Quality Assurance of Emission Inventory Data with the Emissions Modeling Framework and EmisView

Alison M. Eyth, Rajasooriyar Partheepan
Carolina Environmental Program, University of North Carolina at Chapel Hill
137 E. Franklin St., Chapel Hill, NC 27599-6116
eyth@unc.edu

Marc R. Houyoux
Emission Inventory and Analysis Group
U.S. EPA OAQPS (D205-01)
Research Triangle Park, NC 27711
houyoux.marc@epa.gov

ABSTRACT

EPA's new Emissions Modeling Framework (EMF) supports the tracking of quality assurance steps performed for emission inventories and other SMOKE input files. The EMF allows the user to specify a set of steps that should be performed for each type of data that will be quality assured, such as emission inventories and ancillary SMOKE input files. The EMF tracks the steps performed for each dataset, who performed them, when they were performed, their success or failure, and any notes the quality assurance staff wishes to make about the steps. Eventually, the EMF may support automated execution of QA steps using EmisView and other tools like Smkreport. EmisView is an open-source visualization tool developed in FY05 that supports the generation of plots and tables from emission inventories and other related data. The tables and plots created by EmisView can be configured and repeated in an automated fashion. EmisView was released for public use in fall 2005. Some minor updates to the Fall 2005 version have been made in FY06, and additional updates are expected during the summer of 2006. An updated version of EmisView will be released before the end of FY06, in conjunction with a version of the EMF that has more integrated quality assurance features. More information on the EMF and EmisView projects is available at http://emisview.sourceforge.net.

INTRODUCTION

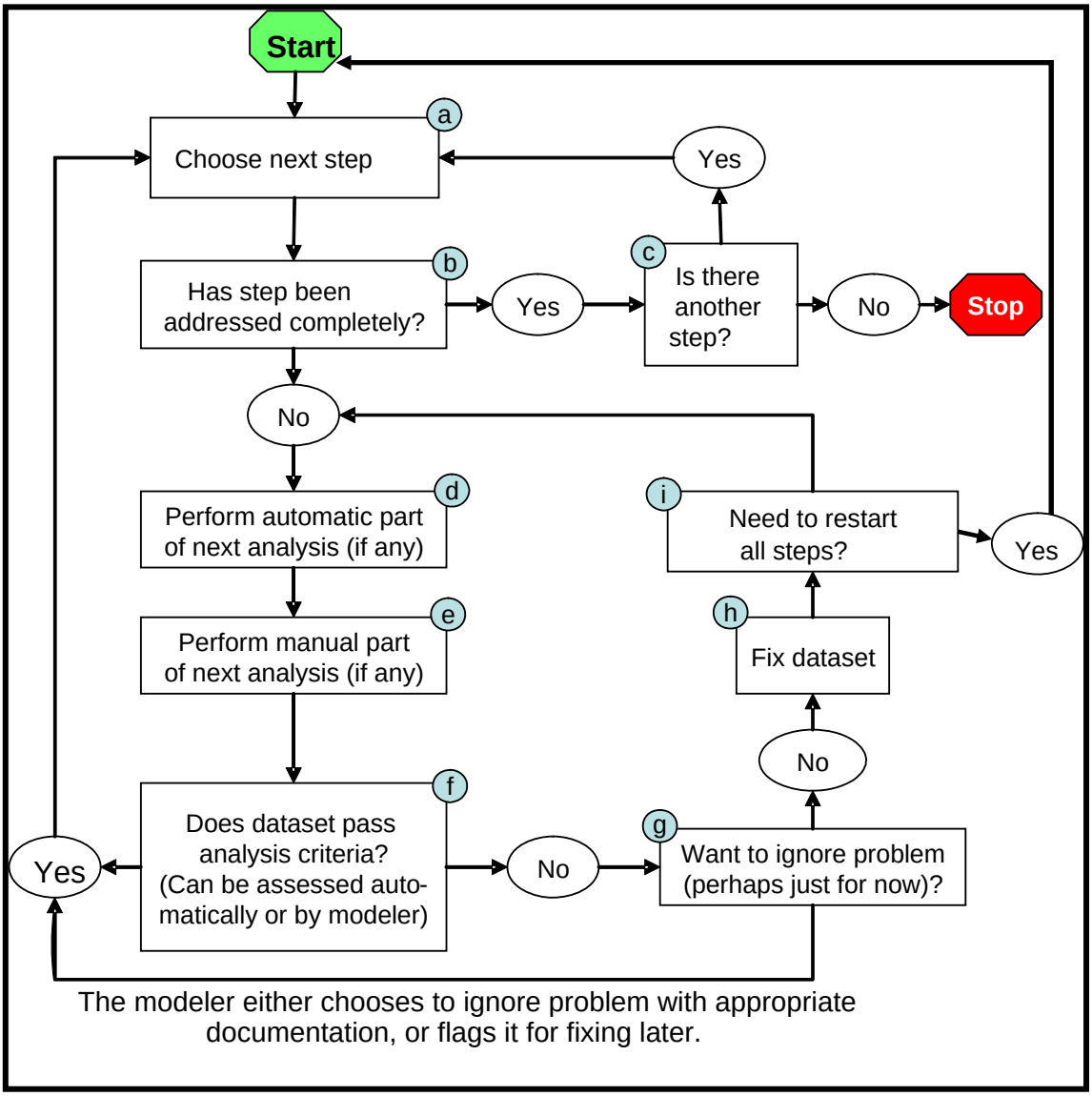
The purpose for developing the Emissions Modeling Framework is to solve many long-standing difficulties in emissions modeling at EPA. The goals of the framework are to (1) remove bottlenecks and prevent errors and reworking by EPA and its contractors, (2) create best practice approaches and protocols for emissions modeling, and (3) develop a software infrastructure for the EPA modeling community that will perform emissions modeling tasks in a consistent way across multiple disciplines, share emissions modeling data across EPA, and provide a transparent and self-documenting approach to emissions modeling. Additionally, it is hoped that the Framework will prove useful for modelers outside of EPA, where such benefits come naturally and without additional resource needs. The EMF will provide integrated quality control processes to foster high quality of emissions results, data handling, organization of data, tracking of emissions modeling efforts, and real-time accessibility of information. One goal of the EMF is to make it easier to track quality assurance steps performed for emission modeling datasets. Quality assurance step tracking features have been added to the EMF in FY06. The quality assurance steps will be used to encode the quality assurance steps EmisView was developed in FY2005 and is expected to become an important tool to facilitate quality assurance of emissions modeling datasets. In the summer of 2006, it is expected that the integration of the EMF and EmisView will be enhanced so that it will be easier to use EmisView with data in the EMF. Details of the EMF Data Management system are provided in a separate paper at this conference.

QUALITY ASSURANCE STEPS

The Concept of Quality Assurance Steps

Most modeling studies have quality assurance procedures that are applied to ensure the quality of the data before it is used for decision making. The level of formality of the quality assurance procedures varies between organizations and modeling studies. The EMF will be used to codify and record the results of quality assurance procedures performed on emission inventories and other datasets needed for emissions modeling. EPA is currently working on a draft Emission Modeling Quality Assurance Protocol document that will codify and standardize the quality assurance steps to be performed on various types of emissions modeling data used by OAQPS. Other groups also have modeling quality assurance protocols to be followed. Information in a modeling QA protocol can be distilled into specific steps to be performed, and those steps can be entered into the EMF as templates for each type of dataset. The defined steps can then be copied into the metadata of the individual datasets, and the EMF allows users to record the results of each step on any version of the dataset. An example of a process for evaluating the steps listed for a dataset is shown in Figure 1.

Figure 1: Flow Diagram for QA Protocol Steps



The boxes in this diagram provide a logical sequence for modelers to step through each of the QA steps (box *a*) for internal consistency checks, while allowing modelers to assess in each case

whether the check has been performed (even prior to receipt of the data, box *b*). If the modeler decides that the step has already been addressed, the modeler can indicate that (box c) and move on to the next step. When the modeler wants to use his own QA steps to address a topic, the modeler can do so (boxes *d* and *e*), and then iterate through each of the analyses available for addressing that topic. If the dataset passes the analysis criteria (as determined in box *f*), then the modeler moves on to perform the next step. If it does not pass the analysis criteria, then the modeler can choose to ignore and flag the problem (box *g*) or fix it (box *h*). If a problem is fixed, the modeler can choose to perform all of the steps again (which would be useful if the fix was an entirely new dataset), or recheck the current step (box *i*).

Below is an excerpt from the draft OAQPS Emissions Modeling Quality Assurance Protocol that describes the checks that should be performed prior to using a newly obtained inventory for modeling (note that more detailed descriptions of each step are available, but are not shown here):

> Proper use of inventories for emissions modeling includes investigating the quality of inventories prior to using them. The two major categories of checks we use in this document are internal checks (this section) and comparisons to other datasets (section III). This section covers those steps that seek to ensure the inventory meets criteria that depend only on the inventory itself. These are:
>
> 1. Is the format of the actual data consistent with its defined format?
> 2. Do high-level (e.g., by state) summaries of the data match summaries provided with the data?
> 3. Do the data include the expected regions, codes, and pollutants?
> 4. Are the data free from duplicate records?
> 5. Do the data cover the complete time period for which it is intended for all regions, processes, and pollutants?
> 6. Have outliers of the emissions data been identified, investigated, and if necessary, corrected?
> 7. Have outliers of other data attributes (such as stack parameters) been identified, investigated, and if necessary corrected?
> 8. Do the data have within-record consistency (e.g., are location coordinates consistent with the reported county)
> 9. Do the data have across-record consistency?
>   a. Are the emissions for multi-component pollutants both complete and free of double-counting?
>   b. Are the pollutants included for a given process similar to those reported for other instances of the same process?
>   c. Are the source categories and pollutants included for a specific geographic region consistent with other regions?
> 10. Are there new, atypical, or innovative attributes of this inventory that need to be considered and evaluated?

Quality Assurance Step Templates in the EMF

Steps described in a QA protocol such as those in the above list can be entered into the EMF as a list of QA Step Templates for a specific type of dataset. The description of and details of the steps can be tailored so that they are relevant to the particular type of dataset being quality assured. For example, evaluation of stack parameters would be included as part of Step 7 for an ORL Point Inventory dataset, but the implementation of Step 7 is different for an ORL Nonpoint Inventory. The EMF distinguishes between the concepts of Dataset Types and Datasets. A Dataset Type has a name and refers to a format

for the data on disk and a schema for tables in the database along with other features. A Dataset has a Dataset Type to indicate the type of data it represents, but it also has a lot of specific "metadata" such as the time period to which it refers, a region or grid, sectors to which it refers, and actual data values. The information associated with a Dataset Type is accessible to a Dataset. For example, "QA Step Templates" are associated with Dataset Types. These templates can be used to define the "QA Steps" that are associated with specific Datasets.

An example of the user interface for adding QA Step Templates to a Dataset Type is shown in Figures 2-5. First, the user brings up the Dataset Type Manager as shown in Figure 2. Next the Dataset Type to which the QA Step Template is to be added is selected by checking the box in the Select column and the Edit button is pressed. This brings up a Dataset Type Edit window for that Dataset Type as shown in Figure 3.

In Figure 3, the list of QA Step Templates is shown in the lower portion of the window. To add a new template, the user presses the Add button, which causes a "New QA Step Template" window to appear. Some templates have already been entered in this example. To remove one or more templates, the user checks the checkboxes in the Select column and then presses the Remove button. To change information about one or more templates, the user checks the checkboxes in the Select column and then presses the Edit button. This causes "Edit QA Step Template" windows to appear for each selected template. An example is shown in Figure 4.

In each "Edit QA Step Template" window, the user can specify the following information: a name for the QA Step Template, a program to use for the step (or "None"), arguments that will be given to the program when it is run, the order in which the step to be run, whether the step is required or optional, and a description of how to perform the QA Step. Note that the order is specified as a floating point number to make it easy to group related steps and to insert new steps as needed. Once the user is done editing the information, pressing the Save button will save and close the window; to ignore any changes, the Close button can be pressed. The "New QA Step Template" window looks just like this window, except that no information is filled in when it first is brought up.

Figure 2. Dataset Type Manager

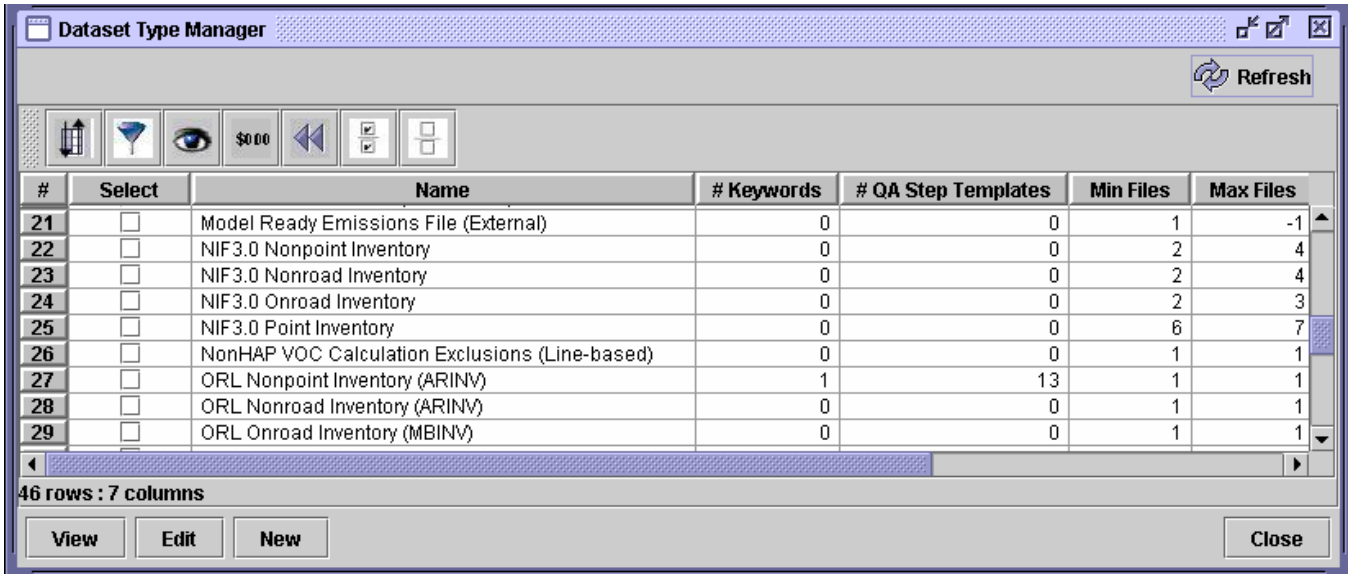


Figure 3. Edit Dataset Type Window

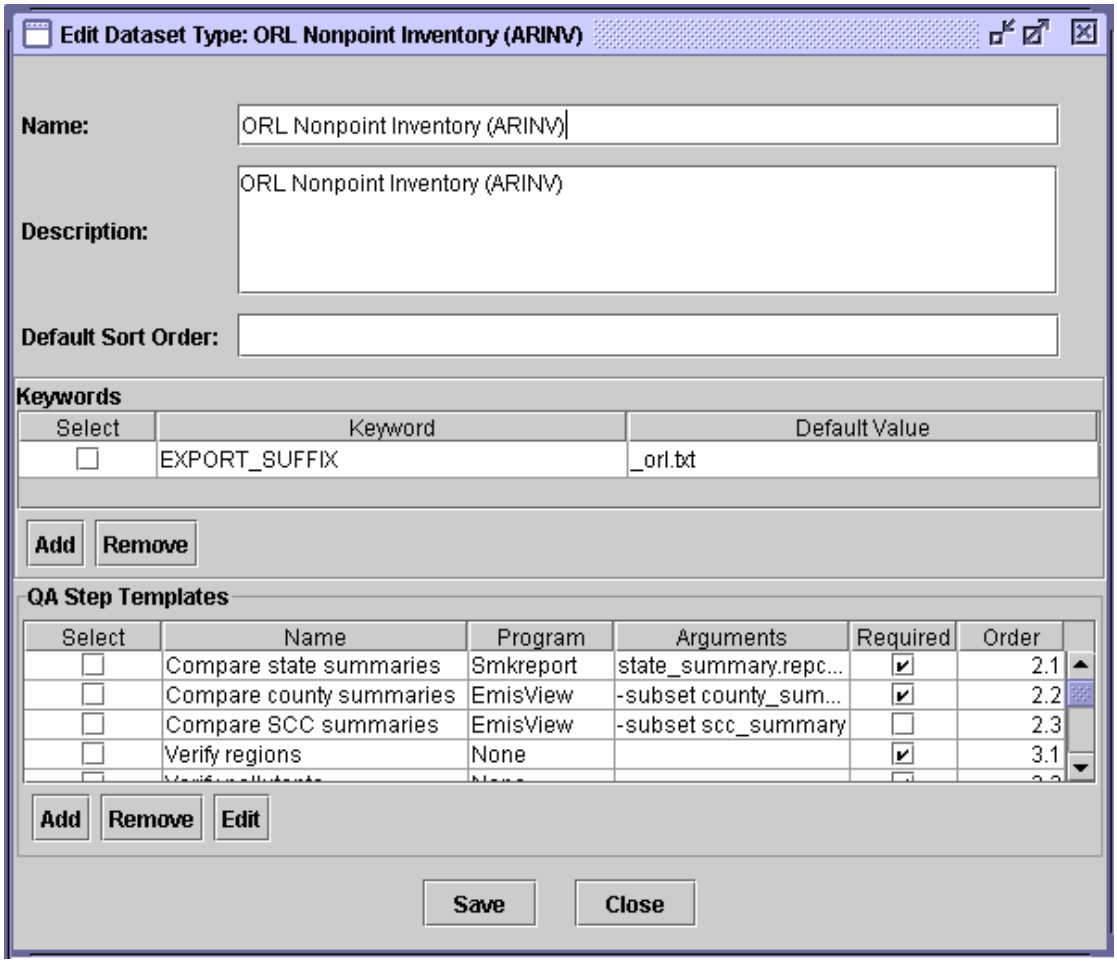
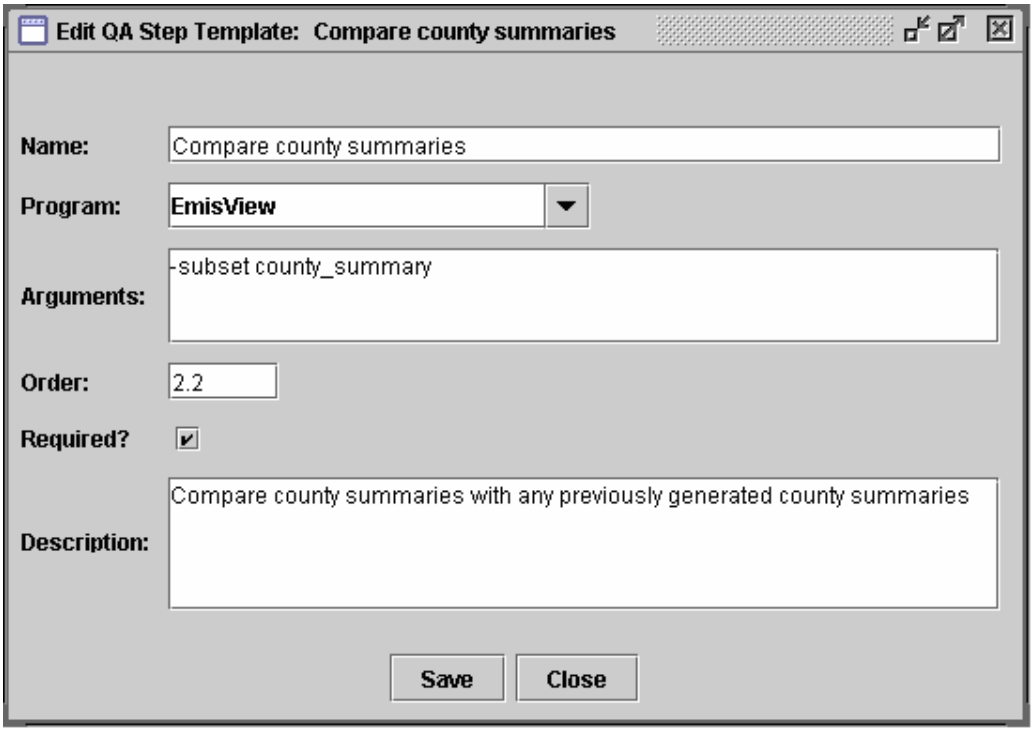


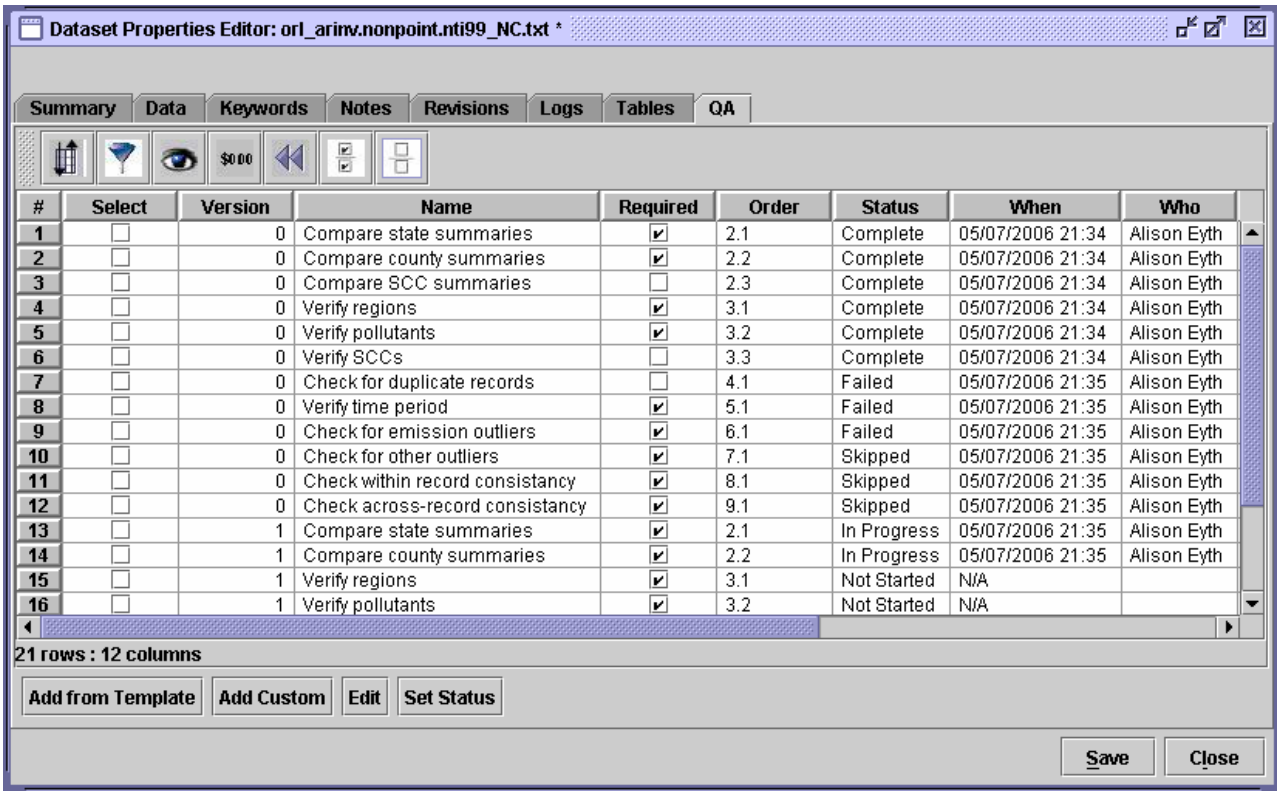
Figure 4. Edit QA Step Template Window



Tracking Quality Assurance Steps for Datasets

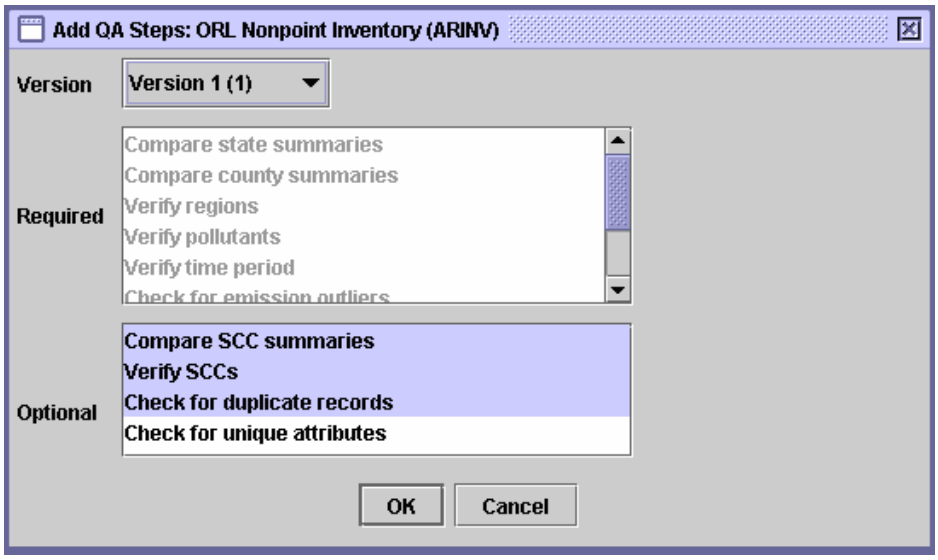
Once the QA Step templates are defined for each Dataset Type, it is then possible to quickly and easily populate the QA Steps for specific Datasets. To add QA Steps for a dataset, the user must bring up the Dataset Properties Editor for the dataset from the Dataset Manager. For more details on this process, see the companion paper on EMF Data Management (Houyoux, 2006). Once the Dataset Properties Editor is opened for the dataset, select the QA tab. An example of a tab with some QA Steps already populated is shown in Figure 5. To add QA Steps from the templates for the dataset type, users press the "Add from Template" button. This causes an "Add QA Steps" dialog to appear such as the one shown in Figure 6.

Figure 5. QA Tab of Dataset Properties Editor



#	Select	Version	Name	Required	Order	Status	When	Who
1		0	Compare state summaries	☑	2.1	Complete	05/07/2006 21:34	Alison Eyth
2		0	Compare county summaries	☑	2.2	Complete	05/07/2006 21:34	Alison Eyth
3		0	Compare SCC summaries	☐	2.3	Complete	05/07/2006 21:34	Alison Eyth
4		0	Verify regions	☑	3.1	Complete	05/07/2006 21:34	Alison Eyth
5		0	Verify pollutants	☑	3.2	Complete	05/07/2006 21:34	Alison Eyth
6		0	Verify SCCs	☐	3.3	Complete	05/07/2006 21:34	Alison Eyth
7		0	Check for duplicate records	☐	4.1	Failed	05/07/2006 21:35	Alison Eyth
8		0	Verify time period	☑	5.1	Failed	05/07/2006 21:35	Alison Eyth
9		0	Check for emission outliers	☑	6.1	Failed	05/07/2006 21:35	Alison Eyth
10		0	Check for other outliers	☑	7.1	Skipped	05/07/2006 21:35	Alison Eyth
11		0	Check within record consistancy	☑	8.1	Skipped	05/07/2006 21:35	Alison Eyth
12		0	Check across-record consistancy	☑	9.1	Skipped	05/07/2006 21:35	Alison Eyth
13		1	Compare state summaries	☑	2.1	In Progress	05/07/2006 21:35	Alison Eyth
14		1	Compare county summaries	☑	2.2	In Progress	05/07/2006 21:35	Alison Eyth
15		1	Verify regions	☑	3.1	Not Started	N/A	
16		1	Verify pollutants	☑	3.2	Not Started	N/A	

Figure 6. Add QA Steps (from Template) Dialog



The “Add QA Steps” dialog allows the user to select the version of the dataset undergoing QA (shown at the top of Figure 6), and it shows the required steps and the optional steps in separate lists. The names of the steps on this dialog and the required and optional specifications are taken from the “QA Step Templates” specified for the Dataset Type. All required steps and any selected optional steps will be added to the table on the QA tab for the specified Dataset and version, except that any steps that have already been added will be ignored. When the user clicks the OK button, all of the information stored as part of the templates is copied into the QA Steps for the specific dataset and version. This information includes the name, program, arguments, required flag, order, and description. Additional fields are added to the steps that either the user or software will set for the specific dataset and version. These fields are the [QA] Status, who set the status, when it was set, a comment on the result of the step, and the name of a dataset that contains a configuration file to be used for the step (if the user wishes to specify this).

A step can also be added to a Dataset using the “Add Custom” button, which brings up a window that is very similar to the “Edit QA Step Template” window shown in Figure 4, except that all the fields are blank and the user must also choose a Version. Once filled in, the information for the QA Step will be added to the table on the QA tab. To edit the information (such as the adding a comment or specifying a configuration) for one or more QA steps, the user selects the steps from the table on the QA tab and then presses Edit. Windows like the one shown in Figure 7 will appear for each selected QA step. When a new status is selected from the Status menu, the software automatically updates the User and Date fields. The available QA statuses are Not Started, In Progress, Skipped, Complete, and Failed. The Configuration field is used to store the name of a Dataset that is used as a configuration file for the QA Step.

To set the status for multiple QA Steps at one time, the user sets the checkboxes for the steps in the Select column and presses the Set Status button. A Set Status for QA Steps window will appear similar to the one shown in Figure 8. The list of QA steps to be set is shown in a scrolled list at the top. The user then chooses the new status and the When and Who fields are automatically filled in, but can be edited if desired. The specified comment is appended to the comment fields for each of the QA Steps after the user presses OK.

Figure 7. Edit QA Step Window

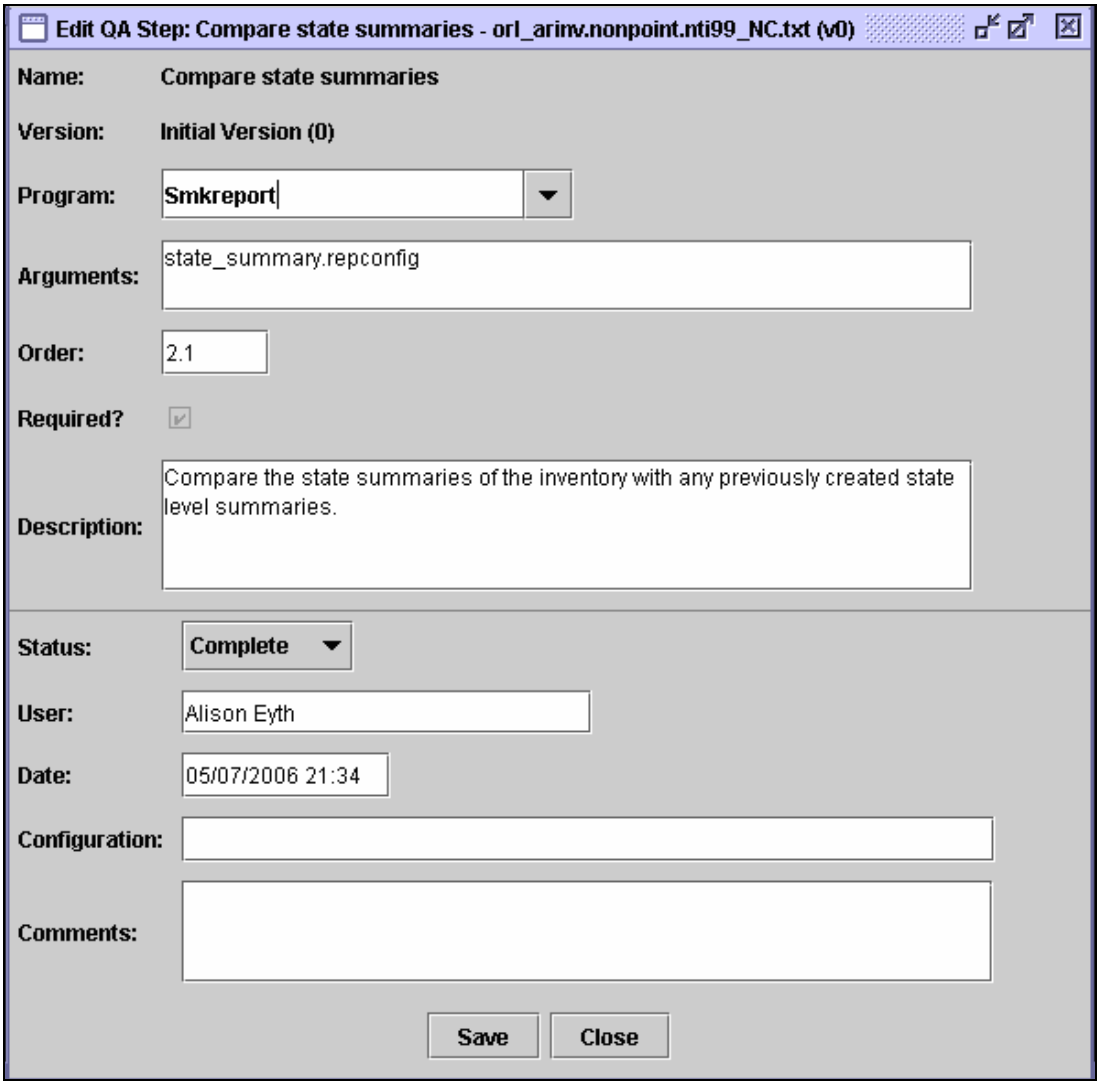
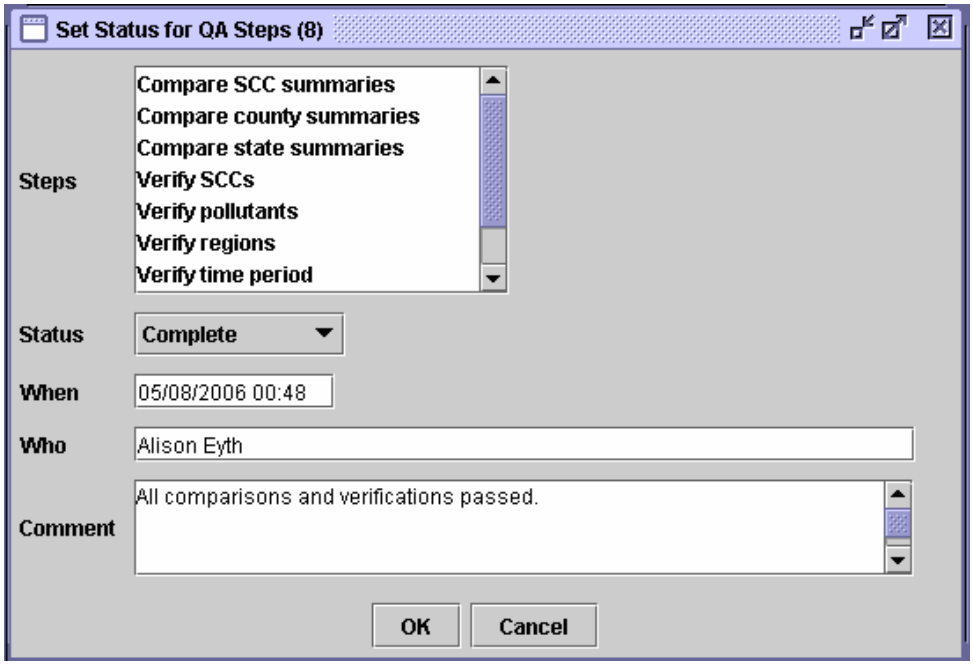


Figure 8. Set Status for QA Steps Window



Performing Quality Assurance Steps with EmisView

There are several tools that are particularly useful for performing quality assurance steps. These include some of the SMOKE programs themselves, such as Smkreport and Smkinven. Smkinven can verify the format of an inventory. Smkreport can perform high level summaries of an inventory. EmisView can also perform analyses commonly needed when quality assuring emission inventories. For example, EmisView can perform the following operations:

- identify outliers in emissions or stack parameters (e.g. find all sources with a stack height > 1000, or emission source > 10000 TPY);
- create high level summaries of emissions by SCC, MACT, NAICS, FIPS, and SIC;
- subselect and report on portions of an inventory to obtain data for only specific SCC codes, states, or counties;
- sort the inventories, or sort in conjunction with applying a filter.

EmisView presents the results of an analysis in an interactive table that supports single and multi-column sorting, filtering, and the creation of many types of plots. Version 1 of EmisView was released in September, 2005, with a minor update in October, 2005. EmisView is released under a GNU Public License and is available at http://sourceforge.net/projects/emisview. It is written in Java and can run on Windows, Linux, or other Unix systems. It can read data in National emission inventory Input Format (NIF3), Inventory Data Analyzer (IDA), and One Record per Line (ORL) formats. It reads emission inventory data from PostgreSQL (http://postgresql.org), and supports NIF inventories that have been imported into the CONsolidated Community Emissions Processing Tool (CONCEPT). For more information on CONCEPT, see http://conceptmodel.org.

To perform an analysis in EmisView, the user must first select a Dataset, Subset, and Product. Analyses are stored in the database, so they can be retrieved and rerun between sessions. The concept of a Dataset is very similar to a Dataset in the EMF – basically it is an inventory file (or set of files in the case of NIF) that has been imported into the EMF. A Subset is a description of how the data should be filtered and grouped (e.g. give data for Kentucky only as county totals by SCC). The Product is the types of tables and/or plots that will be produced. The simplest Product called "Table and Plots" extracts the specified Subset of the Dataset from the database and loads it into the Analysis Engine interactive table without any additional processing. Other products can be defined to sort and filter the data and to create specific plots. A screen shot of the Analyses tab of the EmisView main GUI is shown in Figure 9. Datasets, Subsets, and Products are stored in the database, in addition to the analyses. The other tabs look very similar, only the tables contain information relevant to the tab being shown.

An example of the window used to configure an analysis is shown in Figure 10. This Analysis is used to compute sums for each SCC for each of the Hazardous Air Pollutants (HAPs) in the North Carolina 1999 NTI inventory. The Subset, Dataset, and Product are selected on this window. Instead of selected a Subset, a custom query accessing the database tables directly can be entered. In Figure 10, the analysis will report the results summed by SCC. A unique characteristic of EmisView is the mix-and-match nature of the analyses. If a report summed by FIPS is desired, the user would simply select the ORL by FIPS subset instead. Similarly, Subsets can be applied to multiple datasets. Any datasets that can be used by the selected Subset are shown in the Dataset list. For example, if the All Records Subset is selected, all of the Datasets will appear in the Dataset list. If the user wishes to present the results in a different way, a different product can be selected. For example, the Sort by PM10 product will show the results sorted by the PM10 column (if it exists in the dataset).

Figure 9. Analyses Tab of EmisView Main GUI

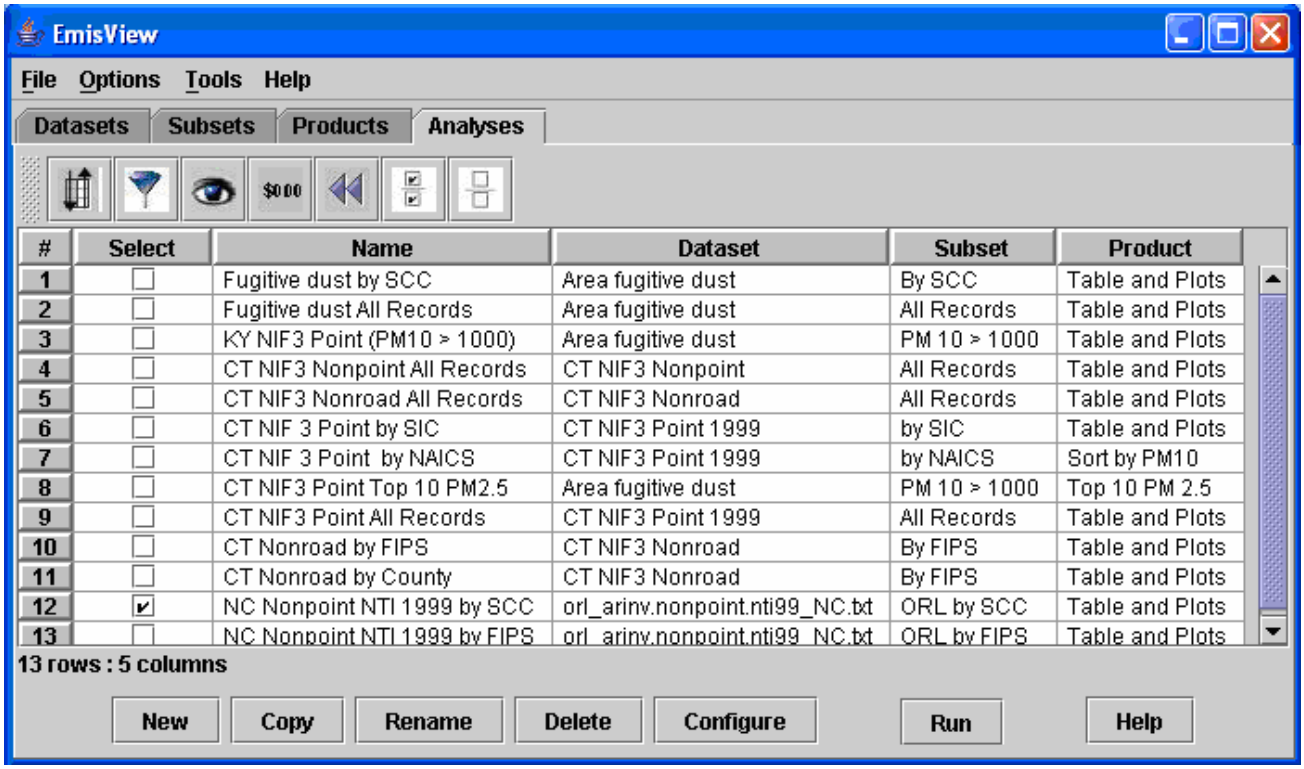
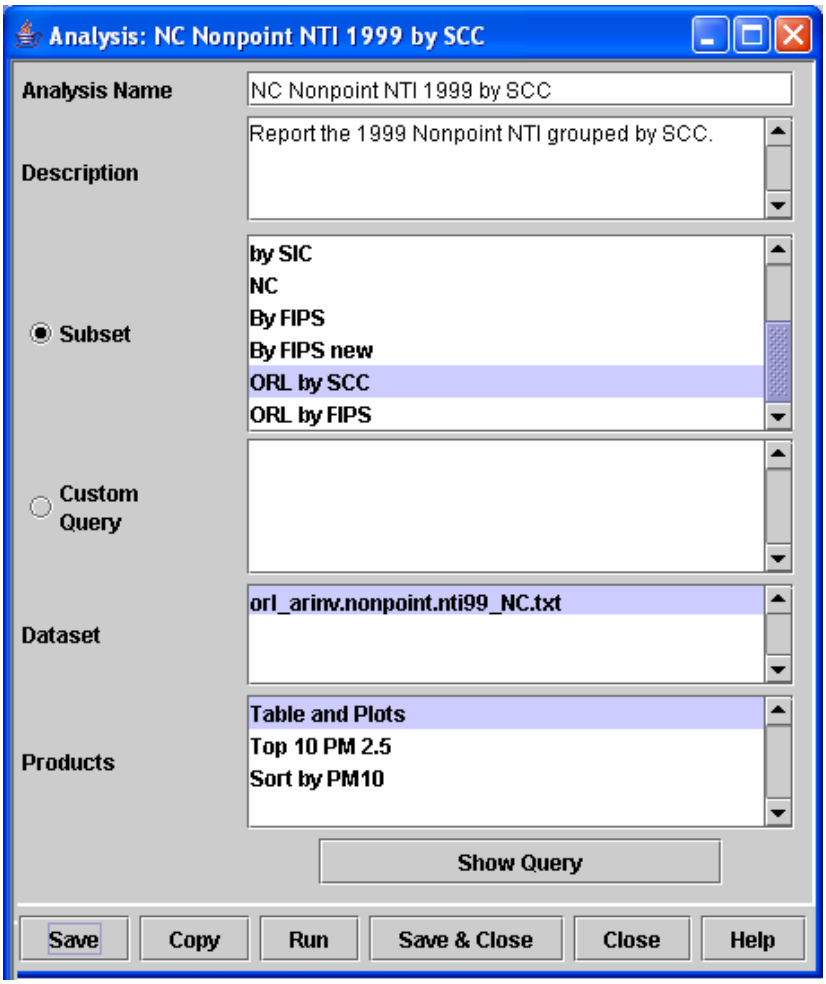
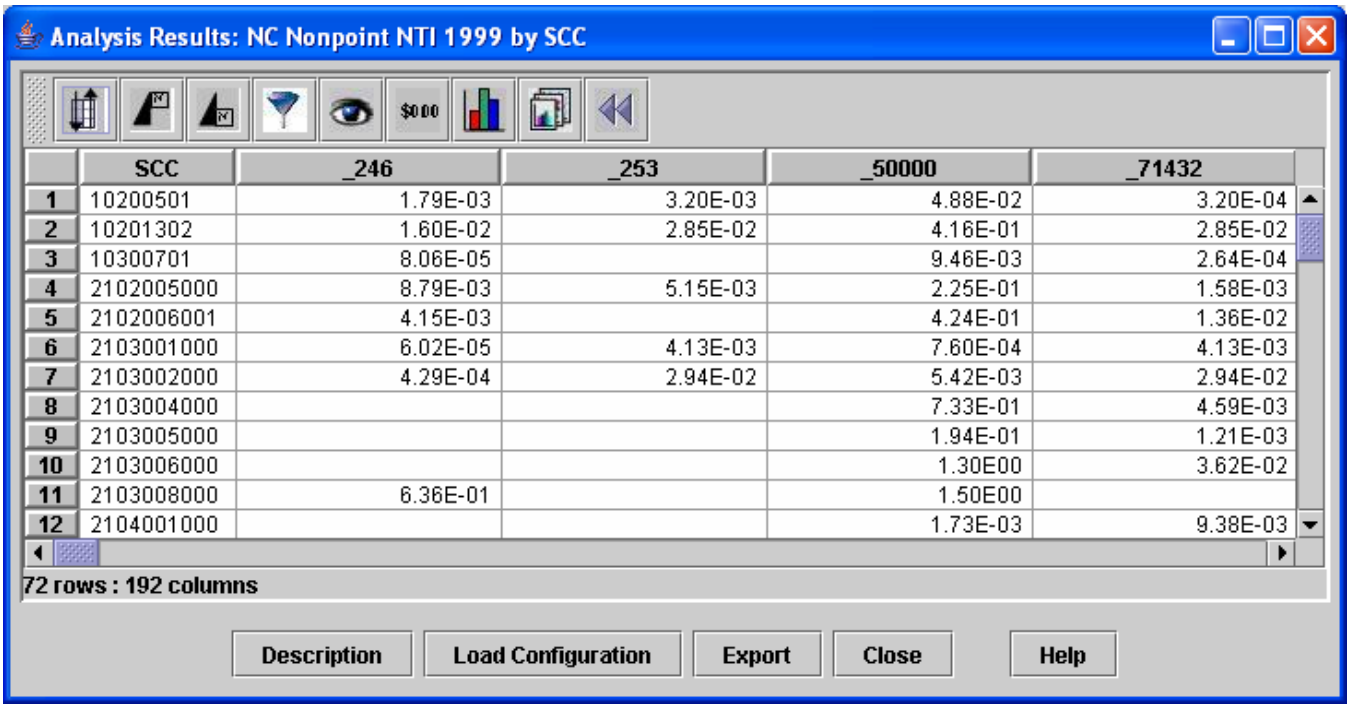


Figure 10. Analysis Configuration Window



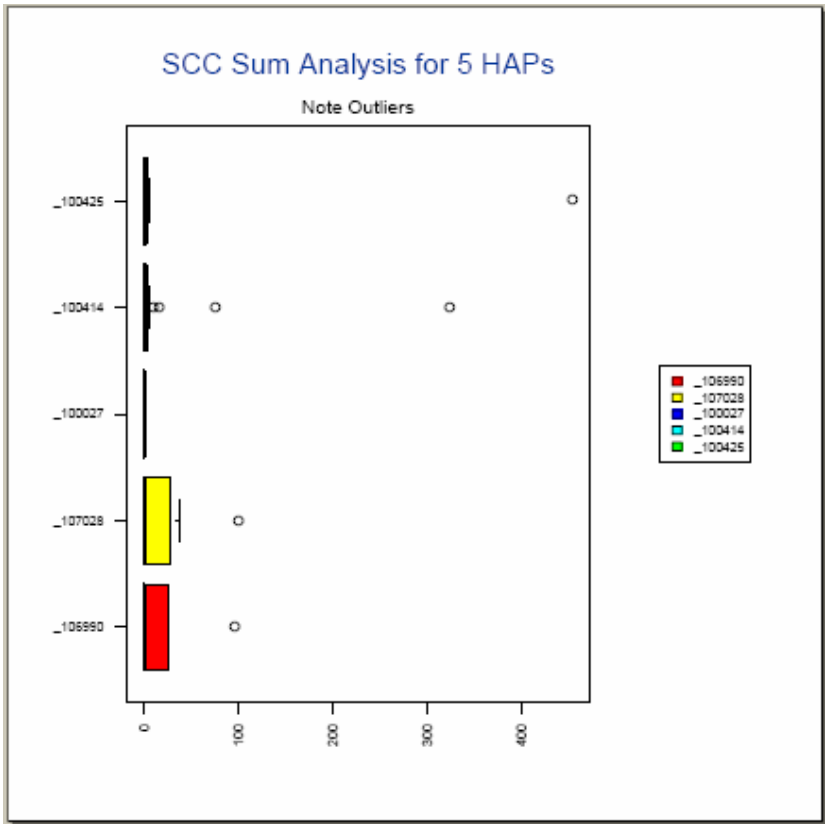
The results of the Figure 10 analysis are shown in Figure 11. The window in which they are presented is an interactive window, from which the data can again be sorted, filtered, and plots can be made using the icons on the toolbar and the popup menu. Because the dataset used is a toxics inventory, the labels for the columns are the CAS numbers of the pollutants in the inventory. An example of a plot that can be created from the Analysis Results window is shown in Figure 12. This plot shows that there are some outliers in the SCC sums that may be worth examining for validity. The other types of plots that can be generated are: CDF, discrete category, histogram, rank order, XY, line, time series, and tornado plots. All of the plots can be configured in many ways. The rendering of the data in the bar plot is especially configurable. The data can be transposed, presented horizontally or vertically and as stacked or adjacent bars. The configuration of the plots and table can be saved to a configuration file to make them easy to repeat at a later date using another set of data. Additional details on the use of EmisView can be found in the EmisView user's guide.

Figure 11. Results of Analysis



	SCC	_246	_253	_50000	_71432
1	10200501	1.79E-03	3.20E-03	4.88E-02	3.20E-04
2	10201302	1.60E-02	2.85E-02	4.16E-01	2.85E-02
3	10300701	8.06E-05		9.46E-03	2.64E-04
4	2102005000	8.79E-03	5.15E-03	2.25E-01	1.58E-03
5	2102006001	4.15E-03		4.24E-01	1.36E-02
6	2103001000	6.02E-05	4.13E-03	7.60E-04	4.13E-03
7	2103002000	4.29E-04	2.94E-02	5.42E-03	2.94E-02
8	2103004000			7.33E-01	4.59E-03
9	2103005000			1.94E-01	1.21E-03
10	2103006000			1.30E00	3.62E-02
11	2103008000	6.36E-01		1.50E00	
12	2104001000			1.73E-03	9.38E-03

Figure 12. Example of a Box Plot



UPCOMING ENHANCEMENTS

Additional work on the quality assurance portion of the EMF is planned for the summer of FY06. Because the framework for managing QA steps is already in place in the framework, the majority of the items deal with improving the integration of EmisView with the EMF. The remaining FY06 work is expected to include some of the following items, and others may be implemented in FY07:

1. EmisView will be upgraded to use the same versions of the importers as the EMF.

2. EmisView will be updated to be able to select datasets and access versions of datasets directly from the EMF database, but the ability to access non-versioned data tables, such as those used by CONCEPT will also be preserved.

3. EmisView will be enhanced to support the analysis of additional types of datasets, and if possible, a generic capability to access any table in a database will be added. This will support analysis of some of the SMOKE and hopefully CONCEPT ancillary files.

4. Additional types of summaries will be added to EmisView for ORL files, such as State and Overall summaries.

5. A simple script based interface to EmisView will be developed so that analyses can be performed and outputs produced without showing the GUI.

6. Improvements will be made to the usability of the top N analysis in configuration files.

7. The EMF will be enhanced to actually reference a dataset as part of a QA step instead of just the name of a datasets.

8. EmisView could be started directly from the EMF client.

9. Access statistical analyses from EmisView which are already available in the Analysis Engine, upon which the EmisView interactive table is based.

10. As part of the case management work in FY07, the EMF could be enhanced to automatically run some QA steps, such as those using EmisView and Smkreport.

11. Export Shapefiles with results of analyses.

CONCLUSIONS

The EMF provides a flexible mechanism for specifying quality assurance steps, such as might be described in a QA modeling protocol. The steps can be described in template form one time for each type of dataset, and can then be quickly copied into the metadata for a particular dataset. As the QA steps are performed, there EMF records the user who performed the step and the time the step was performed. The user can optionally enter a comment to describe any observations relevant to the step. This approach provides a unique, integrated solution to tracking the quality assurance performed on a dataset, and is flexible enough to extend to almost any emissions modeling related dataset. EmisView has been shown to be a useful tool for implementing QA steps that are commonly found in QA protocols. Upcoming enhancements to EmisView and the EMF will provide an even easier to use, more integrated approach to the quality assurance of emissions modeling data.

REFERENCES

Eyth, A., and M. Houyoux. "EmisView: New Software for Visualizing and Quality Assuring Emission Modeling Data". In *Proceedings, Transforming Emission Inventories – Meeting Future Challenges Today*; U.S. Environmental Protection Agency: Las Vegas, NV, 2005.

Eyth, A. EmisView User's Guide. University of North Carolina at Chapel Hill, 2005.

Houyoux, M., M. Strum, R. Mason, and A. Eyth. "Data Management using the Emissions Modeling Framework". 15th Annual Emission Inventory Conference, New Orleans, Louisiana, May 16-18, 2006.

Houyoux, M.R., Strum, M., Possiel, N., Benjey, W.G., Mason, R., Pouliot, G., Loughlin, D., Eyth, A.E., Seppanen, C. "EPA's New Emissions Modeling Framework", 14th Annual Emission Inventory Conference, Las Vegas, Nevada, April 11-14, 2005.

ACKNOWLEDGMENTS

The work described herein was funded under US EPA contract number 68D-02-066. Software development support was provided by Mr. Rajasooriyar Partheepan and Dr. Qun He of the Carolina Environmental Program. Members of the EIAG emissions modeling team including Madeleine Strum, Bill Benjey, and Rich Mason participated significantly in the design of the software described herein.

KEYWORDS

Visualization
Emission Inventory
Quality Assurance
EmisView
Emissions Model
SMOKE
CONCEPT