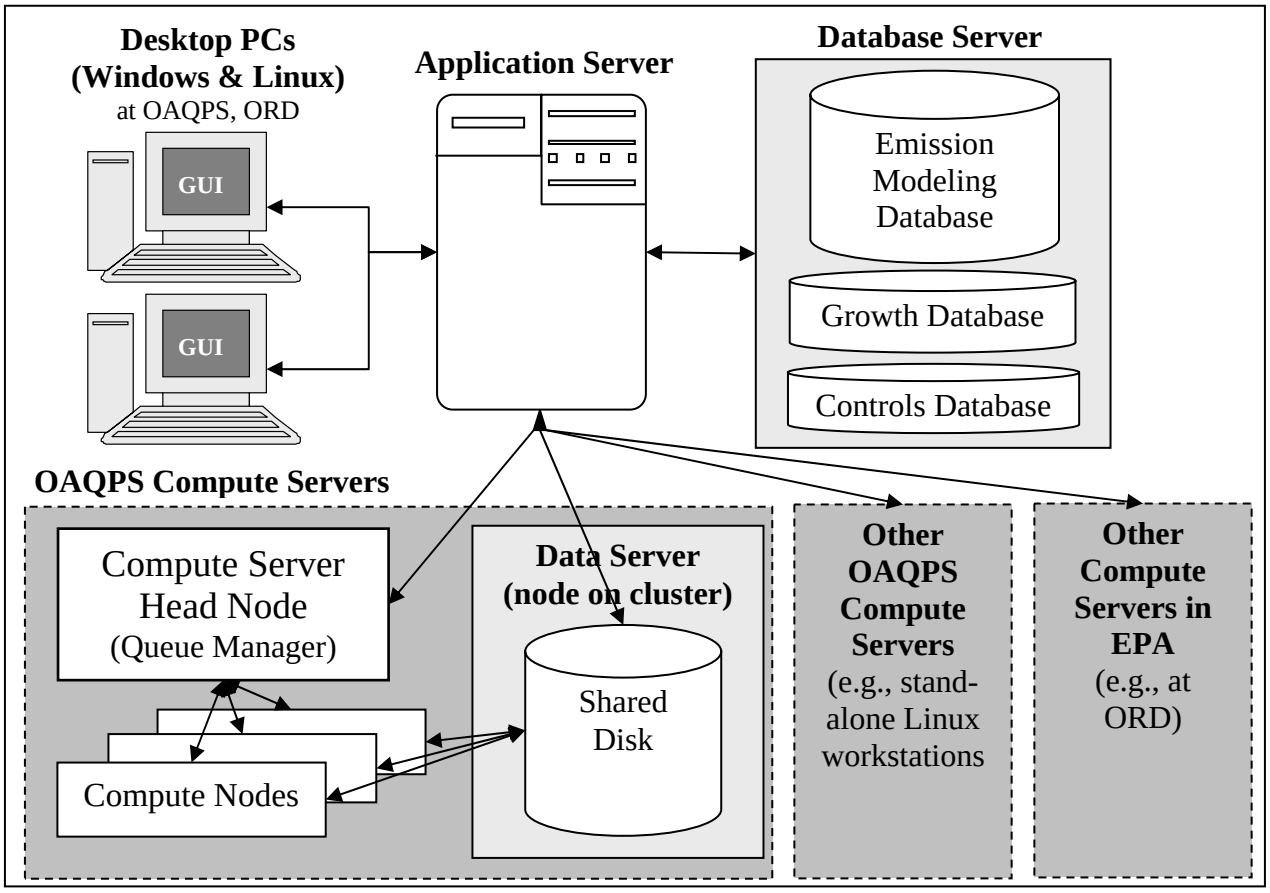
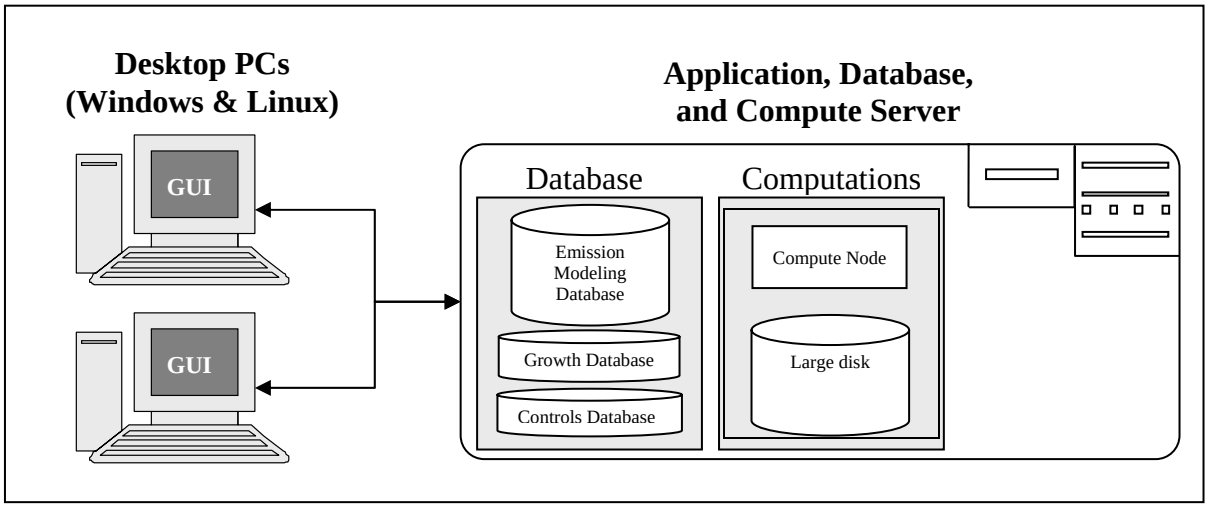


Figure 3: Simplified example of EMF architecture that consolidates all servers onto a single computer.



KEYWORDS

Emissions Modeling Framework

Data Management

SMOKE

Quality assurance

EmisView