

## EMF QA Training Notes

### 3/24/2008

1.	EMF Extended SQL Syntax.....	1
2.	Adding QA Steps from Templates.....	3
3.	Special QA Programs.....	5
4.	Analysis Engine .....	6
4.1	Sorting Data .....	6
4.2	Showing Largest and Smallest N rows .....	7
4.2	Filtering Data .....	9
4.3	Showing and Hiding Data Columns .....	10
4.4	Formatting Data Columns.....	11
4.5	Plotting Data .....	12
4.6	Computing Statistics .....	16
4.7	Exporting Data .....	20
4.8	Analysis Configurations and Plot Templates.....	22
4.	Possible Enhancements .....	24

## 1. **EMF Extended SQL Syntax**

The EMF supports an extended SQL syntax that is used to make the SQL for the QA steps generic so the same queries can be applied to multiple datasets. The SQL used in the EMF queries has a slightly extended form that allows the queries to be made generic across datasets, and to take into account the versioned data that is stored in the EMF. The extensions to standard PostgreSQL are discussed below.

### 1. **\$TABLE [table\_num] a**

The \$TABLE tag is used to refer to a data table for the dataset to which the QA step is attached. The *table\_num* should be replaced with the table number that contains the data of interest in the other dataset, and typically has a value of 1. The 'a' that follows the ']' is a required one-character alias which can be used later in the query to reference columns from the specified dataset. An error is given if the alias does not exist or if it is more than one character. Note that if any of the arguments within the brackets are invalid (e.g., a dataset does not have the specified table number), a meaningful error message is presented to the user.

#### *Example 1: Summarize by Pollutant*

```
select POLL, sum(ann_emis) as ann_emis from $TABLE[1] e group by POLL
order by POLL
```

#### *Example 2: Summarize by U.S. State and Pollutant*

```
select fips.state_name, fips.state_abbr, fips.fipsst, e.POLL,
sum(ann_emis) as ann_emis from $TABLE[1] e inner join reference.fips on
fips.state_county_fips = e.FIPS where fips.country_num = ''0'' group by
fips.state_name, fips.state_abbr, fips.fipsst, POLL order by
fips.state_name, POLL'
```

## 2. **\$DATASET\_TABLE** [ “*dataset name*”, *table\_num*] a

The \$DATASET\_TABLE tag is used to refer to a table in a dataset other than the dataset to which the QA step is attached. The “*dataset name*” should be replaced with the name of the other dataset that contains the data of interest. The *table\_num* should be replaced with the table number that contains the data of interest in the other dataset. Quotes should be placed around the actual dataset name of interest, but not around the table number. For this tag, the default version of the other dataset will be used to generate the results. Note that if any of the arguments within the brackets are invalid (e.g., a dataset does not exist or it does not have the specified table number), a meaningful error message is presented.

## 3. **\$DATASET\_TABLE\_VERSION** [ “*dataset name*”, *table\_num*, *version\_num*] a

The \$DATASET\_TABLE\_VERSION tag is used to refer to a specific version of table in a dataset other than the dataset to which the QA step is attached. The “*dataset name*” should be replaced with the name of the other dataset that contains the data of interest. The *table\_num* should be replaced with the table number that contains the data of interest in the other dataset. The *version\_num* should be replaced by the number of the dataset version you wish to access. Quotes should be placed around the actual dataset name of interest, but not around the table.. Note that if any of the arguments within the brackets are invalid (e.g., a dataset does not exist or it does not have the specified table or version number), a meaningful error message is presented.

## 4. **\$DATASET\_QASTEP** [ “*dataset name*”, “*QA step name*”] a

The \$DATASET\_QASTEP tag is used to refer to a QA step in the current dataset or in a dataset other than the dataset to which the QA step is attached. To refer to a QA step in the current dataset, enter “CURRENT\_DATASET” in place of the “*dataset name*”. Otherwise, you may replace “*dataset name*” with the name of a specific dataset. The “*QA step name*” should be replaced by the name of the QA step that contains the data to be used in the query. Quotes should be placed around the actual *dataset name*, and also around the *QA step name*. Note that if any of the arguments within the brackets are invalid (e.g., the specified QA step name does not exist for the specified dataset), an exception is thrown and an informative status message is presented. For this tag, the QA step with the specified name associated with the default version of the dataset will be used.

## 5. **\$DATASET\_QASTEP\_VERSION** [“*dataset name*”, “*QA step name*”, *version\_num*] a

The \$DATASET\_QASTEP\_VERSION tag is used to refer to a QA step in the current dataset or in a dataset other than the dataset to which the QA step is attached. To refer to

a QA step in the current dataset, enter “CURRENT\_DATASET” in place of the “*dataset name*”. Otherwise, you may replace “*dataset name*” with the name of a specific dataset. The “*QA step name*” should be replaced by the name of the QA step that contains the data to be used in the query, and the *version\_num* should be replaced the version of the dataset with which the specified QA step name is associated. Quotes should be placed around the actual *dataset name*, and also around the *QA step name*, but not around the *version\_num*. Note that if any of the arguments within the brackets are invalid (e.g., the specified QA step name does not exist for the specified version of the dataset), an exception is thrown and an informative status message is presented.

Note that the examples of using the extended SQL syntax are relatively simple when only a single table is involved, but once joins with other tables are included to fill in data from those table, the level of complexity grows quickly and care must be exercised to use the correct form of the join.

*Example 3: Query template to compare emissions inventory summary values to a report*

```
select dis.state_name, dis.poll, a.ann_emis, b.variable, b.value,
  (coalesce(cast(b.value as double precision),0.0) -
  coalesce(a.ann_emis,0.0))
  as emis_diff,
  abs((coalesce(cast(b.value as double precision),0.0)) -
  coalesce(a.ann_emis,0.0)) as abs_diff,
  case
    when (coalesce(a.ann_emis,0.0)) <> 0.0 then
      (abs((coalesce(cast(b.value as double precision),0.0) -
  coalesce(a.ann_emis,0.0))) * 100)/(coalesce(a.ann_emis,0.0))
    else 0
  end
as pctdiff
  from
    (select distinct state_name, poll
      from $DATASET_QASTEP["CURRENT_DATASET", "Summarize by US State
and Pollutant"]
    union
    select distinct state,variable
      from $DATASET_TABLE["alm_annual_2002ac_emf.csv", 1] ) dis
  left outer join $DATASET_QASTEP["CURRENT_DATASET", "Summarize by US
State and Pollutant"] a
    on a.state_name = dis.state_name
   and a.poll = dis.poll
  left outer join $DATASET_TABLE["alm_annual_2002ac_emf.txt", 1] b
    on b.state = dis.state_name
   and b.variable = dis.poll
  order by dis.state_name, a.poll
```

## 2. Adding QA Steps from Templates

While it is possible to design a QA step from scratch every time, QA steps can be easily added to most ORL datasets and datasets of some other types through QA Step Templates that have been added to the EMF to support frequently needed queries. This is the best way to start using QA

steps. The templates are stored along with the dataset types. It is possible to use templates in the EMF because of the extended SQL syntax that allows the same query to be used for multiple datasets of the same type. A QA program must be specified for any QA step. Some QA steps use QA programs (e.g. SQL) that can run inside the EMF, while others (e.g., python scripts) currently run outside the EMF. To add a step from a template to a dataset, run it and view the results, do the following:

1. Choose **Datasets** from the **Manage** menu
2. Select the dataset of interest and click **Edit Properties**.
3. Go to the **QA** tab.
4. Click **Add from Template** to show a dialog with the templates available to the selected dataset.
5. Click on any optional templates for the QA steps you wish to add. Hint: Hold down the Ctrl key to select multiple non-adjacent steps.
6. Check the select checkbox and click **Edit** to open the editor for that QA Step.
7. To see the SQL that will be executed, click the **Set** button next to the Arguments field. Do not change the SQL unless you understand SQL and how your changes will impact the results the query.
8. Click **Run** to run the QA step. Watch your status window for updates on the status.
9. Click **Refresh** to refresh the run status.
10. After the step has completed, click **View Results** to see the QA step data as a table in an Analysis Engine window.
11. From the analysis engine window, you can **sort** and **filter** the data, and you can also **create plots** using the R software package and **compute statistics**. The plots will be created as PDF files and displayed using your default PDF reader.

Some examples of extended SQL templates that are currently available for many ORL types are:

- Summarize by Pollutant
- Summarize by SCC and Pollutant
- Summarize by County and Pollutant
- Summarize by US State and Pollutant
- Summarize by State and Smoke Pollutant Name
- Summarize by Pollutant with Description
- Summarize by SCC and Pollutant with Descriptions
- Summarize by U.S. State and Pollutant with Descriptions
- Summarize by U.S. County and Pollutant with Descriptions
- Summarize by U.S. State, SCC and Pollutant with Descriptions
- Summarize by Mact Code, U.S. State and Pollutant with Descriptions
- Summarize Mercury by U.S. State and MACT, SIC, SCC Codes with Descriptions
- List by Data Source Codes and U.S. State with Description

- List by Data Source Codes, U.S. State and Pollutant with Description
- Get SCC List

If an existing template almost does what you need, but not exactly, you may customize the query within the QA step by clicking the **Set** button next to the Arguments field. Once you have run the query to confirm that it works as you expect, if you believe you may want to use the query again, you should consider adding the new query as a template to the dataset type. Once the template is added, then you and other users can make use of it for other datasets.

### 3. **Special QA Programs**

Aside from the extended SQL queries, there are several special QA Programs that perform special functions within the EMF. These programs can be used by adding QA steps from the QA tab, but instead of clicking “Add from Template”, click “**Add Custom**”. From the dialog that appears, enter a name for the dataset and select one of the following special programs as the QA Program:

1. **Average Day to Annual State Summary:** Converts average day emissions and to monthly total values and then adds the monthly totals together for each selected dataset into a summary by state and pollutant.
2. **Average day to Annual Inventory:** Converts average day emissions and to monthly total values and then adds the monthly totals together for each selected dataset to create data that could be used as an annual inventory (i.e., by FIPS, SCC, and pollutant).
3. **Fire Data Summary (Day-specific):** Totals values from selected day-specific fire inventories at several levels of details (i.e., State, State+SCC, FIPS, and FIPS+SCC).
4. **Multi-inventory sum:** Totals annual and average day emissions from selected ORL inventories at several levels of details (i.e., State, State+SCC, FIPS, and FIPS+SCC).
5. **Multi-inventory column report:** Shows values from selected ORL inventories at several levels of details (i.e., State, State+SCC, FIPS, and FIPS+SCC) with data from each inventory as a column in the report. Currently, only annual emissions are shown, but an option to show average day emissions could be added.
6. **Multi-inventory difference report:** Creates sums of two groups (i.e., base and compare) of ORL inventories and reports statistics on the difference between the two groups, including difference, absolute difference, and percent difference of both annual and average day emissions from the two groups.
7. **SQL:** Indicates that you are specifying a query using the extended EMF SQL syntax described above.
8. **All other programs:** All other QA programs are currently executed external to the EMF, but their status can still be tracked manually by typing into the fields of the QA step.

Once you have added the QA step to the dataset, edit the QA step. When you click the **Set** button next to the **Arguments** field for the first six programs listed above, a special dialog will appear that allows you to specify the datasets and other parameters to use when you run the QA

step. The dialogs for many of the special QA programs are similar. They will allow you to select one or more inventories to use for the query. If you would like to add multiple inventories, you may hold down the Ctrl key while selecting multiple datasets, or you may click Add multiple times. If there are a lot of inventories of your desired type, you can enter a filter in the Dataset name filter field to show only datasets with the specified text in their name.











If you specify an inventory table for the QA program to use, the value in the inventory table will be identified for each pollutant in the inventory, and the specified factor will be applied to the emissions before they are entered into the result table. You can also select the level of detail for the results from the available options (e.g., State, State+SCC, FIPS, and FIPS+SCC).

Once you have run the program, you can click **View Results** to see the result of the query in the Analysis Engine table. Based on your analysis, you can then set the QA status for the QA step and enter any comments you have on what you saw.

#### 4. Analysis Engine


If you view your results from the analysis engine, you can perform functions on your data using the buttons on the toolbar shown in Table 1.

**Table 1. Icons on the Toolbar\***

	Sort		Format Columns
	Show Largest N Rows		Plot
	Show Smallest N Rows		Statistics
	Filter		Configuration
	Show/Hide Columns		Reset

##### 4.1 Sorting Data

There are two ways to sort the data that are loaded in the table application. The first way is to click on the column header for the column of interest. You should see the rows change so that they are sorted in descending order. If you click the column header again, the order of the sort will reverse and the rows will be sorted in ascending order. You can sort based on the values in any column by clicking on the column header. The above example sorted data based on the values in a single column. The second sorting method is based on values in multiple columns:

1. Click on the Sort icon  in the toolbar. The **Sort Columns** dialog should appear.
2. In this dialog, set the drop-down menu in the **Sort By** panel to **Property**. Then click the **Add** button and select a column name from the menu in the **Then Sort By** menu and

uncheck the **Ascending?** checkbox, then click **OK**. You should see that the data are now sorted initially by the first column in ascending order, and the rows that have the same value for the first column are then sorted in descending order.

3. You may add additional columns to the sort by if you so choose to sort by as many columns as you like by clicking the Add button again.

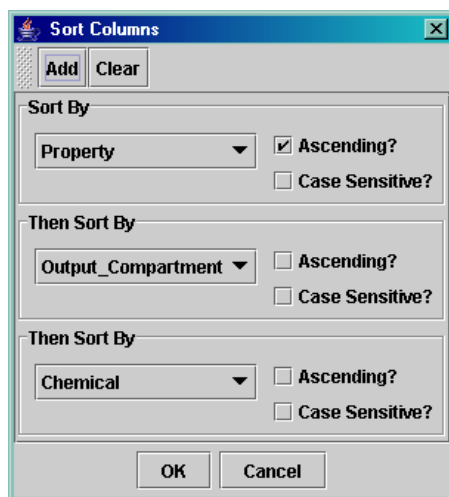



Figure 3. Sort Columns Dialog

## 4.2 Showing Largest and Smallest N rows

Another useful feature of the table application is to show the largest *N* or smallest *N* rows. When performing data analysis, it is useful to examine the rows with the top 10 values. To see this feature with your data set, do the following:

1. If you place the cursor in a column of the data panel and right-click, you will see a pop-up menu that includes the items **Show largest N rows** and **Show smallest N rows**. This will bring up the corresponding dialog with a column already selected. Next, select the number of rows you wish to see. You can do this either by specifying an absolute **Number of rows**, or you may specify a **Percentage of rows**. In this case, enter **10** for the **Number of rows** to show. Now the table in the Table Application window should contain only the 10 rows with the largest values for your selected column.
2. Alternatively, if you click on the Show Largest *N* Rows icon  from the toolbar. The **Show largest rows** dialog (Figure 4) will appear. (Note: when you bring up the dialog, the Data Column will not yet be selected, which is unlike the picture in Figure 4.)

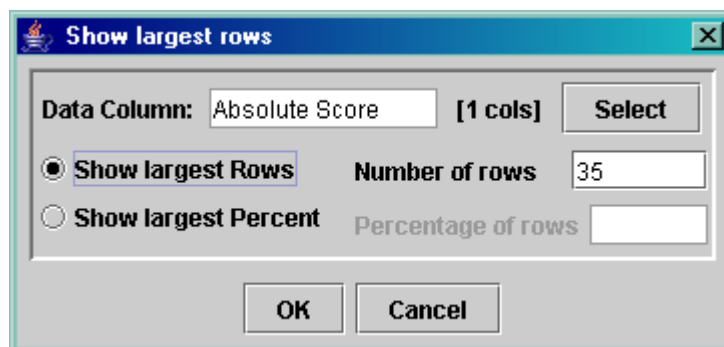


Figure 4. Show Largest Rows Dialog

3. In this dialog, select the data column of interest by clicking on the **Select** button. The **Include/Exclude Columns** dialog will appear (Figure 5). Note that the selection mechanism used in this dialog, explained next, is used repeatedly in the table application.

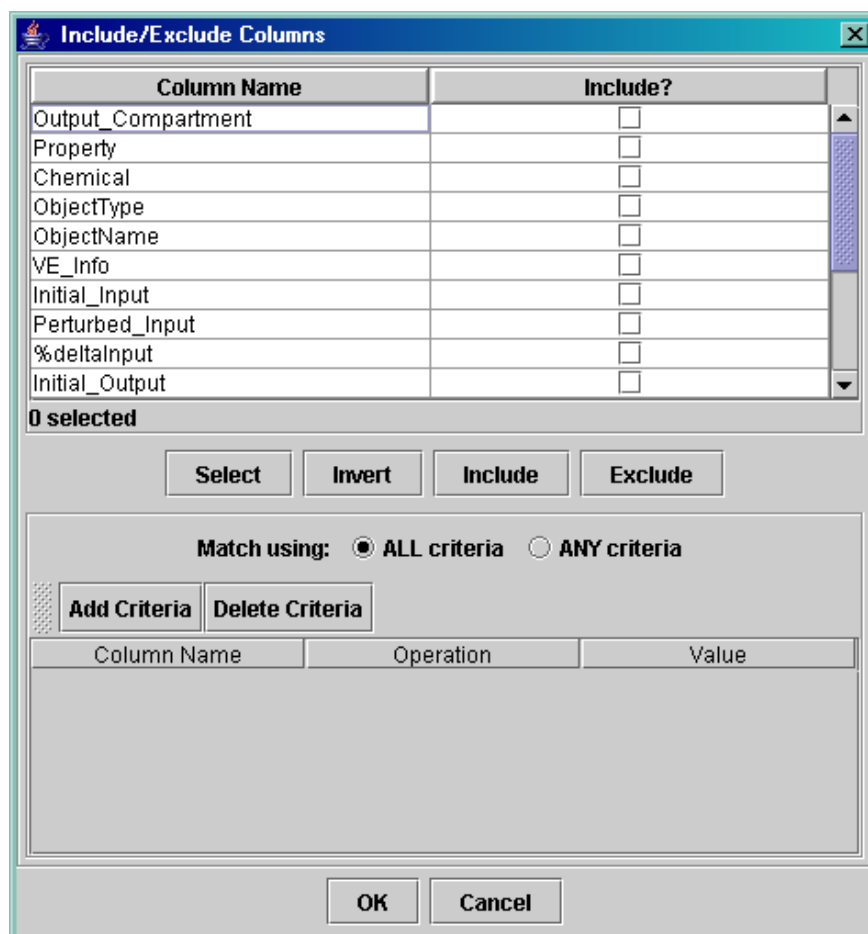





Figure 5. Include/Exclude Columns Dialog



4. In the **Include/Exclude Columns** dialog, scroll down to the column name of interest and click the corresponding checkbox in the **Include?** column (Note: the bottom-most column in this dialog corresponds to the right-most column in the data panel.) Next click **OK**.
5. Next, in the **Show largest rows** dialog, select the number of rows you wish to see. You can do this either by specifying an absolute **Number of rows**, or you may specify a **Percentage of rows**. In this case, enter **10** for the **Number of rows** to show. Now the table in the Table Application window should contain only the 10 rows with the largest values for your selected column.
6. To show all the rows of data again, click the Reset icon  in the toolbar.
7. To see the rows with the smallest values, click the Show Smallest N Rows icon  in the toolbar and follow a similar process.

## 4.2 Filtering Data

The table application allows you to use filter criteria to specify which rows to show. Here are some illustrations of how to specify filter criteria:

1. Click on the Filter icon  in the toolbar. The **Filter Rows** dialog will appear.
2. In this dialog, you may specify multiple criteria against which the rows will be evaluated. To specify the first criterion, click the **Add Criteria** button; a row will be added to the table of criteria.
3. Click on the new row under the **Column Name** label to make a menu of all the column names appear. Select a column of interest. Next, click under **Operation** to make a menu of operations appear. Choose **contains**. Then, under **Value**, type in a part of a value that you are interested in seeing. Note: It is important to hit Enter after you finish typing in this value.
4. Now that you have specified a criterion, Click **OK** on the **Filter Rows** dialog. You should now see only the data rows that contain the specified value.
5. To specify a second criterion, click the Filter icon again (the **Filter Rows** dialog should appear with the same settings you had specified previously), and then click **Add Criteria** to add a new criterion. A second row should be added to the table for you to fill in. After you click **OK**, in the **Filter Rows** dialog you should see only the rows that meet both of the specified criteria.

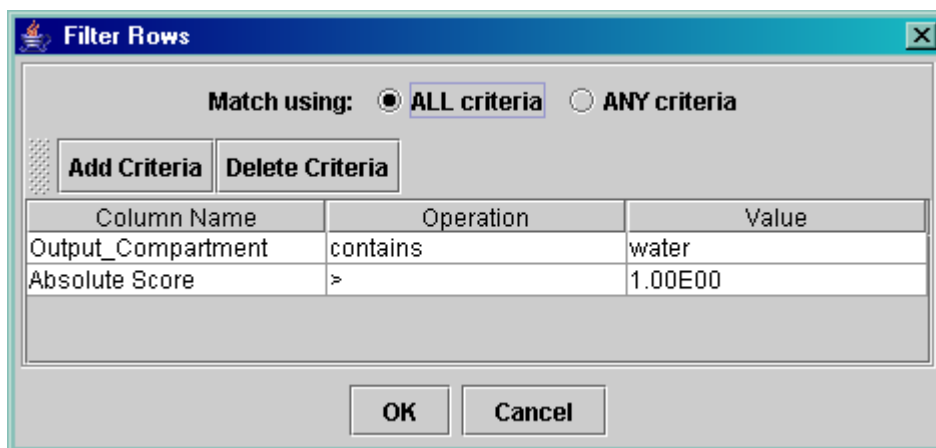



Figure 6. Filter Rows Dialog

- When specifying criteria, you can state that rows should be shown if **ALL criteria** are satisfied, or if **ANY criteria** are satisfied. In this case, if you change the **Match using** selection to **ANY criteria** and click **OK**, then you will see that the number of rows shown increases substantially, and that each row shown satisfies at least one of the criteria.


### 4.3 Showing and Hiding Data Columns

The table application allows you to hide some of the columns in your data set. To see how this works, perform the following steps:

- Click on the Show/Hide Columns icon  in the toolbar. The **Show/Hide Columns** dialog will appear. This dialog looks very similar to the **Include/Exclude Columns** dialog in Figure 5.
- By default, all the columns are shown, so all the checkboxes in the **Show?** column are checked. In this case, we will hide some of the columns we are not interested in by unchecking some of the checkboxes and then clicking **OK**. You will see that the columns are no longer shown in the table.
- Click on the Show/Hide Columns icon again (the **Show/Hide Columns** dialog should appear in the same state as when you closed it). This time, scroll down to the bottom of the table in the **Show/Hide Columns** dialog and select some rows by clicking on the row, holding your mouse button down, and dragging your mouse down until all three column names are highlighted. Then click the **Hide** button. This will cause the checkboxes for the three columns to become unchecked. Now click **OK** and you will see that the three columns are no longer visible in the table.
- To make the hidden columns reappear, click on the Show/Hide Columns icon. In the the **Show/Hide Columns** dialog select all of the column names and then click the **Show** button. After you click **OK** you will see the columns again.

## 4.4 Formatting Data Columns

The table application allows you to specify the format for the columns in your data set. Here are some examples:

1. Click on the Format Columns icon  in the toolbar. The **Format Columns** dialog will appear.
2. To format the last six columns of the data set, scroll down to the bottom of the table in the **Format Columns** dialog. First highlight the columns by clicking on one column name, holding your mouse button down, and dragging your mouse down until all the column names are highlighted. Next click the **Format** button.
3. To make these columns use standard notation instead of exponential notation, click the **Standard Notation** option on the **Numeric Format Options** panel. Click **OK** and note the change in the format for the selected columns.
4. Click on the Format Columns icon again (the **Format Columns** dialog should appear with the same columns still selected for formatting). Now try reducing the number of **Significant Digits** to **2**. Then click **OK** and note the impact on the selected columns.
5. Click on the Format Columns icon again. This time, uncheck all the checkboxes except a couple of columns. On the right side of the dialog, click on the color square next to **Background Color**. In the **Choose Color** dialog that appears, select a light yellow, then click **OK**. Next, set the font **Style** to **Bold**, the **Text Color** to blue, and the **Horizontal Alignment** to **Center**, then click **OK**. You should see that the data in those columns look similar to those shown in Figure 8. Note that additional special options are available for columns that are dates.

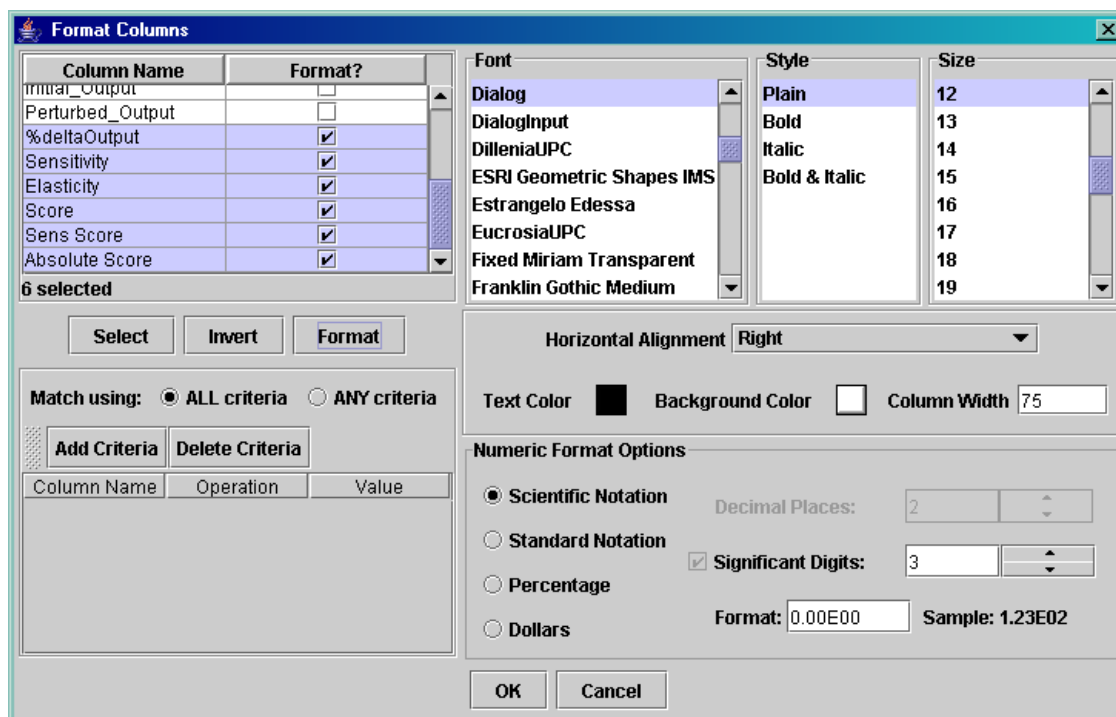



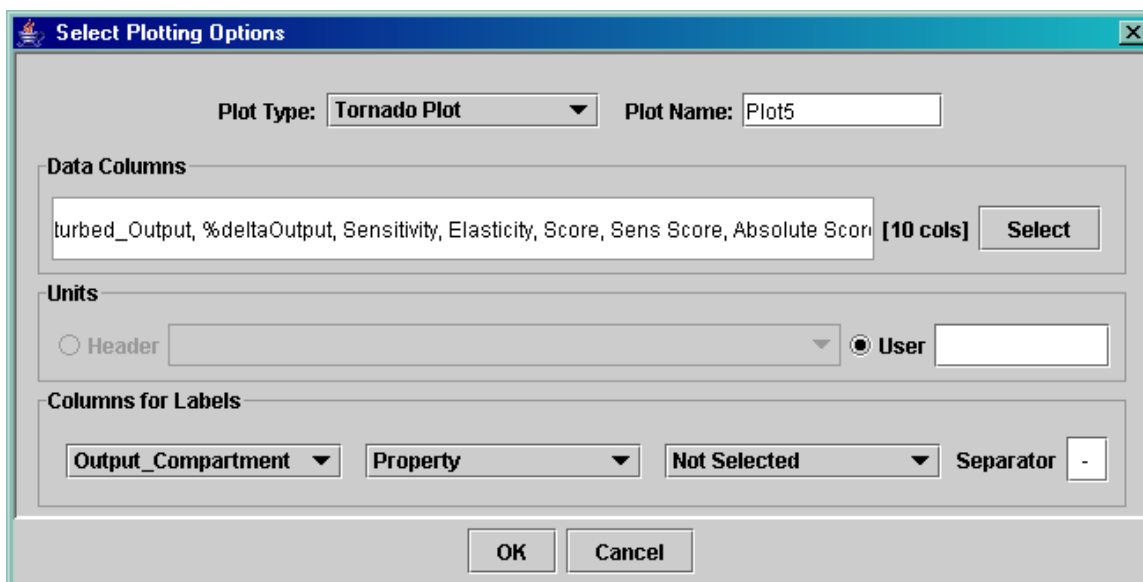
Figure 7. Format Columns Dialog

Sensitivity	Elasticity	Score	Sens Score	Absolute S...
-8.80	-8.80	-8.80	-8.80	8.80
-7.30	-7.30	-7.30	-7.30	7.30
-7.20	-7.20	-7.20	-7.20	7.20
-9.60	-9.60	-2.90	2.90	2.90
-9.30	-9.30	-2.80	-2.80	2.80
-8.10	-8.10	-2.40	2.40	2.40
-8.10	-8.10	-2.40	-2.40	2.40
-8.00	-8.00	-2.40	2.40	2.40
-7.90	-7.90	-2.40	2.40	2.40
-7.90	-7.90	-2.40	-2.40	2.40

Figure 8. Example of Formatted Columns

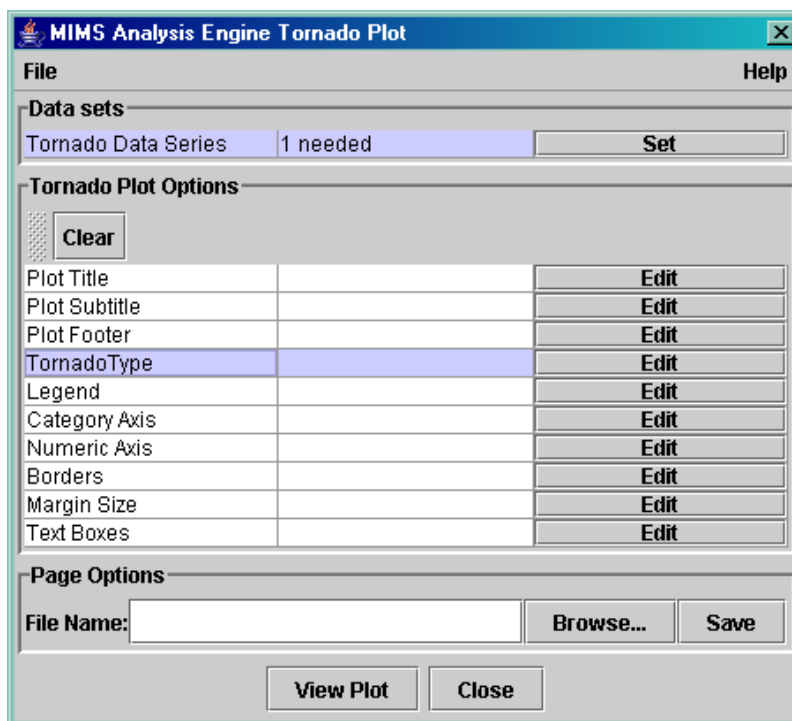
## 4.5 Plotting Data

To plot data, first click on the Plot icon  in the toolbar. The **Select Plotting Options** dialog should appear. An example is shown in Figure 9.



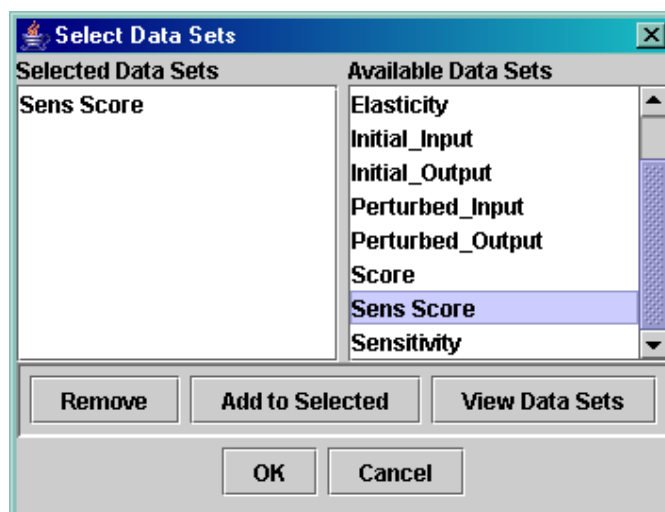
**Figure 9. Select Plotting Options Dialog**

1. In this dialog, specify your **Plot Type**. The numeric data columns are specified by default in the **Data Columns** panel. You may choose a subset of the columns by clicking on the **Select** button in this panel and then choosing the columns of interest from the Column Selection dialog. In this case, we will just use the default selection.
2. Next, set the **Columns for Labels** entry (e.g. to state name if your data are state specific).
3. After you click **OK** in the **Select Plotting Options** dialog, the **MIMS Analysis Engine Plot** dialog will appear for the selected plot (such as the one shown in Figure 9). To specify the data set you wish to plot, click on the **Set** button next to the **Data Series** entry at the top. At this point the **Select Data Sets** dialog will appear. Note: The list of **Available Data Sets** in this dialog corresponds to the columns listed in the Data Columns panel of the **Select Plotting Options** dialog.



**Figure 10. MIMS Analysis Engine Tornado Plot Dialog**

4. In the **Select Data Sets** dialog, select **Sens Score** as the data set to plot by double-clicking on it in the list of **Available Data Sets**. This will cause it to be added to the **Selected Data Sets** list, as shown in Figure 11. Note that since the tornado plot shown in this example can plot only one data set, if you select another data set from the list of available data sets that new selection will replace the item in the list of selected data sets. For plot types that allow multiple data sets, the selected data sets are accumulated in the list on the left. Now click **OK** to return to the **Plot** dialog for the specified plot.



**Figure 11. Select Data Sets Dialog**

5. To see the default plot, click the **View Plot** button. The plot should appear (if R is installed properly), but it may not yet look precisely the way you want it to.
6. To customize the options related to the tornado bars, click on the **Edit** button next to **TornadoType** (or BarType for a Bar Plot). An **Edit Tornado Type Properties** dialog should appear. On this dialog, set the **Space** on the **Bars** panel to 2.0 and the **Size** on the **Labels** panel to 0.3. Click **OK** to close the dialog, then click **View Plot** to see the revised plot.
7. To add a title to the plot, click on the **Edit** button next to **Plot Title**. Then type a title into the text area labeled **Text:** (e.g., Sensitivity Score for the Mass of Methyl Mercury). Note that there are many adjustments that can be made to the look of the title, including the style, color, font size, and position. After you click **OK**, click **View Plot** again to see the plot with the title. It should look similar to the plot in Figure 12.

Sensitivity Score for the Mass of Methyl Mercury

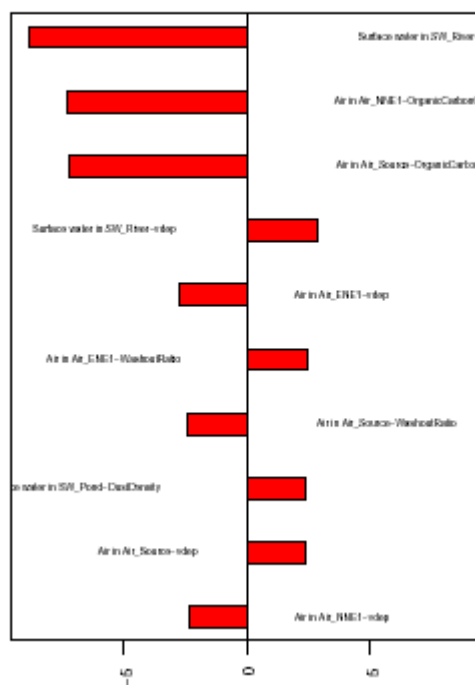


Figure 12. An Example Tornado Plot for Sensitivity Analysis


8. There are many other options that can be configured for the plot. Feel free to explore these. For more information, see the detailed documentation in the MIMS Analysis Engine user's guide, which is available on-line from the **Help** menu in the Table Application window. If you wish to save the plot to a file, click the **Browse** button to

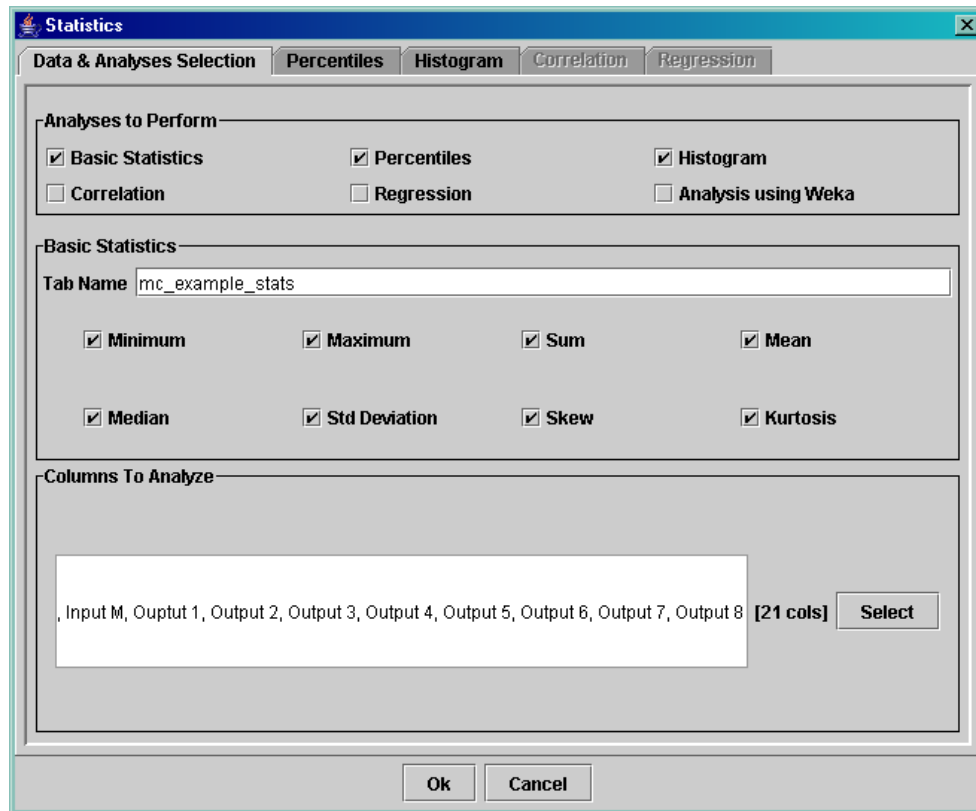
bring up a file browser, browse to the appropriate directory, and enter a file name. The **Files of Type** menu shows you the list of available file types. These include .jpg, .pdf, .png, and .ps. Once you have specified the directory and file name, click the **Save** button in the file browser. Note that if you wish to save the file as .jpg or .png, you should first edit the **Margin Size** analysis option and change the information for **Page Size Width** and **Height** to numbers that make sense in pixels (e.g., 400 x 400). The default units for these are inches, which work fine when you output .pdf and .ps files.

9. When you are finished editing the plot, click on the **Close** button on the **Plot** dialog. There are many types of plots available from the Analysis Engine.

## 4.6 Computing Statistics

The table application can generate a number of different types of statistics that are useful for sensitivity and uncertainty analysis. Follow the steps below to generate some of these.

1. The Monte Carlo example data set is a useful example to illustrate the available statistics. In the Table Application window, click on the tab named **mc\_example** to bring up this data set. Next, click on the Statistics icon  in the toolbar. The **Statistics** dialog will appear (Figure 13).





**Figure 13. Statistics Dialog**



2. In the **Statistics** dialog, the categories of statistics that can be generated are listed in the top panel, named **Analyses to Perform**. The available types of statistics include basic statistics, percentiles, histogram, correlation, regression, and analysis using Weka. The configuration for the basic statistics is done on the **Data & Analyses Selection** tab of the **Statistics** dialog. Configuration options specific to percentiles, histogram, correlation, and regression are done on the corresponding tabs in the dialog. These tabs are activated if the corresponding checkbox is selected in the **Analyses to Perform** panel. If **Analysis with Weka** is checked, the data for the columns selected in the **Columns To Analyze** panel of the **Data & Analyses Selection** tab is passed directly to the Weka Explorer. For more information on Weka, see <http://www.cs.waikato.ac.nz/~ml/weka/>.
3. We will start by computing some basic statistics. First, make sure the **Basic Statistics** box is checked. By default, all of the statistics listed in the **Basic Statistics** panel will be computed. If you do not wish to compute some of these statistics, deselect the corresponding checkboxes in the **Basic Statistics** panel. The statistics will be computed for all columns listed in the **Columns To Analyze** panel. By default, all the numeric data columns are selected. We will use the default setting for this example. Next, click **OK**. A new tab ending with **\_stats** should be added to the Table Application window. Go to that tab. You will find that the tab contains the basic statistics for all the inputs and outputs to the Monte Carlo analysis with the statistics (i.e., minimum, maximum, etc.) as the rows, and the selected data columns as the columns.
4. Go back to the main tab and click the Statistics icon again. This time, deselect Basic Statistics, and select **Percentiles** and **Histogram**. Next, click the **Select** button in the **Columns To Analyze** panel. On the **Include/Exclude Columns** dialog, highlight all the columns that start with Input, then click **Exclude**. Click **OK** to return to the **Statistics** dialog. You should see that the **Columns To Analyze** text box has been updated.
5. Next, go to the **Percentiles** tab, which is shown in Figure 14. The **Frequent Options** panel contains some frequently used options. The default option is **Standard**, which computes the following percentiles: 0.01, 0.05, 0.10, 0.50, 0.90, 0.95, and 0.99. (Note that the percentiles are specified to be between 0 and 1, instead of 0 and 100.) If the standard percentiles are the percentiles you want to compute, you do not need to do anything on this tab. However, if you wish you may select one of the other frequently used options of quartiles, quintiles, or deciles. The **Add Percentiles** panel allows you to add a set of percentiles with a specified minimum, maximum, and step size. For example, if you want to see every 0.01 percentile from 0.9 to 1.0, enter 0.9 as the minimum, 1 as the maximum and 0.01 as the step size. The specified percentiles will be computed for the data in each column selected in the **Columns To Analyze** panel. To modify the items in the Percentiles table directly, perform the following steps:

#	Percentile
1	0.0000
2	0.5000
3	0.9000
4	0.9500
5	0.9900
6	1.0000

**Figure 14. Percentiles Tab of the Statistics Dialog**

- a. Double-click on the 0.0100 under **Percentile** in the table of percentiles. This should allow you to edit the value. Change the value to 0.
  - b. In the table of percentiles, click on the 2 in the row containing 0.0500 and drag your mouse down to also select the row containing 0.100. Then click on the Delete rows icon  in the toolbar right above the table of percentiles. The two rows should be removed from the table of percentiles to compute.
  - c. Click on the last row of the table that contains the value 0.9900 to select it. Then click on the Insert a row below the current row icon  in the toolbar. A new row should be added with the default value of 1.0000. When you are finished, the percentiles tab should look like the one in Figure 14.
6. In the **Histogram** tab (Figure 15), you can specify the bins to use for the histogram and the format of the bin labels. By default the table application will compute the minimum and maximum of the data in the selected data columns, divide the range into 10 equal-sized bins, and compute the frequency count for each bin. If the default does not suit your needs, you can customize the analysis. For example, you can specify the **Type** as Frequency (a count), Percentage (0 to 100), or Probability (0 to 1). In the **Binning** panel, you may change the number of bins, create bins that differ by a factor of 10, or customize the bins. Histogram information will be computed for the data in each column that is specified in the **Columns To Analyze** panel of the **Data & Analyses Selection** tab.

**Data & Analyses Selection** | **Percentiles** | **Histogram** | Correlation | Regression

Tab Name: Histogram 3

Type: ☒ Frequency ☐ Percentage ☐ Probability

Data Range: Minimum value: 1.42E03 Maximum value: 4.46E04

Binning: Bin Type ☒ Equally Spaced ☐ Custom ☐ Factor of 10  
 Lower bound: 0 Upper bound: 50000  
 Number of bins: 10 **Recompute**

Format Labels: Numeric Format Options  
☐ Scientific Notation Decimal Places: 0  
☒ Standard Notation ☒ Significant Digits: 4  
 Format: 0 Sample: 124 **Apply Format**

#	Break Points
1	0
2	5000
3	10000
4	15000
5	20000
6	25000
7	30000
8	35000
9	40000
10	45000
11	50000

Figure 15. Histogram Tab of the Statistics Dialog

The steps below show how to use some of the options on the **Histogram** tab:

- a. Set the **Lower bound** to 0 and the **Upper bound** to 50000, then click **Recompute**. The values in the **Break Points** area of this tab should be updated.
  - b. Click on **Standard Notation** in the **Numeric Format Options** panel and adjust the number of **Decimal Places** down to 0. Then click **Apply Format**. When you are done, the dialog should look like the one in Figure 15.
  - c. Click **OK** on the bottom of the **Statistics** dialog to start the process of computing the percentiles and the histogram.
7. After the histogram and percentiles have been computed, two new tabs will be added to the Table Application window: **Histogram 1** and **Percentiles 1**. Go to the **Histogram 1** tab. If needed, resize the **Bins** column so you can see the full bin labels by putting your mouse between **Bins** and the first data column and dragging it to enlarge the column. The default format for all data is scientific notation, but the frequency histogram information is easier to read as integers. Click the Format Columns icon on the toolbar to change

the formatting. Select the row named **Bins** and click the **Invert** button, then click **Format**. Choose the **Standard Notation** option on the **Numeric Format Options** panel, and reduce the decimal places to 0. Next, set the **Column Width** to 60, then click **OK**. The data in the tab should appear as shown in Figure 16. You may also wish to examine and experiment with the data in the **Percentiles 1** tab.

Row	Bins	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6	Output 7	Output 8
1	0 to 5000	0	0	0	490	490	0	490	490
2	5000 to 10000	0	0	2	0	0	0	0	0
3	10000 to 15000	0	479	289	0	0	0	0	0
4	15000 to 20000	153	11	198	0	0	0	0	0
5	20000 to 25000	337	0	1	0	0	0	0	0
6	25000 to 30000	0	0	0	0	0	0	0	0
7	30000 to 35000	0	0	0	0	0	221	0	0
8	35000 to 40000	0	0	0	0	0	135	0	0
9	40000 to 45000	0	0	0	0	0	134	0	0
10	45000 to 50000	0	0	0	0	0	0	0	0

**Figure 16. Output from a Histogram Analysis**

Additional types of statistical analyses are Correlation and Regression analyses, but they are beyond the scope of this document.

## 4.7 Exporting Data

The table application allows you to export the data that are loaded into it in the form in which they are currently being viewed. For example, if in one of the tabs you have applied a sort and a filter to the data, and you have hidden some of the columns, you can save the data in the form in which it appears in the tab. To export the data, choose **Export** from the **File** menu. The **Export Files** dialog will appear and should look similar to the one in Figure 17.

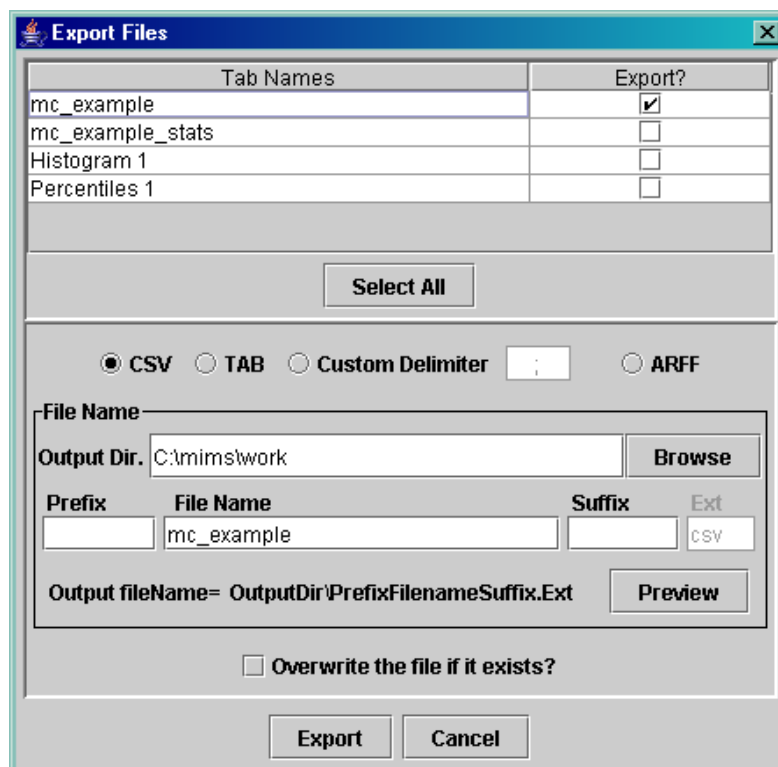


Figure 17. Export Files Dialog

To export any tabs of interest, perform the following steps:

1. Select the corresponding checkboxes in the **Export?** column
2. Specify the format as comma separated values (CSV), tab delimited, custom delimited, or ARFF (a Weka format). Note that in the October 2004 version of the table application, only ASCII formats are supported. Eventually, we may support RTF and HTML, which could also preserve the formatting of the data and include the plots.
3. Specify the file name(s) by setting the **Output Directory** and, if desired, setting a prefix and suffix to use for the file names. For example, if you have sorted the files, you may want to add a suffix of \_sort. If you wish to see the complete file name, click the **Preview** button.
4. If you wish to overwrite the file if it already exists, check the **Overwrite file if it exists?** checkbox.
5. When you are finished specifying all the settings, click the **Export** button.

## 4.8 Analysis Configurations and Plot Templates

The table application has a very useful feature called analysis configuration. When you are performing an analysis, the system tracks what you have done and then allows you to save the analysis to a file and reapply it at a later time. An analysis configuration for the `mc_example.csv` file has been provided as part of this tutorial. Follow the steps below to see how analysis configurations can be used.

1. Close the table application and then restart it (see Section 2.1 for instructions on how to do this). Choose the **Recent Files** item from the **File** menu. In the **Import Recent Files** dialog, double-click on the row with the file name `mc_example.csv`. The file should then be loaded into the table application.
2. Next, you will load a configuration that was prepared for the purposes of this tutorial. This configuration contains plots that would be useful to generate in the context of a sensitivity or uncertainty analysis. To load this file, choose the **Load Configuration** item from the **File** menu. From the file browser that appears, choose the file `mc_example.cfg` from the tutorial directory of the MIMS installation. The **Load Configuration Preview** dialog should appear (Figure 18).

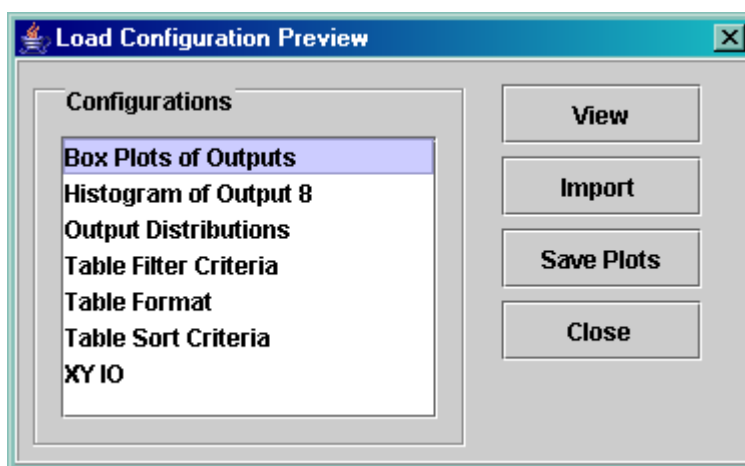



Figure 18. Load Configuration Preview Dialog

3. In the **Load Configuration Preview** dialog, click on **Box Plots of Outputs** to select it and then click the **View** button. A box plot of the model outputs should appear. From this plot you can see the basic ranges of the model outputs. Next, click on **Histogram of Output 8** and then click the **View** button. This will generate a histogram plot for the data values for Output 8. Note: this is just an example of a histogram plot – there is nothing in particular about Output 8 that caused us to generate this plot for it.

4. Next, select all the items in the **Configurations** list by clicking on the first item, then holding the shift key down and clicking on the final item. Next, click the **Import** button and then the **Close** button. (Note: If we had not looked at any of the plots first and instead had immediately clicked the **Import** button with no items selected, all of the configuration items would have been automatically imported.)
5. To see that the items in the configuration were imported, click on the Configuration icon  in the toolbar. The **Analysis Configuration** dialog should appear and should look like the one in Figure 19. This dialog shows the configuration that is currently applied to your data set, whereas the Load Configuration Preview dialog showed the items in the configuration file that were available for import.

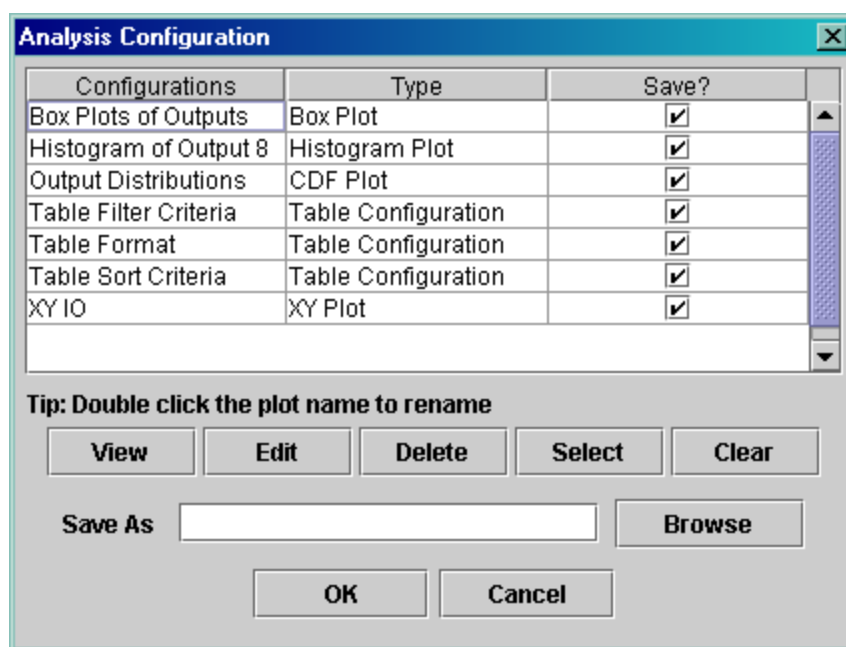


Figure 19. Analysis Configuration Dialog

6. You can view plots from the **Analysis Configuration** dialog as we did from the **Load Configuration Preview** dialog. For example, click on the **Type** column header to sort based on the contents of that column. Then you can select the four plots by clicking on the **Box Plot** and dragging your mouse down to the **XY plot**. Then click **View** and the four plots should be generated. If you are running on Windows, you can use the navigation keys in Acrobat to scroll back and forth through the plots. Please take a moment to examine the plots that were produced.

#### **4. Possible Enhancements**

Based on discussions, the following enhancements to the system are suggested:

1. Allow users to access the export and load configuration options that exist in the analysis engine (and perhaps the help system also).
2. Have a QA program that allows a user to select QA step results and builds a query for that. For example: difference two versions of a dataset, or difference QA reports for two versions of the same or different datasets.
3. Support an independent where clause to let users limit reports to particular rows of interest (although this can sometimes be done by generating the full report and then filtering with the analysis engine).
4. Use the multi-inventory sum to also sum the ptday files.
5. Add an option on the multi-column report to remove the POLL column and sum over the SMOKE name when using an inventory table.