EPA/600/R-18/021 January 2018 | www.epa.gov/ord

United States Environmental Protection Agency

Temporal-Spatial Ambient Concentrator Estimator (T-SpACE): Hierarchical Bayesian Model Software Used to Estimate Ambient Concentrations of NAAQS Air Pollutants in Support of Health Studies

Eric S. Hall

Office of Research and Development
National Exposure Research Laboratory

EPA/600/R-18/021
January 24, 2018

Temporal-Spatial Ambient Concentrator Estimator (T-SpACE): Hierarchical Bayesian Model Software Used to Estimate Ambient Concentrations of NAAQS Air Pollutants in Support of Health Studies

Eric S. Hall

Systems Exposure Division
National Exposure Research Laboratory
Office of Research and Development
US Environmental Protection Agency
Research Triangle Park, North Carolina 27711

Table of Contents

Appendix A - Table of Contents

Appendix B - Table of Contents

1.0 Introduction to Temporal-Spatial Ambient Concentrator Estimator (T-SpACE) Software

1.1 EPA's NAAQS and criteria pollutants

To fulfill its mission to protect human health and the environment, EPA has established National Ambient Air Quality Standards (NAAQS) on six selected air pollutants known as criteria pollutants: ozone (O_3); carbon monoxide (CO); lead (Pb); nitrogen dioxide (NO_2); sulfur dioxide (SO_2), and; particulate matter (PM). The states are primarily responsible for maintaining and improving air quality and complying with the NAAQS. Ozone (O_3) and particulate matter (PM) are two of the criteria pollutants whose levels are regulated by NAAQS. Extensive air pollution monitoring networks have been set-up across the US to understand the levels (concentrations) of these air pollutants across the US and over time. Once collected, this information is reported to EPA and is made available publicly after the data is reviewed and analyzed for quality and accuracy.

When assessing air quality, the most direct way is to utilize unbiased ground-level air pollution measurements from the existing surface monitoring network stations located across the US. However, a good portion of the US, especially rural areas, have very sparse or irregularly spaced monitoring stations, resulting in large areas of the country with no information available on air quality. Air quality models estimate the spatial and temporal gradients of air pollution based on emissions inventories and meteorological information. These models, while providing estimates over large regions at relatively low cost, have (statistical) bias and have greater error (variance) than air pollution monitoring networks. These methods include Models-3/Community Multiscale Air Quality (CMAQ) numerical model and the Comprehensive Air Quality Model with Extensions (CAMx). Various statistical methods have been investigated that utilize existing monitoring data and spatial modeling techniques to develop a concentration surface that (retrospectively) estimates the distribution of daily air pollution levels within a specified region of the US. Kriging is one such method. Another technique is Hierarchical Bayesian Modeling (HBM).

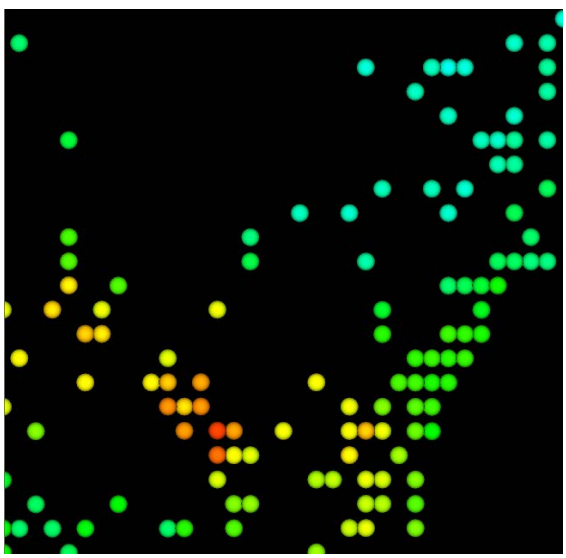


Figure 1. Plot of $PM_{2.5}$ Measurements on a Specific Monitoring Day from FRM Monitors in the NAMS/SLAMS Network Located in the Northeast US

The HBM approach combines air monitoring data from the 2,545 National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) with results from model estimates, such as from the CMAQ and CAMx air quality models. The approach develops daily estimated surfaces for both fine particulate matter ($PM_{2.5}$) and ozone (O_3) concentrations.

This report describes the HBM approach developed to model concentration surfaces for O_3 and $PM_{2.5}$, and the accompanying software that has been developed to produce the surface estimates.

1.2 Need for the software

The Temporal-Spatial Ambient Concentrator Estimator (T-SpACE) software, that was jointly developed by EPA and Battelle through EPA Contract EP-D-04-068 (Work Assignment #54), requires data preparation and model parameter input assignments from the model user to run a simulation to produce an estimated air pollution concentration surface. The T-SpACE software was designed to streamline model set-up and minimize the amount of work required to manually implement the data and model preparation steps, including allowing the user to prepare the input data for the model run, choose the model specifications, run the model, receive output files containing the estimated surface, and run a set of validation procedures to compare the T-SpACE output concentrations with a standard kriging method and model-based output concentrations such as those from CMAQ and CAMx. The T-SpACE software provides a single interface for running a simulation. The user is asked for particular choices for input data sets, model parameters, and names for output files and is guided through the simulation. **Note:** This model was used in the preparation of two peer-reviewed journal articles: Weber, S. A., Insaf, T. Z., **Hall, E. S.**, Talbot, T. O., Huff, A. K., 2016, "Assessing the Impact of Fine Particulate Matter ($PM_{2.5}$) Sources on Respiratory-Cardiovascular Chronic Diseases in the New York City Metropolitan Area using Hierarchical Bayesian Model Estimates", Environmental Research, Volume 151, November 2016, pp. 399-409, DOI: 10.1016/j.envres.2016.07.012 (**Weblink:** http://www.sciencedirect.com/science/article/pii/S0013935116302961), and, "Assessing the Impact of Fine Particulate Matter ($PM_{2.5}$) on Respiratory-Cardiovascular Chronic Diseases in the Baltimore Maryland Metropolitan Area using Hierarchical Bayesian Model Estimates" (***in preparation***). The United States Environmental Protection Agency and its Office of Research and Development did not perform, or provide funding, for data collection and analysis of the health outcome data described in the two above-mentioned journal articles.

1.2.1 Brief overview of the steps to develop the surfaces

In the software, there are five distinct steps to develop the surfaces and the final step prepares the visual rendering of the generated air pollution concentration surface.

- Step 1: Choose Time/Grid
- Step 2: Prepare Model Input Data
- Step 3: Model Specification
- Step 4: Launch Model
- Step 5: Launch Validation
- Step 6: Visualize Surface

Each estimated air pollution concentration surface is calculated for a single calendar year. In Step 1, the area covered by the air pollution concentration surface is defined by the positional grid coordinates utilized by the selected air quality model. Step 1 directs the user to choose the time frame over which to calculate the surfaces, and the positional grid coordinates from the selected air quality model (CMAQ or CAMx) to provide a boundary for the air pollution concentration surface estimates. Step 2 transforms the latitude/longitude based monitor data to the positional grid coordinates of the selected air quality model and develops an analysis data set to model the air pollution concentration estimates. Step 3 directs the user to specify the air quality model and the parameters of the simulation. Step 4 launches T-SpACE using a Markov Chain Monte Carlo (MCMC) simulation. This step produces the file that contains the estimated

air pollution concentration surfaces. Step 5 validates the model results against kriging and the estimates from one of the air quality models. Step 6 generates a graphical view of the surface.

1.2.2 Expected use of the resulting air quality surfaces

The resulting air quality concentration surfaces are useful to public health researchers, epidemiologists, air quality assessors, air quality scientists, and those with a basic understanding of statistical/mathematical modeling.

2.0 Design and nominal operational behavior of the T-SpACE software

The T-SpACE software was designed to easily guide the user through the many steps that were previously manually implemented utilizing SAS and R.

2.1 Design of the software

There were several design trade-offs that were made in the design of this system. The system was developed from an existing set of SAS programs, an MCMC model executable that required a particular text structure for the input information and had a specified structure for the output results, and monitor and model-based concentration data that had a specified structure. Throughout the software development process, decisions were made to allow the T-SpACE model to work within the established design constructs, taking into account EPA requirements, while developing software that was easy for the user.

2.1.1 Transforming monitor data and model-based data to simulation input data

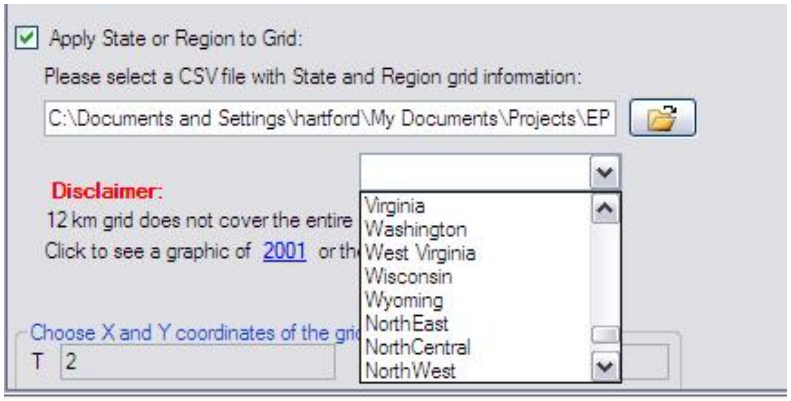
Early in the development of T-SpACE, the following design and operational constructs were implemented using previous model development work that had been completed:

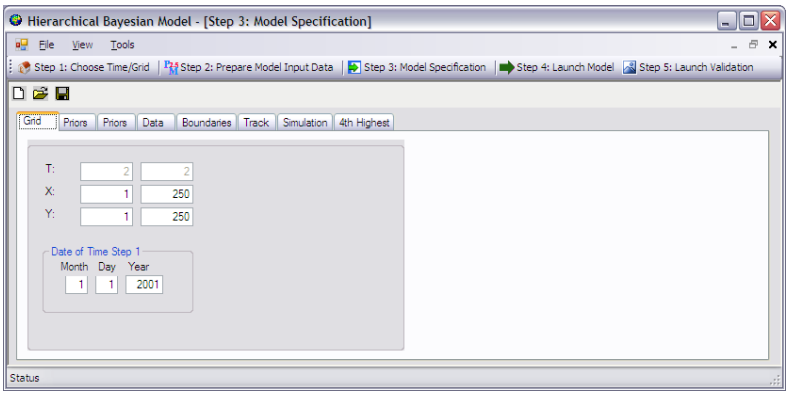
- Conversion of the Network Common Data Form (NetCDF) model-based concentration files to the required format for input into the MCMC model executable;
- Conversion of the AQS monitor data to the air quality model-based grid, and subsequently to the required format for input into the MCMC model executable;
- Adding the latitude and longitude to the MCMC simulation output file.

The decision was made to ensure that the data transformations used for these data types were implemented according to EPA standards such as utilization of the Lambert projection transformations and the quality assurance (QA) information from the AQS monitor files.

2.1.1.1 Air Quality Monitoring Network Monitor Data

For the initial input air quality monitor data, the user must supply a text file that is in the exact same format as EPA's AQS RD (raw data) monitor data.





2.1.1.2 Air Quality Model-Based Data

For the initial input air quality model data, the user must supply a file that is formatted like the CMAQ NetCDF files. A file formatted like the CMAQ NetCDF files is required for model execution. When the CAMx model is used in T-SpACE, its format is converted into the CMAQ NetCDF file format.

2.1.2 Choosing the parameters for the simulation

The software has been designed to allow the user to choose the parameters for the simulation. These choices are made in Steps 1 and 3 of the software. The choices the user can make are described below.

2.1.2.1 Choice of the region for simulation

The software provides users with the ability to identify the appropriate grid coordinates for a particular region or state.

2.1.2.2 T-SpACE parameter choices

There are several tabs in the software to guide the user through the choices for the model parameters. By sequentially following the tabs, the user is specifying priors of the model including the number of control points in the bias spline calculation, the distributional properties of the mean level of the Conditional Auto Regressive (CAR) process, the vector of coefficient for the bias, the precision of the measurement error in the monitor observations, the precision of the measurement error in the computer observations, the precision of the mean process, and the temporal autocorrelation parameter of the mean process. In addition, the user specifies the boundaries which define the neighborhood structure of the model, the random seeds for initializing the simulation, the frequency of sampling during the simulation, the number of initial model runs (burn-ins) for the simulation, and the total number of simulation loops that will be completed. Finally, to provide the user information about the simulation, the user can choose to track the chain of the MCMC simulation.

2.1.3 Results from the simulation

The T-SpACE software stores the results in a tabular (comma-delimited) format. One row exists in this file for each grid cell within the region and for each day within the time period of interest. The relevant variables in this file, in the order in which they appear (and are shown in the column headings), are as follows:

Day	XCoord	YCoord	Longitude	Latitude	PredAvg	PredStd	BiasAvg	BiasStd	CovarAvg	CovarStd	MonitorData
1/2/2001	41	65	-91.2294	35.24114	3.56199	1.34186	-0.90992	0.940508	-999	-999	-999
1/2/2001	41	66	-91.0884	35.34224	3.55762	1.34159	-0.95024	0.924785	-999	-999	-999

- **Day**: Date represented by the data given in this row, in MM/DD/YYYY format.

- **XCoord**: The cell's x-coordinate (east-west) within the grid (where the coordinate value increases by 1 with each grid cell as you move from west to east).
- **YCoord**: The cell's y-coordinate (north-south) within the grid (where the coordinate value increases by 1 with each grid cell as you move from south to north).
- **Longitude**: The x-coordinate value transformed to longitude (degrees).
- **Latitude**: The y-coordinate value transformed to latitude (degrees).
- **PredAvg**: The (natural log-transformed) model-predicted value of the mean concentration for the location given by the specified latitude and longitude.
- **PredStd**: The (natural log-transformed) model-predicted value of the standard error for the mean predicted concentration.
- **BiasAvg**: Mean of the "bias surface," or the estimate of bias in the CMAQ or CAMx response surface relative to the monitoring station data (which represents "true" concentrations).
- **BiasStd**: Standard error of the mean of the bias surface.
- **CovarAvg**: This parameter is obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua and Terra satellites. The number -999 denotes a missing value. [This is implemented in the Model 6 version of T-SpACE.]
- **CovarStd**: This parameter is obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua and Terra satellites. The number -999 denotes a missing value. [This is implemented in the Model 6 version of T-SpACE.]
- **MonitorData**: The (natural log-transformed) air pollutant measurement from any air pollution monitoring station present at the given location (when monitor is present).

For variables with data values of -999, this indicates missing values. Values of -999 for the variable MonitorData indicate that no monitoring station exists at the location(s) represented in these (table) rows.

2.1.4 Summary statistics and validation of the resulting air quality surfaces

The user can run validation statistics at the completion of the simulation run. The first sets of statistics calculated are the distributional summary statistics for each of the model parameters, a summary of the number of iterations, the mean of the simulated values, and the standard error observed from the simulated values. In addition, the validation compares the T-SpACE results with a Krige model that makes predictions solely from monitoring data (using an exponential variogram approach in which results from daily variograms are combined) and predictions based solely on the CMAQ (or CAMx) modeling system (i.e., one of the input data sources to the T-SpACE). The summary statistics provided include the mean square error (MSE) and bias for the T-SpACE, kriging, and the CMAQ (or CAMx) model surface, the estimated percentage of time (across all monitoring days and locations) that the calculated prediction intervals (i.e., 95% credible intervals for the T-SpACE) actually included the observed monitor value, considering data for those locations in the validation dataset that contain monitors, and the prediction intervals that represent bounds on the range of measurements that are expected to contain the (unknown) true value a certain percent of the time (typically, 95%) if the model assumptions are correct.

2.1.5 Three (consecutive) "year" average of 4th highest concentration surface (O_3)

Along with predicting average concentrations within each grid cell on a daily basis, the T-SpACE has the capability to predict the fourth highest daily maximum 8-hour average ozone (O_3) concentration value that would be expected to occur within the time period of interest (e.g., one year) within each grid cell and to average three such surfaces together to create a 3 consecutive "year" average of the 4th highest concentration surface known as the EPA 'design value' for O_3. "Year" is defined as the time frame of interest. The time frame could range from January 1 to December 31 for a year.

Within a time period of interest (i.e., one year), the T-SpACE software determines the fourth highest average daily concentration value for ozone (O_3) within each grid cell across all simulated surfaces that the software generates for each day within the time period (as noted in Section 4.1). The results are averaged and smoothed to generate a single response surface. The software generates a comma-separated value (CSV) file that contains the results. There are only four columns included in this CSV file:

- **XCoord**: The cell's x-coordinate (east-west) within the grid (where the coordinate value increases by 1 with each grid cell as you move from west to east).
- **YCoord**: The cell's y-coordinate (north-south) within the grid (where the coordinate value increases by 1 with each grid cell as you move from south to north).
- **PredAvg**: The (natural log-transformed) T-SpACE-estimated value of the fourth highest average daily concentration for the given grid cell.
- **PredStd**: The (natural log-transformed) T-SpACE-estimated value of the standard error for the fourth highest average daily concentration for the given grid cell.

XCoord	YCoord	4HAvg	4HStd
1	1	3.76241	0.173166
1	2	3.8159	0.16131
1	3	3.89272	0.139333

In particular, because the air pollution concentration response surface for the average of the 4th highest daily maximum 8-hour averaged ozone O_3 concentration represents the entire time period of interest (i.e., one year), a date column is not necessary in this ***.CSV** file.

2.2 Operational behavior of the software

2.2.1 System requirements

The system requirements for the Hierarchical Bayesian (Air Pollution) Model (**Model 5:** T-SpACE) software are as follows:

- The target environment is the Windows XP or higher operating system for both the 32 and 64 bit platforms. The software will also work on Windows Vista platform, but was not developed specifically for this platform.
- SAS 9.1.3 or higher is required on the system to run the data validation routines.
- For the initial input monitor data, the user must supply a text file that is in the exact same format as EPA's AQS RD (raw data) monitor data.

- For the initial input model based data, the user must supply a file that is formatted like the CMAQ NetCDF files.
- A file formatted like the CMAQ NetCDF files is required for model execution, even if initial input model based data file already exists.
- The user must understand the T-SpACE model to make informed decisions about the simulation model parameters.
- The user will need to disable real-time file scanning anti-virus software. If the user chooses to run a simulation for a full calendar-year of data and for a full model-based grid, the output files that are created are large. They are opened and closed several times. When a file is accessed on a system on which anti-virus scan software is running in real-time, the anti-virus scan software will lock the T-SpACE simulation, and the T-SpACE software will not run correctly.

2.2.2 Typical processing times

The computer that was used to determine the operational behavior of T-SpACE was configured as follows:

- Intel Core 2 Duo E6750 Processor, @ 2.66 GHz
- 3.48 GB of RAM, @ 2.66 GHz
- Windows XP Professional, Service Pack 2

2.2.2.1 Preparing the input data for the simulation (Step 2)

CMAQ Model Input File

Running a 2001 simulation for **ozone** using T = (1, 365), X (1, 213), Y (1, 188) and the **ozone_2001_12.conc** CMAQ input file the time needed was 11 minutes and 16 seconds.

Running a 2001 simulation for **$PM_{2.5}$** using T = (1, 365), X (1, 213), Y (1, 188) and the **pm25_2001_12km.ioapi** CMAQ input file the time needed was 11 minutes and 53 seconds.

Monitor Model Input File

Running a 2001 simulation for **ozone** using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_44201_2001-0.txt** monitor input file the time needed was 3 minutes and 8 seconds.

Running a 2001 simulation for **$PM_{2.5}$** using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_88101_2001-0.txt** monitor input file the time needed 1 minute and 51 seconds.

2.2.2.2 Full year/Full grid simulations (Step 4)

Using the **ozone** files created as described above, the time needed to run the model was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps**:
  - Read input files – 35 seconds
  - Allocate variables – 1 minute, 15 seconds
  - Burn-in steps – 3 minutes, 27 seconds
  - Allocate variables – 55 seconds
  - ***T-SpACE model simulation steps* – 17 minutes, 8 seconds**

    - Write files – 8 minutes, 35 seconds
    - Add latitude and longitude – 7 minutes, 22 seconds
    - Write File – 1 minute, 24 seconds
    - Create validation chain file – 25 seconds
- **100 burn-in steps and 2500 simulation steps**:
    - Read input files – 53 seconds
    - Allocate variables – 1 minute, 15 seconds
    - Burn-in steps – 3 minutes, 26 seconds
    - Allocate variables – 1 minute, 16 seconds
    - ***T-SpACE model simulation steps* – 14 Hours, 21 minutes, 17 seconds**
    - Write files – 9 minutes, 52 seconds
    - Add latitude and longitude – 7 minutes, 49 seconds
    - Write File – 1 minute, 31 seconds
    - Create validation chain file – 34 seconds

Using the **$PM_{2.5}$** files created as described above, the time needed to run the model was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps**:
    - Read input files – 41 seconds
    - Allocate variables – 11 seconds
    - Burn-in steps – 3 minutes, 27 seconds
    - Allocate variables – 13 seconds
    - ***T-SpACE model simulation steps* – 17 minutes, 7 seconds**
    - Write files – 8 minutes, 9 seconds
    - Add latitude and longitude – 7 minutes, 45 seconds
    - Write File – 29 seconds
    - Create validation chain file – 11 seconds
- **100 burn-in steps and 2500 simulation steps**:
    - Read input files – 41 seconds
    - Allocate variables – 17 seconds
    - Burn-in steps – 3 minutes, 25 seconds
    - Allocate variables – 24 seconds
    - ***T-SpACE model simulation steps* – 14 Hours, 16 minutes, 50 seconds**
    - Write files – 9 minutes, 34 seconds
    - Add latitude and longitude – 7 minutes, 52 seconds
    - Write File – 32 seconds
    - Create validation chain file – 13 seconds

2.2.2.3 Validation (Step 5)

The time needed to validate the model output for the **ozone** simulation previously described was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps** – ***14 seconds***
- **100 burn-in steps and 2500 simulation steps** – ***5 minutes, 44 seconds***

The time needed to validate the model output for the **$PM_{2.5}$** simulations previously described was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps – *17 seconds***
- **100 burn-in steps and 2500 simulation steps – *6 minutes, 58 seconds***

3.0 Operation of and the statistical basis underlying the operation of the T-SpACE software model

The Bayesian approach featured by the T-SpACE software model is centered around *Bayes' Theorem,* an axiom on statistical probability, first developed by Thomas Bayes and presented pubically in 1763, but which has received considerable attention in recent years by statisticians and non-statisticians alike. Specifically, Bayes' Theorem relates how new evidence (data) is used to update or revise existing knowledge of a given unknown (random) process or model, thereby improving the ability to make more accurate and precise conclusions using that process or model, such as predicting the likelihood that something of interest will occur, or obtaining the best estimate of some unknown parameter such as air pollution levels over a given region. For example, if *A* represents such a random process, then existing knowledge of the possible values that *A* can hold and their probabilities of occurrence is specified by placing a *prior distribution* on *A*. Note that the selection of the prior distribution of *A* is ***subjective***, as it is based on existing knowledge at that time. Now, if *B* represents a random variable from which values are generated in an experiment, and the values of *B* are determined in part by the process or model represented by *A*, then Bayes' Theorem indicates that the probability of a specific outcome of *A* occurring is proportional to its probability under the prior distribution, multiplied by the likelihood of observing the particular set of data that was generated from *B* in the experiment, given that outcome of *A*. This updated likelihood is called the *posterior distribution* for *A*. The terms "prior" and "posterior" are introduced because they represent the basic concept behind Bayesian statistics and the T-SpACE model.

The basic statement of Bayes' Theorem is expanded in order to represent the situation handled by the T-SpACE model. Specifically, the likelihood that certain events occur can be expressed as a series of conditional probabilities. For example, if *A*, *B*, and *C* represent three different types of events, then the probability that *A*, *B*, and *C* occur simultaneously can be expressed as the product of the following probabilities:

- The probability that *C* occurs.
- The probability that *B* occurs given (or "conditional") that *C* occurs.
- The probability that *A* occurs given that both *B* and *C* occur.

Note that this breakdown of the problem is hierarchical in nature. Therefore, if the joint probability of {*A, B, C*} is very difficult or impossible to calculate, such as when it represents a complex multidimensional process over space and time (as with air pollution), such a hierarchical breakdown expresses the problem as a series of probabilities that can each be more easily calculated. While this concept has long been recognized in theoretical statistics, it has only recently been considered for modeling complicated environmental processes such as that used by T-SpACE for modeling air pollution levels.

In hierarchical Bayesian modeling, a common setup is to express the primary parameter of interest ("*A*" in the example above) as a function of available data possibly from multiple sources ("*B*"), which in turn can be expressed is a function of some random (often unmeasurable) process

and/or parameters from a model ("*C*"). (This is a very simplistic and basic expression of a common setup that can take different forms.) This begins to show why T-SpACE is a conceptually appealing approach to addressing a wide array of technical problems that are not easily addressed with standard statistical techniques. T-SpACE allows the user to

- Break complex multidimensional problems into simpler, logical, and intuitive hierarchies;
- Combine multiple sources of information, including expert opinion and differing data sources;
- Perform mathematical analyses consistent with the scientific method, and
- Implement "learning algorithms" on the unknown entity of interest that are further refined as more data become available.

In recent years, the rapid decrease in the cost of high performance computer equipment and significant focused research effort by the statistical community on Markov Chain Monte Carlo (MCMC) methods for fitting T-SpACE air pollution concentration surfaces (described below) have changed T-SpACE from a theoretical concept to a practical tool for use in decision sciences and health exposure assessment.

In this context, the T-SpACE model provides a system to predict air quality data for a specific time and spatial scale using monitoring and modeling data as input, while minimizing the limitations either of these methods when applied separately. The primary advantage of the T-SpACE model is increased flexibility and the ability to predict pollution gradients (and accompanying uncertainty in the prediction) that might otherwise be unknown using interpolation results (e.g., from kriging) based solely on relatively sparse monitoring data. Spatial maps of bias in numerical models are another useful output that will allow modelers to improve their models to minimize bias. Its major disadvantage is its computational burden.

The T-SpACE model considers the following series of random (unknown) variables and observed data, each of which is indexed in the range of space and time that is of interest to the user:

- W : (unobservable) log-transformed **true concentrations** of the pollutant (e.g., $PM_{2.5}$);
- X : (observable) ***unbiased*** log-transformed concentration **measurements** obtained **from monitoring** stations;
- Y : (observable) ***biased*** log-transformed concentration ***'measurements'*** **generated by an air quality (simulation) model**;
- D : (observable) set of variables that characterize the degree of bias present in the **modeled** (simulated) concentration values (relative to their representation of the true concentrations W); and
- C : (observable) covariates associated with the observed concentration measurements.

The first stage of the hierarchical bayesian process in T-SpACE expresses the observed (log-transformed) concentration measurements X and Y, as (each) originating from normal distributions with (unknown) mean and variance as follows:

$$X \sim N(W, \sigma_X^2) \qquad Y \sim N(W + D\beta_D, \sigma_Y^2)$$

While measurements originating from monitoring stations are assumed to be unbiased estimates of the true concentrations at their respective locations, the modeled (simulated) concentrations are assumed to be biased, as represented by the (unknown) parameters in β_D.

The <u>second stage</u> of T-SpACE modeling expresses the unknown, but true, (log-transformed) concentrations W, as an unknown overall mean, plus a random process Z, which consists of multiple components, and has a multivariate normal distribution with a mean of zero, and a covariance matrix that accounts for the presence of correlation among air pollution measurements that are collected successively in both space and time. Realizations of the random process Z for successive points in space and time may be correlated. Therefore, the covariance matrix for Z is characterized by two (unknown) parameters: σ_Z^2 and Δ_Z. The parameter σ_Z^2 represents the inherent variability in Z at a given point in space and time, and Δ_Z represents the correlation in values of Z that occur at successive points in time. In addition, correlation in values of Z that occur at neighboring points in space is represented by a "neighborhood autocorrelation structure", which the user of the T-SpACE model specifies.

Finally, the <u>third stage</u> of T-SpACE modeling expresses all of the unknown parameters above as random variables with distributions that are initially specified as "prior" distributions (i.e., before any data are collected). If all of the unknown parameters above are represented by θ, then Bayes' Theorem states the joint distribution of the unknown true concentrations (W) and the unknown model parameters (θ) can be expressed as the product of two prior distributions:

$$[W, \theta] = [W \mid \theta] \cdot [\theta] \qquad (1)$$

(We use the notation $[x]$ to denote the joint distribution of multiple unknown quantities and $[x \mid y]$ to denote the conditional distribution of x given another set of variables y.). This abstraction looks simple, however, in reality, the relationship between the components can be complicated, and the individual components have a large number of dimensions. For example, the covariate representation (C) is in fact a computational complexity parameter embedded in the model $[W \mid \theta]$. Statistical algorithms to deal with these and other model complexities are relatively recent, the most notable of them being Markov Chain Monte Carlo (MCMC) algorithms (e.g., Gilks et. al., 1998; Gelman et. al., 2004).

In the T-SpACE model, the prior distribution (1) is updated through observation of monitoring data (X) and pollutant data generated by an air quality (simulation) model (Y). As noted above, both are modeled independently given the true pollutant concentrations (W) and some components of the unknown parameter vector: $[X \mid W, \theta]$ and $[Y \mid W, \theta]$. Bias present in the simulated concentration measurements (D) is a complexity factor embedded in the model $[Y \mid W, \theta]$.

The T-SpACE model makes its predictions through characterizing the "posterior distribution" $[W, \theta \mid X, Y]$, which captures all available information about the true underlying pollutant

concentrations as well as the parameters governing the relationships among these concentrations and the observed quantities:

$$[W,\theta \mid X,Y] \propto [X,Y \mid W,\theta]\cdot[W,\theta] \qquad (2)$$

3.1 Grid size

Because the air quality model-based data are retrospective 'predictions' (i.e., estimates) across an entire two-dimensional spatial domain, the set of model predictions at a given point in time resembles a smooth "surface", when it is graphically displayed, rather than simply a set of points at various locations. This surface, called a *response surface*, appears three-dimensional, in that it covers the two-dimensional spatial region and has a third dimension representing the magnitude of the predicted concentrations.

The air quality model-based estimates, such as from CMAQ or CAMx, are made relative to a specified rectangular locational *grid* that covers the entire region of interest. The grid is composed of a matrix of rectangular *cells.* Within each grid cell, the modeling system predicts average pollution levels at a given point in time. Each cell has equal area which is expressed as *h* x *v*, where *h* and *v* denote each cell's horizontal and vertical dimensions, respectively (in kilometers). Under the CMAQ modeling system, grid cells can either be 36 km x 36 km (which represents the "parent domain" covering the entire continental US) or 12 km x 12 km (which represents primarily the eastern and Midwest regions of the US up to 2007, and the entire US from 2008 and onward). Thus, as the values of *h* and *v* decrease for a fixed area, the CMAQ modeling system generates a greater density of predictions, which increases the resolution of the area's prediction surface. Currently, EPA recommends that locational grid resolutions not be smaller than a 12 km x 12 km resolution, as any smaller/finer resolution (e.g., 4 km x 4 km, which has been used to characterize urban core areas), when applied to a large region, can result in unacceptably high uncertainties in emissions and meteorological information.

The user is asked to specify the lower x and y and upper x and y coordinates of the grid over which the simulation will be run. As mentioned earlier, the user is provided guidance on the x, y-coordinates that cover specific regions of the US.

3.2 Timeframe

T represents the time frame over which the simulation will be run. The study period in this version of the software is limited to one calendar year. The user cannot choose a time frame that spans or splits between two different calendar years. The value T represents the number of days (i.e., Julian [calendar] Days) from January 1 to the value T that is chosen, where January 1 is day number 1.

3.3 Bias Spline

The bias is evaluated as a linear combination of 2nd order uniform bias spline functions defined over a 3-dimensional lattice of uniform knots. Using bias splines as basis functions for the bias allows the user the ability to control the degrees of freedom of the bias through the number of control points. Furthermore, the piece-wise nature of the bias spline functions respects the principle of locality, that is, local information does not affect regions far from the region where

the local information is defined. On the numerical side, bias splines allow a tensor factorization of the bias matrix into three matrixes $B^{r}_{i_r j_r} = b_{j_r}(i_r), r = 1,2,3$ for a total dimensionality of $N_1M_1 + N_2M_2 + N_3M_3$, very much less than the total dimension of the full D-matrix, which is $N_1M_1 \cdot N_2M_2 \cdot N_3M_3$.

Guidance on choosing the bias spline parameters and the number of control points.

Variable	Bias Spline Min	Bias Spline Max	Number of Control Points
T	First day of the time range over which the bias will be calculated. By default, this is the first day of the year (i.e., Jan 1st).	Last day of the time range over which the bias will be calculated. By default, this is the last day of the year (i.e., Dec 31st).	This defines the degrees of freedom in the model for time. By default, 4 is chosen representing a control point for each season of the year.
X	Represents the min x-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the min x-coordinate is 1, for the 12 km grid. If the user is using the 36 km grid, then the default should be 13.	Represents the maximum x-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the maximum possible x-coordinate for the 12 km is given, 213. If the user is working with the 36 km grid, then this should be 142.	This is the degrees of freedom for the x-coordinate in the bias function. 8 is the default entry. Please note that if this number gets too high the performance of the simulation may be slow.
Y	Represents the min y-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the min x-coordinate is 1, for the 12 km grid. If the user is using the 36 km grid, then the default should be 15.	Represents the maximum y-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the maximum possible x-coordinate for the 12 km is given, 188. If the user is working with the 36 km grid, then this should be 94.	This is the degrees of freedom for the x-coordinate in the bias function. 7 is the default entry. Please note that if this number gets too high the performance of the simulation may be slow.

3.4 Control Points

The following should be considered when choosing the number of control points:

- Since the number of control points is roughly equal to the number of degrees of freedom of the simulation, the user wants to be conscious of not over- or under-fitting the model by choosing too many or too few control points

- For T, time, the number of control points has been chosen to represent the four seasons. If the user believes there to be a monthly bias in the concentrations, then they may choose this value to be 12.

- For X and Y, the number of control points has been chosen to represent an estimated rate on which bias changes in each direction. A metric that may be used to choose the number of control points for X and Y could be the length of the grid over which the simulation is to run divided by the average scale of the spatial bias (every 25 grid cells). It is recommended that, in general, there should be ***less than 100 control points*** in the x-y coordinate system.

- Finally, it is recommended that a user of the system run sensitivity analyses to understand how sensitive the surface predictions are to the number of control points.

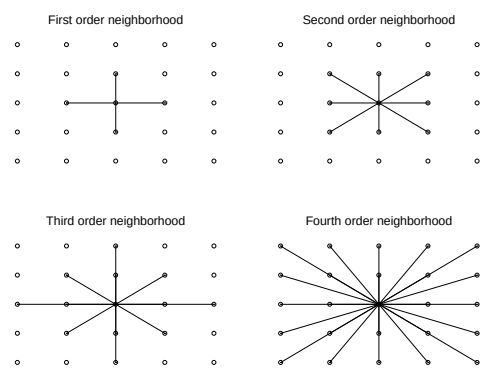
3.5 Priors

In Bayesian modeling, unknown model parameters are assumed to represent random quantities that originate from some underlying distribution. In the absence of any new data, the modeler selects a "prior distribution" for each parameter based on existing knowledge or their 'best judgment'. In the T-SpACE model, prior distributions are assigned as follows (to the model parameters introduced earlier):

- For the unknown variance parameters σ_X^2, σ_Y^2, and σ_Z^2, their reciprocals are known as τ_X^2, τ_Y^2, and τ_Z^2 respectively (where τ is the Greek letter "tau"), and are each assigned a gamma distribution (with shape and scale parameters specified by the user).
- The parameter D and the unknown "slope" parameters β_D are each assigned a normal distribution (with mean and variance specified by the user)
- The temporal correlation parameter Δ_Z is assigned a uniform distribution (with lower and upper bound specified by the user).

3.6 Neighborhood Boundary

The neighborhood boundary defines the neighborhood structure of the model. There are four possible neighborhood structures that are available for the simulation, first through fourth order neighborhoods. As seen in the figure, as the order increases, the number of neighbors used in the calculation increases. The default in the software is a first order neighborhood.



3.7 Burn-in/Loop

Users of the T-SpACE software encounter the reference to a "burn-in" period for the MCMC. This signifies the number of initial model runs, in the iterative process, whose results are discarded from consideration in the model computations, as they are assumed to represent estimates that deviate considerably from the final posterior distribution. The ideal situation is to construct an MCMC algorithm that has a short burn-in period and yields stable 'draws' (selections) from the posterior distribution.

3.8 Markov Chain Monte Carlo (MCMC) Simulation

Because it is not possible to obtain a closed form of a posterior distribution on the air pollutant concentrations (i.e., fitting the T-SpACE model) using standard statistical modeling techniques, an alternative, computationally-intensive procedure is needed to obtain a good approximation of the posterior distribution. This procedure involves performing a computer simulation centered on a Markov Chain process in an effort to converge numerically to the proper distribution. A series of random variables $\{X_1, X_2, \ldots, X_n\}$ representing some ordered process (e.g., values over time) is known as a *Markov Chain* if the probability that one of the variables (X_i) holds a specific value depends only on the value of the variable that immediately precedes it (X_{i-1}) and is independent of the values of all other prior variables ($X_i, \ldots, X_{i-2}$). For example, if you are about to roll a die to determine how many spaces to advance in a board game, the probability that you will land on a particular space as a result of that roll is dependent only on where you currently

are on the board, and not on the outcome of any prior rolls that determined how you got to your current position. This is a simple example of a Markov Chain process.

In the MCMC process used in T-SpACE, the parameters of the posterior distribution [W, θ | X, Y] are random variables that must be estimated. They are assigned initial estimates (based on the assigned 'prior' distribution), which are then repeatedly updated in a linear, stage-wise Markov Chain process, where the estimates in one stage of the process depend only on the estimates in the prior stage. The stage-wise process allows for the estimates to slowly improve until a state of "stationarity" is achieved, that is when the estimates have converged to a final set of values, and ***no further significant improvement*** can be made. In statistics, the term "Monte Carlo" refers to the use of random simulation techniques to characterize some unknown physical or mathematical system. The Monte Carlo component of the MCMC is the series of simulation runs in which one moves in a probabilistic fashion from one stage to the next in the Markov Chain process to get to the final posterior distribution. The most widely used simulation technique in MCMC is known as the *Gibbs sampler* (e.g., Gelfand and Smith, 1990). To perform MCMC, the T-SpACE model uses an algorithm based on the Gibbs Sampler. In this algorithm, the posterior distribution [W, θ | X, Y] is sampled by simulating successively from the steps:

$$\begin{aligned} &[W \mid \theta, X, Y] \\ &[\theta \mid W, X, Y] \end{aligned} \tag{3}$$

Each step is conditioned on the latest values obtained from the previous step. These distributions are referred to as the 'full conditional' distributions. When these full conditional distributions are not in a recognizable form and can only be calculated up to a normalizing constant, the algorithm incorporates a *Metropolis*-type step (Robert and Casella, 2004). Software such as WinBugs is available to perform MCMC techniques, although specially-prepared software has been developed to handle the MCMC necessary to fit the T-SpACE model.

4.0 CMAQ and CAMx Air Quality Models Used in T-SpACE

4.1 CAMx model description

The Comprehensive Air Quality Model with Extensions (CAMx) is an Eulerian (gridded) photochemical dispersion model that allows for integrated "one-atmosphere" assessments of tropospheric air pollution (e.g., ozone, particulates, air toxics) over spatial scales ranging from neighborhoods (local) to continents (global). It is a modern, open-source system that is computationally efficient, flexible, and pubically available. CAMx's Fortran source code is modular and well-documented. The Fortran binary input/output file formats are based on the Urban Airshed Model (UAM) convention, and are compatible with many existing pre- and post-processing tools. Meteorological input fields are supplied to CAMx from separate weather prediction models. CAMx specifically supports the Weather Research and Forecasting (WRF) model, the Mesoscale Model 5 (MM5) and the Regional Atmospheric Modeling System (RAMS). All emission inputs are supplied from external pre-processing systems (e.g., the Sparse Matrix Operator Kernel (SMOKE) system and the Emissions Processor System Version 3 (EPS3).

CAMx simulates the emission, dispersion, chemical reaction, and removal of pollutants by 'marching' the Eulerian continuity equation forward in time (t) for each chemical species (l) on a system of nested three-dimensional grids. The continuity equation specifically describes the time dependency of volume-average species concentration within each grid cell as a sum of all physical and chemical processes operating on that volume. CAMx can perform simulations on four types of Cartesian map projections: Lambert Conic Conformal, Polar Stereographic, Mercator, and Universal Transverse Mercator (UTM). CAMx also offers the option of operating on a geodetic latitude/longitude grid system. The vertical grid structure is defined externally, so layer interface heights may be specified as any arbitrary function of space and/or time. This flexibility in defining the horizontal and vertical grid structures allows CAMx to be configured to match the grid of any meteorological model that is used to provide environmental input fields. Detailed information on the CAMx system can be found at: http://www.camx.com.

4.2 CMAQ model description

The Models-3/Community Multiscale Air Quality (CMAQ) modeling system goals are to improve: 1) the environmental management community's ability to evaluate the impact of air quality management practices for multiple pollutants at multiple scales, and; 2) the scientist's ability to better probe, understand, and simulate chemical and physical interactions in the atmosphere. Traditionally, the CMAQ modeling system has been used to predict air quality across an entire regional or national domain, and then to simulate the effects of various changes in emission levels for policy-making purposes. However, for health studies, it is frequently applied to provide supplemental information on predicted air quality in areas where few or no monitors exist. It also accounts for meteorological conditions to better evaluate health outcomes. Detailed information on the CMAQ modeling system is available at https://www.epa.gov/cmaq and at https://www.cmascenter.org.

Because the CMAQ modeling system makes predictions across an entire two-dimensional domain, the set of model predictions at a given point in time resembles a smooth "surface" when it is graphically displayed, rather than simply a set of points at various locations. This surface, called a *response surface*, appears three-dimensional in that it covers the two-dimensional region and has a third dimension representing the magnitude of the predicted concentrations. The CMAQ model predictions are made relative to a specified rectangular *grid* that covers the entire region of interest. The grid is composed of a matrix of rectangular *cells.* Within each cell, the CMAQ modeling system predicts average pollution levels at a given point in time. Each cell has equal area which is expressed as h x v, where h and v denote each cell's horizontal and vertical dimensions, respectively (in kilometers). Under the CMAQ modeling system, grid cells can either be 36 km x 36 km (which represents the "parent domain" covering the entire continental US) or 12 km x 12 km (which represents primarily the eastern and Midwest regions of the US up to 2007, and the entire US from 2008 and onward). Thus, as the values of h and v decrease for a fixed area, the CMAQ modeling system generates a greater density of predictions, which increases the resolution of the area's prediction surface.

5.0 References

[1] T.W. Tesche, Ralph Morris, Gail Tonnesen, Dennis McNally, James Boylan, Patricia Brewer, "CMAQ/CAMx annual 2002 performance evaluation over the eastern US", Atmospheric Environment 40 (2006) 4906–4919.

[2] Jinyou Liang, Philip T. Martien, Su-Tzai Soong, and Saffet Tanrikulu, A Photochemical Model Comparison Study: CAMx and CMAQ Performance in Central California, Proceedings of the Thirteenth Conference on the Application of Air Pollution Meteorology with the Air and Waste Management Association, Vancouver Canada; American Meteorological Society, Boston (2004)

[3] McMillan, N.J., Coutant, B.W., Morara, M., Young, G.S., and Zhou, J. (2005). "Combining Monitored and Modeled $PM_{2.5}$ and Satellite Aerosol Optical Depth Information Across Space and Time Using Bayesian Modeling" Technical Memorandum from Battelle to US EPA, Office of Air Quality Planning and Standards, Contract No. 68-D-98-030, Work Assignment 3-06.

[4] User's Guide: Hierarchical Bayesian Space-Time Modeling of Air Pollution Data. April 2009. Version 5_4h.0.1.21.

[5] US EPA Final Report: Overview of EPA's Hierarchical Bayesian Model For Predicting Air Quality Patterns In The United States Over Space And Time, For Use With Public Health Tracking Data. Contract No. EP-D-04-068, Work Assignment 54, Task 1, April 15, 2009.

[6] Draft Final Report: Hierarchical Bayesian Model Software Architecture/Design Document. April 30, 2009. Contract No. EP-D-04-068, Work Assignment 54, Task 4.

[7] USEPA, Draft Report Hierarchical Bayesian Model Evaluation For Ozone Data, April 30, 2009, Contract No. EP-D-04-068, Work Assignment 54.

[8] User's Guide: Comprehensive Air Quality Model With Extensions Version 6.40, Ramboll Environ, December 2016, http://www.camx.com/files/camxusersguide_v6-40.pdf.

Appendix A

TEMPORAL-SPATIAL AMBIENT CONCENTRATOR ESTIMATOR (T-SPACE) SOFTWARE ARCHITECTURE/DESIGN

1. INTRODUCTION

1.1 SOFTWARE OVERVIEW

1.1.1 Description

EPA has developed a method to combine air pollution monitoring network data and simulation model output to maximize the advantages offered by both data sources while minimizing the disadvantages of each when predicting fine particulate matter ($PM_{2.5}$) and ozone (O_3) concentrations. Air quality (simulation) models estimate the spatial and temporal gradients of air pollution based on emissions inventories and meteorological information. These models, while providing estimates over large regions at relatively low cost, have been found to have statistical bias and have greater errors than air pollution monitoring networks. Air pollution monitoring networks generally provide relatively accurate, unbiased results. However, because monitoring networks are sparsely and irregularly spaced over large spatial domains, they are not able to produce large-scale predictions.

EPA developed the T-SpACE model that combines monitoring data from the 2,545 National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) with estimated air pollution concentrations from air quality models such as the Models-3/Community Multiscale Air Quality (CMAQ) and the Comprehensive Air Quality Model with Extensions (CAMx), which maximizes the advantages offered by both data sources, while minimizing the disadvantages of each.

1.1.2 Purpose

The T-SpACE model requires data preparation and model parameter input from the user to run a simulation to produce an estimated air pollution concentration surface. To streamline the model set-up, the user is allowed to prepare the input data for the model run, choose the model specifications, run the model, receive output files containing the estimated air pollution concentration surface, and run a set of validation procedures to compare the T-SpACE model run with a standard kriging method, and the model-based air pollution output concentrations from CMAQ or CAMx. To minimize the amount of work for the user, a graphical user interface (GUI) “wrapper” was developed that provides a single interface for configuring and running a simulation. The user is asked for particular choices for input data sets, model parameters, and names for the output files, and is guided through the simulation.

1.1.3 Scope

This document was generated to provide a detailed description of the architecture of the Hierarchical Bayesian Space-Time Model for Air Pollution, Version 5-4h.0.1.21 (T-SpACE). In particular, this document provides information on the physical structure of the source code, the logical structure of the classes, the input data, the output data, and the expected behavior of the application.

1.2 GLOSSARY

1.2.1 Definitions

Abstract Class	An object-oriented programming class that provides basic methods (functions/procedures) and data attributes (variables/parameters) which are specifically defined in lower-level classes through the mechanism of inheritance. This is a 'template' for building a functional class.
Application Programming Interface	A set of protocols for using a programming library.
Assembly	A compiled low-level machine code (library).
Base Class	A functional object-oriented programming class that is derived from an abstract object-oriented programming class. It provides defined methods (functions/procedures) and data attributes (variables/parameters) which can be inherited in lower level (child[ren]) classes. This is a 'template' for building an 'object' (object = and 'instance'/working example of the base class).
Bitmap	Image file format for storing digital images.
C#	An object-oriented programming language developed by Microsoft.
Console Application	A program that has no user interface. It is run from the computer command line.
DOS Batch File	A Disk Operating System (DOS) file that contains a series of command line instructions to be run in sequence.
DOS Short File Name	A file name compatible with the DOS operating system, allowing up to 8 characters in the file name, 3 characters in the extension and a limited allowable character set.

Dynamic-Linking Library	A shared library for use on the Windows platform. Files in the Dynamic-Linking Library are designated by the '.dll' extension.
Kriging	A geostatistical method that uses interpolation to develop a response (air pollution concentration) surface from the nearest set of monitors.
.NET Framework	A software framework for the Windows platform.
Singleton Class	A object-oriented programming class that generates a single instance (object) of itself at any given time.
Static Method	A method that is shared by all instances of a class. It can be invoked without creating an instance of the class.
Visual Basic	An object-oriented programming language.
Visual Studio	An integrated development environment used for programming in C# and Visual Basic.
Windows Console Project	A Visual Studio project for creating a console application.
Windows Deployment Project	A Visual Studio project for creating an installation package for an application.
Windows Forms Project	A Visual Studio project for creating a Windows application with a GUI (Graphical User Interface).
Windows Library Project	A Visual Studio project for creating a DLL.
Windows Long File Name	A file name that permits up to 255 characters with an expanded allowable character set.

1.2.2 Acronyms

API	Application Programming Interface
AQS	Air Quality System
CAMx	Comprehensive Air Quality Model with Extensions
CASTNET	Clean Air Status and Trends Network
CMAQ	Community Multiscale Air Quality

CSV	The file extension for a Comma Separated Value file
DLL	The file extension for a Dynamic-Link Library file
DOS	Disk Operating System
GUI	Graphical User Interface
HBM	Hierarchical Bayesian Model
IOAPI	Input/Output Application Programming Interface
MCMC	Markov Chain Monte Carlo
MSI	The file extension for a Microsoft Installer file
NAMS	National Air Monitoring Stations
NetCDF	Network Common Data Format
OOP	Object-Oriented Programming
POC	Parameter Object Code
PNG	The file extension for a Portable Network Graphic file
RAM	Random Access Memory
SAS	Statistical Analysis System
SLAMS	State and Local Air Monitoring Stations
TauX	The precision of the measurement error in the air quality monitor observations
TauY	The precision of the measurement error in the air quality model ‘observations’
TauZ	The precision of the mean process
T-SpACE	Temporal-Spatial Ambient Concentrator Estimator
TSV	The file extension for a Tab-Separated Value file
UC	User Controls

1.2.3 Abbreviations

BMP	The file extension for a Bitmap file
EXE	The file extension for an Executable file
GHz	Gigahertz
ICO	The file extension for an Icon file
K	Kilobyte. 1K = 1,024 bytes
KM	Kilometer. 1KM is approximately 0.625 miles
MB	Megabyte. 1MB = 1,024K = 1,048,576 bytes
O_3	Ozone (pollutant)
O3	Ozone (file-name)
$PM_{2.5}$	Fine Particulate Matter less than 2.5 microns in (aerodynamic) diameter
TXT	The file extension for a Text file

1.3 REFERENCES

[1] McMillan, N.J., Holland, D.M., Morara, M., and Feng, J., Combining different sources of particulate data using Bayesian space-time modeling, Environmetrics, 2010, Volume 21, pp 48 – 65, DOI: 10.1002/env.984

[2] USEPA, Draft Report Overview of EPA's Hierarchical Bayesian Model For Predicting Air Quality Patterns In The United States Over Space And Time, For Use With Public Health Tracking Data, March 13, 2009, Contract No. EP-D-04-068, Work Assignment 54.

[3] User's Guide: Hierarchical Bayesian Space-Time Modeling of Air Pollution Data, March 2009, Version 5-4h.0.1.21.

2. DESIGN DESCRIPTIONS

This chapter discusses the design of the T-SpACE Model software. Software design is the process of problem-solving and planning for a software solution.

2.1 DESIGN OVERVIEW

This section provides an overview of the design of the T-SpACE Model software. Each subsection details an aspect of the design.

2.1.1 Requirements

The requirements for the T-SpACE Model software are as follows:

- The target environment is the Windows XP, or higher operating system for both the 32- and 64-bit platforms. The software will work on the Windows Vista platform, but was not developed specifically for this platform.

- SAS 9.1.3 or higher is required on the system to run the data validation routines.

- For the initial input air quality monitor data, the user must supply a text file that is in the exact same format as EPA's AQS RD (raw data) monitor data.

- For the initial input air quality model data, the user must supply a file that is formatted like the CMAQ NetCDF files.

- A file formatted like the CMAQ NetCDF files is required for model execution. When the CAMx model is used in T-SpACE, its format is converted into the CMAQ NetCDF file format by T-SpACE.

- The user must understand the T-SpACE model to make informed decisions about the simulation model parameters. There are three documents that can help with this understanding: [1], [2], [3].

- The user will need to disable real-time file scanning/anti-virus software. If the user chooses to run a simulation for a full calendar-year of data for a full model-based grid, the output files that are created are large. They are opened and closed several times. When a file is accessed on a system on which file scanning/anti-virus software is running in real-time, the file scanning/anti-virus software will 'lock' the T-SpACE simulation file, and the T-SpACE model software will not run correctly.

2.1.2 Tradeoffs

There were several design trade-offs that were made in the design of this system. The system was developed from an existing set of SAS programs, a Markov Chain Monte Carlo (MCMC) model executable that required a particular structure for the input information and had a specified structure for the output results, and monitor and model-based concentration data that had a

specified structure. Throughout the development process, decisions were made to allow the T-SpACE model to work within the established design constructs, and implement EPA requirements. The system design trade-offs included the following:

- A decision was made to have EPA scientists develop the algorithms for:
  - Conversion of the NetCDF model-based concentration files to the required format for input into the MCMC model executable;
  - Conversion of the AQS monitor data to the air quality model-based grid, and subsequently to the required format for input into the MCMC model executable;
  - Adding the latitude and longitude to the MCMC simulation output file.

  The decision was made to ensure that the data transformations used for these data types were implemented according to EPA standards. The code that performs the data transformations had been previously designed by EPA, and was easily updated. When changes to the format of the input or output data was made during the design process, additional routines were written to open up the files, make the requested changes, and close the files. This led to the issue with the real-time, virus scanning software mentioned above.

  The designed model processes are pollutant-specific, limiting the simulations to $PM_{2.5}$ and ozone. For instance, there are two separate executables to process ozone and $PM_{2.5}$ CMAQ or CAMx data. Additional air pollutants will require additional simulation algorithms and associated components.

- Utilization of SAS and the previously developed SAS validation programs to run the validation routine.
  - This decision requires that any computer system that will be using the validation portion of the software must have SAS installed on that system.

  This decision was made because the statistical functions in SAS have been validated, and the previously developed SAS validation programs for T-SpACE have been verified.

2.2 USER INTERFACE

Each subsection provides a brief description of the forms that were developed as part of the T-SpACE user interface. The subsection headings represent the name of the form, visible in the form's title bar.

2.2.1 Hierarchical Bayesian Melding

This form acts as a container for the forms needed to prepare, run and validate simulations. This form also contains the buttons used to walk through the modeling process.

2.2.2 Step 1: Choose Time/Grid

This form is used for Step 1 of the process, where the user specifies the temporal and spatial grids for the simulation, as well as the folder that will hold the files created in Step 2.

2.2.3 Step 2: Prepare Model Input Data

This form is used for Step 2 of the process, where the user generates the air quality model-based and air pollution monitor files that will be input to the simulation. The raw air quality model and air pollution monitor data files are transformed into the required format in this step.

2.2.4 Step 3: Model Specification

This form is used for Step 3 of the process, where the user chooses the simulation parameters.

2.2.5 Step 4: Launch Model

This step does not have, nor does it require, a configurable GUI/'form' setup process for the user to select/modify simulation parameters. All of the necessary configuration setup steps for the model run (implemented in Step 4) have been performed in the previous 3 Steps (and other associated pre-Step 4 preparation/activities). When the Step 4 button is pushed, a DOS window opens and at this point, the T-SpACE model simulation run begins.

2.2.6 Step 5: Launch Validation

This form is used for Step 5 of the process, where the user specifies the kriging and network files used to validate the simulation output.

2.2.7 Average 4th Highest Surface Files

This form is where the user specifies three consecutive 4th highest (daily maximum 8-hour average ozone O_3 concentration) surface files to average into a single ozone concentration surface file.

2.2.8 File Compare

This form is where the user specifies 2 AQS monitor files, or 2 air sites (airsite.tsv) files, to compare in order to determine differences in monitor locations.

2.2.9 Apply States to Grid

This form is where the user can create a state grid file by applying state and region definitions to cells in a (locational) grid file.

2.2.10 Step 6: Visualize Surface

Step 6 of the process is where the user generates a three-dimensional (3-D: latitude, longitude, concentration) graphical representation of the air pollution concentration surface generated in T-SpACE.

2.3 SYSTEM DESIGN

The system design section provides a detailed listing of the physical structure of the application's source code modules and support files. There are two components: Projects and Other Files.

The projects listed below are each a set of source code modules that compile into a single executable or DLL.

The other files are those files developed elsewhere that are required to run the software.

2.3.1 Projects

The subsection heading is the project name as found in the source code. Each module in the project is listed with a brief description.

2.3.1.1 **BayesianMelding**

This is a Windows Forms project written in Visual Basic 2005. The output of this project, **GUI 2.1.exe**, is referenced by the main application. BayesianMelding is used to provide the user with a friendly interface with which to maintain the cryptic parameter files used by the model.

Modules:

- **frmParams.vb** – The form as well as all logic needed to verify input, load parameter files and save parameter files.
- **StringTokenizer.vb** – Class used to parse (split) strings into individual tokens based on a specified delimiter.

2.3.1.2 **BayesianMonitorCheck**

This is a Windows Console project written in C# 2005. The output of this project, **BayesianMonitorCheck.exe**, is located in the application folder. This application is used to post-process the monitor $PM_{2.5}$ file (containing $PM_{2.5}$ monitor concentration values) created in Step 2 so that it can be used in the model. This post-processing consists of:

- Removing any data record that has a negative value for row and/or column
- Replaces "m" in the logpm column with an empty string (this represents a missing value)

Modules:

- **Program.cs** – Contains the program's entry point and all logic needed to complete the processing

2.3.1.3 **CMAQColumnRenamer**

This is a Windows Console project written in C# 2005. The output of this project, **CMAQColumnRenamer.exe**, is located in the application folder. This application is used to post-process the CMAQ or CAMx (Ozone and $PM_{2.5}$) files created in Step 2 so that the column names are consistent. These column names are:

- Day
- Column
- Row
- CMAQ log (ozone)/CMAQ log (pm) – **<u>This is a generic designation that incorporates</u>:** CAMx log (ozone)/CAMx log (pm)

Modules:

- **Program.cs** – Contains the program's entry point and all logic needed to complete the processing.

2.3.1.4 *CMAQPostProcessor*
This is a Windows Console project written in C# 2005. The output of this project, **CMAQPostProcessor.exe**, is located in the application folder. This application is used to post-process the CMAQ $PM_{2.5}$ file created in Step 2 so that it can be used in the model. This post-processing consists of:

- Replaces the value in the average column with the natural log of the average.

Modules:

- **Program.cs** – Contains the program's entry point and all logic needed to complete the processing.

2.3.1.5 **CreateValidationChainFile**
This is a Windows Console project written in C# 2005. The output of this project, **ValidationChainFileCreator.exe**, is located in the application folder. This application is used to post-process the chain file created in Step 4 so that it can be used in the validation step. This post-processing consists of:

- Removes all columns except for TauX, TauY, and TauZ

Modules:

- **Program.cs** – Contains the program's entry point and all logic needed to complete the processing.

2.3.1.6 **EPAWA***41State*
This is a Windows Library project written in C# 2005. The output of this project, **EPAWA41State.dll**, is a reference of the main application. This class serves as a repository for values that are needed throughout program execution. These values are accessible as properties of the **WA41_StateTracker** object.

Modules:

- **WA41_StateTracker.cs** – Contains the StateTracker's (**WA41_StateTracker** object) properties as well as the methods/procedures that access these properties.
- **DosLongToShortPathConverter.cs** – Contains static methods that:
    - Convert a file's Windows long file name to its DOS short file name
    - Convert a file's DOS short file name to its Windows long file name

2.3.1.7 **MonitorColumnRenamer**
This is a Windows Console project written in C# 2005. The output of this project, **MonitorColumnRenamer.exe**, is located in the application folder. This application is used to post-process the Monitor (Ozone and $PM_{2.5}$) files created in Step 2 so that the column names are consistent. These column names are:

- Day
- Column
- Row
- monitor log (ozone) / monitor log (pm)

Modules:

- **Program.cs –** Contains the program's entry point and all logic needed to complete the processing.

2.3.1.8 **Bayesian** *Air Pollution Model*
This is a Windows Deployment project written in Visual Studio 2005. The output of this project, **Bayesian Air Pollution Model.msi** and **setup.exe**, are used to install the application on the target computer.

It has the following detected dependencies (i.e., files that must be already generated/present and contain the requisite data required to initiate the process):

- **.NET Framework 2.0**
- **EPAWA41State.dll**
- **GUI 2.1.exe**

2.3.1.9 **EPAWA***41*
This is a Windows Forms project written in C# 2005. The output of this project, **EPAWA41.exe**, is the main application. The EPAWA41 application is used to provide the user with a single interface to execute the steps needed to run a simulation. Since this project has several modules, multiple subfolders have been created for organization. The module list is provided by subfolder.

Modules:

- Application folder
  - **program.cs** – Contains the program's entry point and launches the main form.
- BatchFileProcessors subfolder
  - **BatchFileCreator.cs** – Class for creating DOS batch files needed in Step 2.
  - **BatchFileCreatorAddLL.cs** – Class for creating DOS batch files needed to execute **ADDLL.exe** (Add Latitude and Longitude executable file).
  - **BatchRunner.cs** – Class for executing DOS batch files.
- CSVFileProcessors subfolder
  - **CSVProducer.cs** – Abstract class used as a basis for the batch file creator classes.
  - **CSVProducerNetCDF.cs** – Class that contains the interoperability functions that allows the application to use the NetCDF API library.
  - **IOAPILatLonInserter.cs** – Class that inserts latitude and longitude information from CMAQ/CAMx IOAPI files into the CMAQ/CAMx and monitor files created in Step 2.
  - **RowColToLatLongConverter.cs** – Class that inserts latitude and longitude information from CMAQ/CAMx IOAPI files into the surface file created by the model.
  - **TextLatLonInserter.cs** – Class that inserts latitude and longitude information from a cells text file (cells*.txt) into the CMAQ/CAMx and monitor files created in Step 2.
- Forms subfolder
  - **Average4thHighest** – Form that allows the user to select three consecutive, yearly 4th Highest Surface files and average them into a single concentration surface file.
  - **FileCompare** – Form that allows the user to compare two AQS monitor files or two site TSV files to see which monitor site locations are missing from one of the files.

- **Main** – This is the application's main form, the one that opens when the application starts.
- **Step2Ozone** – Form that allows the user to create the CMAQ/CAMx and/or monitor file for ozone. This is a Step 2 element.
- **Step2PM25** - Form that allows the user to create the CMAQ/CAMx and/or monitor file for $PM_{2.5}$. This is a Step 2 element.
- **Validation** – Form that allows the user to run a validation SAS program against the model output. This is a Step 5 element.

  WelcomeIntro – Form that allows the user to specify the time period and geographic area of interest for the simulation. This is a Step 1 element.

- Resources subfolder
  - **CSVFile3.bmp** – Bitmap representing a CSV file.
  - **Earth.bmp** – Bitmap representing the Earth.
  - **Earth.ico** – Application's icon.
  - **GreenArrow.bmp** – Bitmap representing a green arrow.
  - **GreenGoArrowBitmap.bmp** – Bitmap representing a large green arrow.
  - **GridToEarthModify.bmp** – Bitmap representing the application of a grid to the Earth.
  - **GridToEarthModifyPNG.png** – PNG representing the application of a grid to the Earth.
  - **Home2.bmp** – Bitmap representing a home.
  - **HomeInPS.bmp** – Bitmap representing a home.
  - **OpenFile.bmp** – Bitmap representing the opening of a file.
  - **OpenFolder.png** – Bitmap representing the opening of a folder.
  - **PM25.bmp** – Bitmap representing the $PM_{2.5}$ air pollutant.
  - **PM25Red.bmp** – Bitmap representing the $PM_{2.5}$ air pollutant.
  - **SmallDisk.bmp** – Bitmap representing a diskette.
- Tools subfolder
  - **AQSFileComparer.cs** – Class that compares two AQS monitor files.
  - **FileComparer.cs** – Base class that provides functionality for all file comparers.
  - **MonitorSite.cs** – Class that represents a monitor site from an AQS monitor file or an air pollution monitor site TSV file.
  - **SurfaceCell.cs** – Class that represents a cell from a surface file.
  - **SurfaceFileAverager.cs** – Class that averages 4th highest surface files.
  - **TSVFileComparer.cs** – Class that compares two sites TSV files.
- UserControls subfolder
  - **CalendarRangeSelector** – Contains the two calendar controls used in Step 1.
  - **TabPageGoBetween** – Container for the two tabs in the Step 2 form.
  - **UC-CMAQ** – The CMAQ/CAMx interface for Step 2.
  - **UC-IntroScreen** – The interface for Step 1.
  - **UC-Monitor** – The monitor interface for Step 2.
- Utility subfolder
  - **AppSettings.cs** – Class that saves application settings between program sessions.
  - **TextFileCreator.cs** – Class for saving batch file output to text files.
  - **ValidationAppSettings.cs** – Class that saves validation application settings between program sessions.

2.3.2 Other Files

This section lists the files that are written to the computer when T-SpACE is installed. These files are different from the output of the project files listed previously. This list is presented by folder. A description of each file is provided. The following files are placed in the application folder upon installation of the T-SpACE software. The application folder uses 45.8 MB of disk space.

Application Folder:

- **addll.exe** – Add latitude and longitude (addll) executable provided by the EPA. It inserts latitude and longitude information in the air pollution concentration surface file created by the T-SpACE software. This is automatically executed when the model completes.
- **airsites.tsv** – Tab-delimited file provided by the EPA. It contains the locations of the $PM_{2.5}$ monitors. This is used by monpm25.exe.
- **AqVis4.exe** – Surface visualization software.
- **cmaqpm25.exe** – Executable provided by the EPA. It creates the CMAQ file for $PM_{2.5}$ in Step 2.
- **cmaqpm25O3.exe** – Executable provided by the EPA. It creates the CMAQ file for ozone in Step 2.
- **Model5-4H.0.1.19_Mc2_Win32.exe** – Model executable for the 32-bit platform. This is executed in Step 4.
- **Model5-4H.0.1.19_Mc2_Win64.exe** – Model executable for the 64-bit platform. This is executed in Step 4.
- **monO3.exe** – Executable provided by the EPA. It creates the monitor file for ozone in Step 2.
- **monO3airsites.tsv** – Tab-delimited file provided by the EPA. It contains the locations of the ozone monitors. This is used by monO3.exe
- **monpm25.exe** – Executable provided by the EPA. It creates the monitor file for $PM_{2.5}$ in Step 2.
- **netcdf.dll** – API for accessing NetCDF format data files.
- **tz.csv** – Comma-delimited file provided by the EPA. It contains time zone information. It is used by cmaqpm25.exe and cmaqpm25O3.exe.
- **ValidationTemplate.sas** – This is a template of a SAS program that is used to validate the model output files. The template is completed with user selected files and then executed in Step 5.

CellsFiles subfolder:

- **cells_12km_2001.txt** – Cells text file for the year 2001, 12 km grid.
- **cells_12km_2002.txt** – Cells text file for the year 2002, 12 km grid.
- **cells_36km_2001.txt** – Cells text file for the year 2001, 36 km grid.
- **cells_36km_2002.txt** – Cells text file for the year 2002, 36 km grid.

Help subfolder:

- **2001CMAQ12km.PNG** – Graphic depicting the 12 km grid for the year 2001.
- **2002CMAQ12km.PNG** – Graphic depicting the 12 km grid for the year 2002.
- **HBMUsersGuide.doc** – The User's Guide.

OtherDataFiles subfolder:

- **aod_fake.csv** – An empty file representing Aerosol Optical Depth (AOD) data. This is a remnant of the manual, multi-step process.
- **StateLatLon.csv** – This is a CSV representation of the States.txt file, including only the data needed to create a StateGrid.csv file.
  States.txt – Contains states with their latitude and longitude ranges. The source of this information is http://www.netstate.com.

StateFiles subfolder:

- **StateGrid2001-12km.csv** – Contains a list of states and regions, with their associated X, Y ranges according to the 2001 cells text file for the 12 km grid.
- **StateGrid2001-36km.csv** – Contains a list of states and regions, with their associated X, Y ranges according to the 2001 cells text file for the 36 km grid.
- **StateGrid2002-12km.csv** – Contains a list of states and regions, with their associated X, Y ranges according to the 2002 cells text file for the 12 km grid.
- **StateGrid2002-36km.csv** – Contains a list of states and regions, with their associated X, Y ranges according to the 2002 cells text file for the 36 km grid.

ValidationDataFiles subfolder:

- **castnet_grid_o3_2001_12km.sas7bdat** – SAS dataset containing CASTNET ozone data for the 12 km grid for the year 2001. This file can be used in validation (Step 5).
- **castnet_grid_o3_2001_36km.sas7bdat** – SAS dataset containing CASTNET ozone data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).
- **castnet_grid_o3_2002_12km.sas7bdat** – SAS dataset containing CASTNET ozone data for the 12 km grid for the year 2002. This file can be used in validation (Step 5).
- **castnet_grid_o3_2002_36km.sas7bdat** – SAS dataset containing CASTNET ozone data for the 36 km grid for the year 2002. This file can be used in validation (Step 5).
- **krigepredozoneerror122001.sas7bdat**– SAS dataset containing kriged ozone data for the 12 km grid for the year 2001. This file can be used in validation (Step 5).
- **krigepredozoneerror122002.sas7bdat**– SAS dataset containing kriged ozone data for the 12 km grid for the year 2002. This file can be used in validation (Step 5).
- **krigepredozoneerror362001.sas7bdat**– SAS dataset containing kriged ozone data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).
- **krigepredozoneerror362002.sas7bdat**– SAS dataset containing kriged ozone data for the 36 km grid for the year 2002. This file can be used in validation (Step 5).
- **pm25_improve_validation_2001_36.sas7bdat** – SAS dataset containing IMPROVE $PM_{2.5}$ data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).
- **pm25_krige_predict_error_2001_36.sas7bdat** – SAS dataset containing kriged $PM_{2.5}$ data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).
- **pm25_stn_validation_2001_36.sas7bdat** – SAS dataset containing STN $PM_{2.5}$ data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).
- **validation_pm25_2001_36.sas7bdat** – SAS dataset containing IMPROVE and STN $PM_{2.5}$ data for the 36 km grid for the year 2001. This file can be used in validation (Step 5).

2.4 LOGICAL DESIGN

Each subsection lists the object-oriented programming (OOP) classes defined in each project. A description of each is provided, as well as the objects that use the class. See Figures 1a and 1b for a graphical representation of the logical structure of the software.

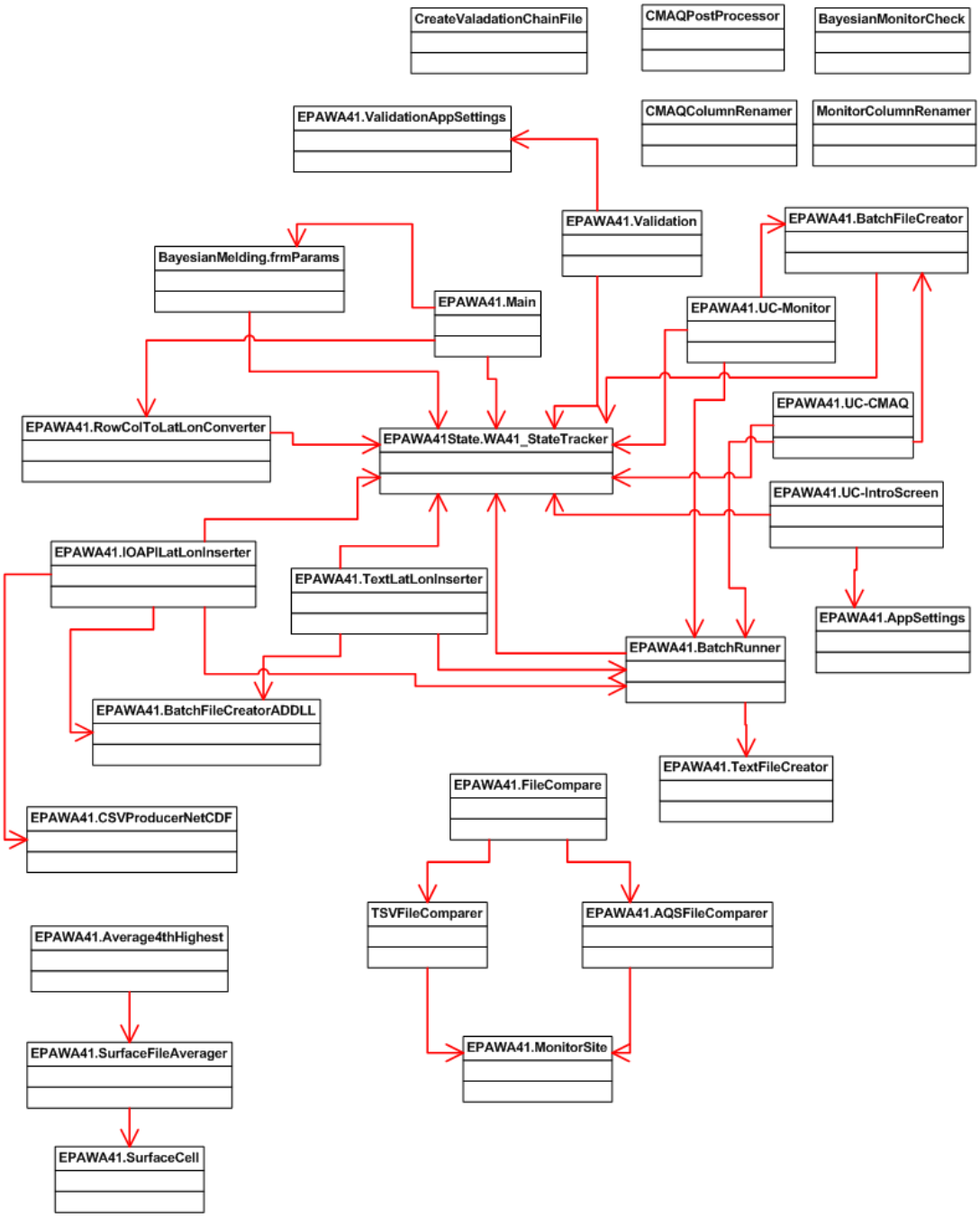


Figure 1a. Class Usage Diagram (Class A ---uses---> Class B).

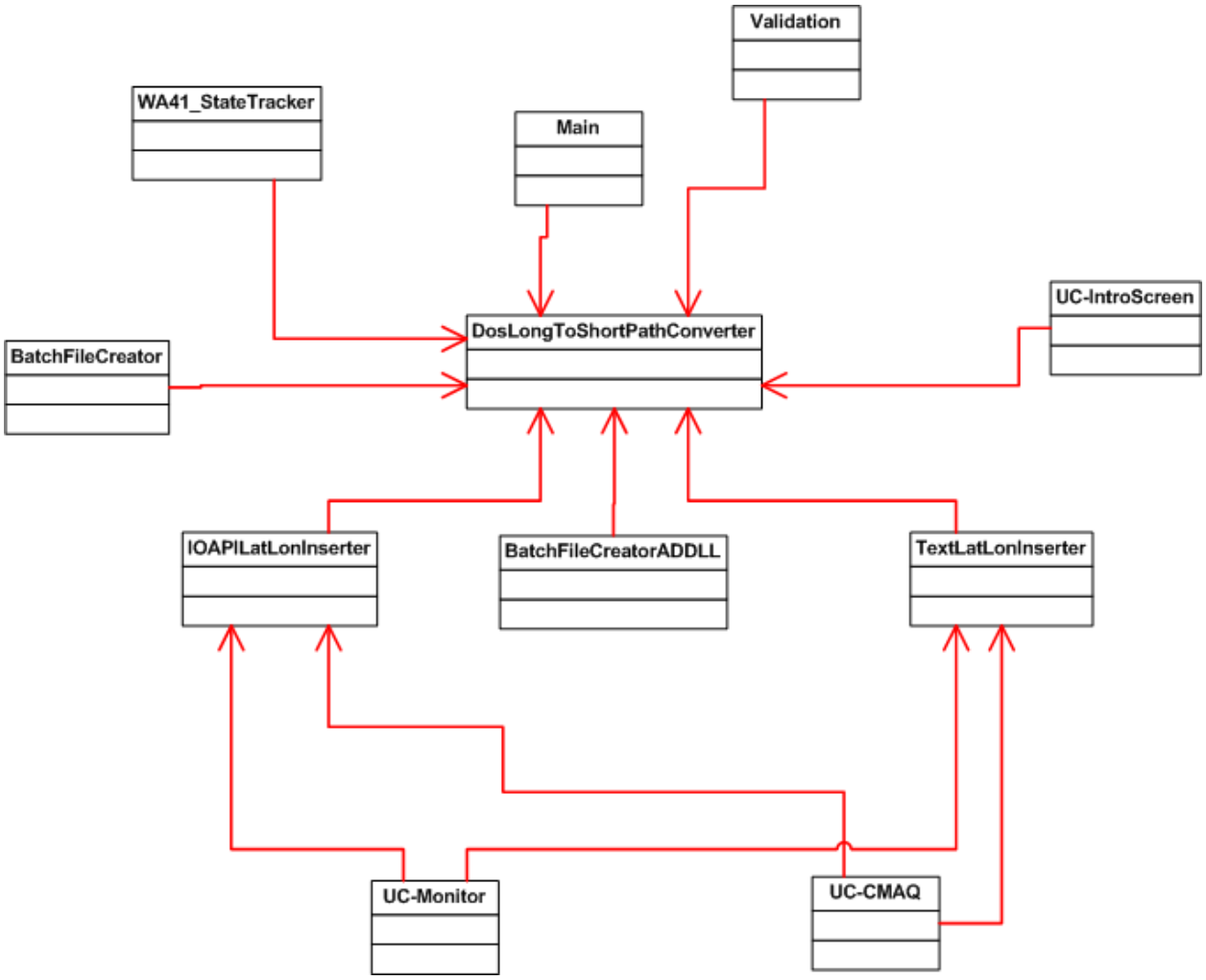


Figure 1b. Class Usage Diagram (Class A ---uses---> Class B) (Continued).

2.4.1 BayesianMelding Classes

- **frmParams** – The form as well as all logic needed to verify input, load parameter files and save parameter files.
    - Used by EPAWA41.Main
- **StringTokenizer** – Parses (splits) strings into individual tokens based on a specified delimiter.
    - Used by **BayesianMelding.frmParams**

2.4.2 BayesianMonitorCheck Classes

- **Checker** – Performs the post processing on the monitor $PM_{2.5}$ file created in Step 2
    - Used by **BayesianMonitorCheck.Main**

2.4.3 CMAQColumnRenamer Classes

- **CMAQColumnRenamer –** Performs the post processing of the CMAQ/CAMx files created in Step 2.
    - Used by **CMAQColumnRenamer.Main**

2.4.4 CMAQPostProcessor Classes

- **CMAQPostProcessor –** Performs the post processing of the CMAQ/CAMx $PM_{2.5}$ file created in Step 2.
    - Used by **CMAQPostProcessor.Main**

2.4.5 CreateValidationChainFile Classes

- **ValidationChainFileCreator –** Performs the creation of the validation chain file.
    - Used by **ValidationChainFileCreator.Main**

2.4.6 EPAWA41State Classes:

- **WA41_StateTracker –** This is the repository for values that are needed throughout program execution.
    - Used by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.BatchFileProcessors.BatchFileCreator**
        - **EPAWA41.BatchFileProcessors.BatchRunner**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.RowColToLatLongConverter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Main**
        - **EPAWA41.Forms.Validation**
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-IntroScreen**
        - **EPAWA41.UserControls.UC-Monitor**

- **DosLongToShortPathConverter –** Contains two static methods allowing the application to obtain either the DOS short name (8.3, no spaces) or the Windows long name of a file from the other one.
    - Used by:
        - **EPAWA41.BatchFileProcessors.BatchFileCreator**
        - **EPAWA41.BatchFileProcessors.BatchFileCreatorAddLL**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Main**
        - **EPAWA41.Forms.Validation**
        - **EPAWA41.UserControls.UC-IntroScreen**
        - **WA41State.WA41_StateTracker**

2.4.7 MonitorColumnRenamer

- **MonitorColumnRenamer** – Performs the post processing of the monitor files created in Step 2.
    - Used by **MonitorColumnRenamer.Main**

2.4.8 Bayesian Air Pollution Model Classes

None.

2.4.9 EPAWA41

- BatchFileProcessors
    - **BatchFileCreator** – Creates batch files used in Step 2.
        - Used by:
            - **EPAWA41.UserControls.UC-CMAQ**
            - **EPAWA41.UserControls.UC-Monitor**
    - **BatchFileCreatorAddLL** – Creates batch file to run addll.exe.
        - Used by:
            - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
            - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
    - **BatchRunner** – Executes batch files created by **EPAWA41**. **BatchFileProcessors.BatchFileCreator**.
        - Used by:
            - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
            - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
            - **EPAWA41.UserControls.UC-CMAQ**
            - **EPAWA41.UserControls.UC-Monitor**

- CSVFileProcessors
    - **CSVProducer** – Base class to provide functionality to other classes.
        - Inherited by:
            - **EPAWA41.BatchFileProcessors.BatchFileCreator**
            - **EPAWA41.BatchFileProcessors.BatchFileCreatorAddLL**
    - **CSVProducerNetCDF** – Provides an interface to the NetCDF library.
        - Used by:
            - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
    - **IOAPILatLonInserter** – Inserts latitude and longitude into the CMAQ/CAMx and monitor files created in Step 2.
        - Used by:
            - **EPAWA41.UserControls.UC-CMAQ**
            - **EPAWA41.UserControls.UC-Monitor**
    - **RowColToLatLongConverter** – Inserts latitude and longitude values into the surface file created by the model
        - Used by: **EPAWA41.Forms.Main**
    - **TextLatLonInserter** – Inserts latitude and longitude values into the CMAQ/CAMx and monitor files created in Step 2.
        - Used by:
            - **EPAWA41.UserControls.UC-CMAQ**

            - **EPAWA41.UserControls.UC-Monitor**
- Tools
    - **AQSFileComparer** – Compares two AQS monitor files
        - Used by: **EPAWA41.Forms.FileCompare**
    - **FileComparer** – Base class to provide functionality to other classes.
        - Inherited by:
            - **EPAWA41.Tools.AQSFileComparer**
            - **EPAWA41.Tools.TSVFileComparer**
    - **MonitorSite** – Object representing a monitor site
        - Used by:
            - **EPAWA41.Tools.AQSFileComparer**
            - **EPAWA41.Tools.FileComparer**
            - **EPAWA41.Tools.TSVFileComparer**
    - **SurfaceCell** - Object representing a cell from surface file
        - Used by: **EPAWA41.Tools.SurfaceFileAverager**
    - **SurfaceFileAverager** – Averages 3 consecutive (annual) 4th Highest surface files
        - Used by: **EPAWA41.Forms.Average4thHighest**
    - **TSVFileComparer** – Compares two air pollution monitor sites tsv files
        - Used by: **EPAWA41.Forms.FileCompare**
- Utility
    - **AppSettings** – Saves application settings for use in future program sessions
        - Used by: **EPAWA41.UserControls.UC-IntroScreen**
    - **TextFileCreator** – Creates text files that capture execution messages from batch files
        - Used by:
            - **EPAWA41.BatchFileProcessors.BatchRunner**
            - **EPAWA41.Forms.Main**
    - **ValidationAppSettings** – Saves validation settings for use in future program sessions
        - Used by: **EPAWA41.Forms.Validation**

2.5 DATA

Each subsection lists a category of data used by the application. A description of each data file is provided, as well as the listing of the files that fall into each category.

2.5.1 Permanent Storage

Each subsection lists a category of permanent data files. These are the data files that persist when the application is closed.

2.5.1.1 Input Files

Each subsection lists a data file that must exist before a simulation is started. These do not include the previously described data files that are installed to the application folder. These data files are not part of the application's installation but are obtained otherwise.

A description of each data file is provided, as well as the names and the representative data files.

2.5.1.1.1 CMAQ/CAMx File

These files are in NetCDF format, also referred to as IOAPI files. These files are specific to each air pollutant, year, and grid size. The air pollutant (ozone/O3 or PM/PM25) and grid size (12/12km or 36/36km) are indicated by the file name. These files are generated and supplied by EPA. These files are input for the CMAQ/CAMx (i.e., air quality model) portion of Step 2. This file is also necessary for Step 4 to obtain latitude and longitude information for the air pollution concentration surface file created by the model. This file is also required in the Monitor portion of Step 2 if the user decides to place IOAPI latitude and longitude information into the monitor file created there. Table 1 lists a few sample CMAQ files.

Table 1. Sample CMAQ Files.

CMAQ File Name	Type of Data	Size
ozone_2001_12.conc	Ozone, 2001, 12 km^2	1338.22 MB
2002ac_v4.61_L3b_EUS_12km.combine.hourly.O3.total365	Ozone, 2002, 12 km^2	2237.66 MB
2002ac_v4.61_L3b_us36b.combine.hourly.O3.total365	Ozone, 2002, 36 km^2	553.99 MB
pm25_2001_12km.ioapi	$PM_{2.5}$, 2001, 12 km^2	1338.22 MB
2002ac_v4.61_l3b_eus_12km.dailyavg.pm25	$PM_{2.5}$, 2002, 12 km^2	93.25 MB

2.5.1.1.2 Monitor File

These files are text files, delimited by the pipe character ('|'). These files are specific to air pollutant and year. A portion of a file is provided to understand the required format of the file. These files are available for download from the AQS download page at https://www3.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm. These files are input for the Monitor portion of Step 2. Table 2 lists the names of several sample monitor files.

Table 2. Sample Monitor Files.

Monitor File Name	Pollutant	Size
RD_501_44201_2001-0.txt	Ozone	521.26 MB
RD_501_44201_2002-0.txt	Ozone	531.41 MB
RD_501_88101_2001-0.txt	$PM_{2.5}$	12.19 MB
rd_501_pm_fine_spec_2001-0.txt	$PM_{2.5}$	36.89 MB
RD_501_88101_2002-0.txt	$PM_{2.5}$	12.30 MB
rd_501_pm_fine_spec_2002-0.txt	$PM_{2.5}$	76.83 MB

2.5.1.1.3 Cells File

These files are text files, delimited by spaces. These files are specific to a year and locational grid, providing latitude and longitude values for each cell of the grid. These files are provided by EPA and are required in the CMAQ/CAMx and Monitor portions of Step 2, if the user decides to place the latitude and longitude information for each cell in the created files. Table 3 lists the names of several sample cells files.

Table 3. Sample Cell Files.

Cell File Name	Size
cells_12km_2001.txt	4.58 MB
cells_12km_2002.txt	7.66 MB
cells_36km_2001.txt	1.90 MB
cells_36km_2002.txt	1.90 MB

2.5.1.2 Output Files

Each subsection lists a file that is created during a simulation. These files are not deleted when the simulation is complete. They are available at the completion of the simulation. A description of each file is provided, as well as file sizes for a typical simulation.

2.5.1.2.1 CMAQ Model Input file

This is a CSV file created in the CMAQ/CAMx portion of Step 2, and is an input into the T-SpACE model. This file has 4 columns: Day, Column, Row, and CMAQ log (i.e., the log-transformed value of air quality model [CMAQ or CAMx] estimated concentration). The number of rows, and total size, is dependent upon the grid selections (T, X, Y) made in Step 1.

- T = time of simulation in calendar days (start, end)
- X = rows of grid cells (start, end)
- Y = columns of grid cells (start, end)

Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **ozone_2001_12.conc** CMAQ/CAMx input file; the resultant CMAQ/CAMx model input file is 245.00 MB in size.

Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **pm25_2001_12km.ioapi** CMAQ/CAMx input file; the resultant CMAQ/CAMx model input file is 399.65 MB in size.

2.5.1.2.2 Monitor Model Input File
This is a CSV file created in the monitor portion of Step 2, and is an input to the T-SpACE model. This file has 4 columns: Day, Column, Row, and monitor log (i.e., the log-transformed air pollution monitor concentration value). The number of rows, and total size, is dependent upon the grid selections (T, X, Y) made in Step 1.

Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_44201_2001-0.txt** monitor input file; the resultant monitor model input file is 3.66 MB in size.

Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_88101_2001-0.txt** monitor input file; the resultant monitor model input file is 5.35 MB in size.

2.5.1.2.3 CMAQ Model Input file with Latitude and Longitude
This is a CSV file that is optionally created in the CMAQ/CAMx portion of Step 2. This file is identical to the CMAQ Model Input File previously described, except that it has 2 additional columns: Longitude and Latitude.

Adding the latitude and longitude values to the ozone file example previously described (**ozone_2001_12.conc**) resulted in a file that is 509.12 MB in size.

Adding the latitude and longitude values to the $PM_{2.5}$ file example previously described (**pm25_2001_12km.ioapi**) resulted in a file that is 660.81 MB in size.

2.5.1.2.4 Monitor Model Input file with Latitude and Longitude
This is a CSV file that is optionally created in the monitor portion of Step 2. This file is identical to the Monitor Model Input File previously described, except that it has 2 additional columns: Longitude and Latitude.

Adding the latitude and longitude to the ozone file example previously described (**RD_501_44201_2001-0.txt**) resulted in a file that is 7.34 MB in size.

Adding the latitude and longitude to the $PM_{2.5}$ file example previously described (**RD_501_88101_2001-0.txt**) resulted in a file that is 11.82 MB in size.

2.5.1.2.5 Chain File
This is a CSV file that is created by the T-SpACE model in Step 4. This file contains the values for TauX, TauY, TauZ, RhoZ (i.e., the temporal autocorrelation parameter of the mean process) and each Mu (the mean level of the CAR [conditional autoregression] process) for each simulation step.

The chain file is helpful in understanding if the choices made by the user for the number of burn-in steps and number of simulations steps is appropriate. By plotting the Tau parameters, by each iteration, the user can see where the Tau parameter values are increasing or decreasing, indicating that those iterations should be included in the (number of) burn-in steps. If the user sees that the Tau parameters jump between a range (e.g., bouncing back and forth), this would indicate the point at which the simulation steps should begin, and where the number of burn-in steps should end. See Tables 4, 5, 6 and 7 for typical chain file sizes.

2.5.1.2.6 Global File

This is a CSV file that is created by the T-SpACE model in Step 4. This file contains a brief summary of the number of burn-in steps, simulation loop steps, and 'thinning' choices made for the simulation. Thinning is the process of saving a sample every given number of steps (n). It is used to reduce the self-correlation in the chain file. See Tables 4, 5, 6 and 7 for typical Global file sizes.

2.5.1.2.7 Simulation Data File

This is a machine readable file that is created by the T-SpACE model in Step 4. This file is for Aerosol Optical Depth (AOD) satellite input, which is implemented in the Model 6 version of T-SpACE. See Tables 4, 5, 6 and 7 for typical simulation data file sizes.

2.5.1.2.8 Summary File

This is a CSV file that is created by the T-SpACE model in Step 4. This file contains summary statistics for all calculated parameters of the model including the bias surface. Appendix G, the T-SpACE User's Guide, provides a detailed description of the contents of this file. See Tables 4, 5, 6 and 7 for typical summary file sizes.

2.5.1.2.9 Surface File

This is a CSV file that is created by the T-SpACE model in Step 4. This file is used by the validation process (Step 5). This file contains the estimated surface from the simulation. The T-SpACE User's Guide provides a detailed summary of the contents of this file. See Tables 4, 5, 6 and 7 for typical air pollution concentration surface file sizes.

Table 4. Sizes of Model Output Files Using Ozone Model Input Files for 10 Burn-in Steps and 50 Simulation Steps.

Model Output File	Size
Chain.csv	146 K
Globals.csv	< 1 K
Simulation.dat	1,334.50 MB
Summary.csv	1,642 MB
Surface.csv	1,479.30 MB

Table 5. Sizes of Model Output Files Using Ozone Model Input Files 100 Burn-in Steps and 2500 Simulation Steps.

Model Output File	Size
Chain.csv	6.97 MB
Globals.csv	< 1 K
Simulation.dat	1,334.50 MB
Summary.csv	1,664.34 MB
Surface.csv	1,106.42 MB

Table 6. Sizes of Model Output Files Using $PM_{2.5}$ Model Input Files Previously Described, 10 Burn-in Steps and 50 Simulation Steps.

Model Output File	Size
Chain.csv	146 K
Globals.csv	< 1 K
Simulation.dat	1,334.50 MB
Summary.csv	1,659.52 MB
Surface.csv	1,531.19 MB

Table 7. Sizes of Model Output Files Using $PM_{2.5}$ Model Input Files Previously Described, 100 Burn-in Steps and 2500 Simulation Steps.

Model Output File	Size
Chain.csv	6.97 MB
Globals.csv	< 1 K
Simulation.dat	1,334.50 MB
Summary.csv	1,681.85 MB
Surface.csv	1,145.24 MB

2.5.1.2.10 Validation Chain File

This is a CSV file that is created by a process that runs immediately after completion of the T-SpACE model run in Step 4. This file is used by the validation process (Step 5). This file is a subset of the chain file previously described, containing only the values for TauX, TauY, and TauZ. The size of a validation chain file is typically 10% of the size of the parent chain file from which it was created.

2.5.1.2.11 Validation Report File

This is a text file that is created by running the validation process (Step 5). This file contains the reports generated by the SAS validation program. The size of this file is approximately 20 K.

2.5.1.2.12 Validation Log File

This is a text file that is created by running the validation process (Step 5). This file contains the logs generated by the SAS validation program. The size of this file is approximately 90 K.

2.5.1.2.13 Validation SAS Program

This is a text file that is created by running the validation process (Step 5). This file contains the SAS source code generated and executed by the T-SpACE model. The size of this file is 16 K.

2.5.1.3 Intermediate Files

Each subsection lists a file that is created during a T-SpACE simulation run. These files are deleted automatically by the system when no longer needed. A description of each file is provided, as well as file sizes for a typical process.

2.5.1.3.1 CMAQ Ozone Column Rename Temp File

This CSV file is a duplicate of the CMAQ Ozone Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that changes the columns names in the CMAQ Ozone Model Input File. This temp file, which is the same size as the CMAQ Ozone Model Input File, is deleted when the file with the desired column names is generated.

2.5.1.3.2 Monitor Ozone Column Rename Temp File

This CSV file is a duplicate of the Monitor Ozone Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that changes the column names in the Monitor Ozone Model Input File. This temp file, which is the same size as the Monitor Ozone Model Input File, is deleted when the file with the desired column names is generated.

2.5.1.3.3 CMAQ $PM_{2.5}$ Column Post-Process Temp File

This CSV file is a duplicate of the CMAQ $PM_{2.5}$ Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that performs post-processing on the CMAQ $PM_{2.5}$ Model Input File. This temp file, which is the same size as the CMAQ $PM_{2.5}$ Model Input File, is deleted when the post-processed file is written.

2.5.1.3.4 CMAQ $PM_{2.5}$ Column Rename Temp File

This CSV file is a duplicate of the CMAQ $PM_{2.5}$ Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that changes the columns names in the CMAQ $PM_{2.5}$ Model Input File. This temp file, which is the same size as the CMAQ $PM_{2.5}$ Model Input File, is deleted when the file with the desired column names is generated.

2.5.1.3.5 Monitor $PM_{2.5}$ Column Post-Process Temp File

This CSV file is a duplicate of the Monitor $PM_{2.5}$ Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that performs post-processing on the Monitor $PM_{2.5}$ Model Input File. This temp file, which is the same size as the Monitor $PM_{2.5}$ Model Input File, is deleted when the post-processed file is generated.

2.5.1.3.6 Monitor $PM_{2.5}$ Column Rename Temp File

This CSV file is a duplicate of the Monitor $PM_{2.5}$ Model Input File created in Step 2. It is used by the process that runs automatically in Step 2 that changes the columns names in the Monitor $PM_{2.5}$ Model Input File. This temp file, which is the same size as the Monitor $PM_{2.5}$ Model Input File, is deleted when the file with the desired column names is generated.

2.5.1.3.7 CMAQ Reformatted File for ADDLL Temp File

This CSV file is created when, in Step 2, the user chooses to have latitude and longitude information added to a CMAQ Model Input File. This file contains all of the data from CMAQ Model Input File created in Step 2, reformatted to match what ADDLL.exe expects as input. This file is deleted once **ADDLL.exe** completes execution. The file is generated as an ***intermediate data file***, which allows ***air quality model concentration surfaces*** to be generated for individual states and regions in the US. With ozone, the size of this file is approximately 150% of the size of the CMAQ Model Input File from which it is created. With $PM_{2.5}$, the size of this file is approximately 200% of the size of the CMAQ Model Input File from which it is created.

2.5.1.3.8 CMAQ Reformatted File with Lat/Lon Temp File

This CSV file is created when, in Step 2, the user chooses to have latitude and longitude information added to a CMAQ Model Input File. This file is created by **ADDLL.exe** and contains all of the data from the temporary file previously described, plus latitude and longitude for each record. Another process reformats this file into the CMAQ Model Input file with Latitude and Longitude previously described. The file is generated as an ***input data file***, which allows ***air quality model concentration surfaces*** to be generated for individual states and regions in the US. Once the file reformatting is complete, the file is deleted. With ozone, the size of this file is approximately 175% of the size of the temporary file from which it is created. With $PM_{2.5}$, the size of this file is approximately 150% of the size of the temporary file from which it is created.

2.5.1.3.9 Monitor Reformatted File for ADDLL Temp File

This CSV file is created when, in Step 2, the user chooses to have latitude and longitude information added to a Monitor Model Input File. This file contains all of the data from Monitor Model Input File created in Step 2, reformatted to match what **ADDLL.exe** expects as input. This file is deleted once **ADDLL.exe** completes execution. The file is generated as an ***intermediate data file***, which allows ***air quality monitor concentration surfaces*** to be generated for individual states and regions in the US. With ozone, the size of this file is approximately 150% of the size of the Monitor Model Input File from which it is created. With $PM_{2.5}$, the size of this file is approximately 150% of the size of the Monitor Model Input File from which it is created.

2.5.1.3.10 Monitor Reformatted File with Lat/Lon Temp File

This CSV file is created when, in Step 2, the user chooses to have latitude and longitude information added to a Monitor Model Input File. This file is created by ADDLL.exe and contains all of the data from the temporary file previously described, plus latitude and longitude for each record. Another process reformats this file into the Monitor Model Input file with Latitude and Longitude previously described. Once that reformat is complete, this file is deleted. The file is generated as an ***input data file***, which allows ***air quality monitor concentration surfaces*** to be generated for individual states and regions in the US. With ozone, the size of this file is approximately 200% of the size of the temporary file from which it is created. With $PM_{2.5}$, the size of this file is approximately 225% of the size of the temporary file from which it is created.

2.5.1.3.11 Raw Surface Temp File

This CSV file is created by the T-SpACE model in Step 4. It is used by the process that runs automatically when the T-SpACE model simulation completes which creates the final air pollution concentration surface file. This process takes the raw surface temp file, adds latitude and longitude to each record, and converts the date on each record to a recognizable format. This file is deleted once the post-processing successfully completes. The file allows ***final (monitor plus model) concentration surfaces*** to be generated for individual states and regions in the US. The size of this file is approximately 75% of the size of the final surface file based on this temporary file.

2.5.2 Volatile Storage

This subsection lists the properties contained in the volatile storage (i.e., computer memory) used during a simulation. This storage disappears when the application is closed. A description of each property is provided, as well as the classes that initialize the property and read the property. Volatile storage is provided by the ***state tracker*** (**Singleton)** class, which acts as a repository for values that are needed throughout the application. When the T-SpACE application starts, a single object of this class (***state tracker***) is instantiated (created/generated), and a reference ('pointer') to this object is passed to objects that need to initialize or retrieve values stored within them. No computer hard disk input/output (I/O) is utilized with the ***state tracker***.

The ***state tracker*** properties are:

- **ApplicationLocation** – Application executable's folder.
  - Initialized by:
    - **EPAWA41.Forms.Main**
  - Read by:
    - **EPAWA41.BatchFileProcessors.BatchFileCreator**
    - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
    - **EPAWA41.CSVFileProcessors.RowColToLatLonConv**
    - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
    - **EPAWA41.Forms.Validation**
    - **EPAWA41.UserControls.UC-CMAQ**
    - **EPAWA41.UserControls.UC-Monitor**
- **BayesianMeldingStartDate** – Start date in format needed by **Bayesian Melding**.
  - Initialized by:
    - **EPAWA41.UserControls.UC-IntoScreen**
  - Read by:
    - **BayesianMelding.frmParams**
- **BayesianDay** – Day of BayesianMeldingStartDate.
  - Initialized by:
    - **BayesianMelding.frmParams**
  - Read by:
    - **BayesianMelding.frmParams**
- **BayesianMonth** – Month of BayesianMeldingStartDate.
  - Initialized by:
    - **BayesianMelding.frmParams**
  - Read by:

        - **BayesianMelding.frmParams**
- **BayesianYear** – Year of BayesianMeldingStartDate.
    - Initialized by:
        - **BayesianMelding.frmParams**
    - Read by:
        - **BayesianMelding.frmParams**
- **CMAQPath** – Location of the CMAQ/CAMx file created in Step 2.
    - Initialized by:
        - **EPAWA41.UserControls.UC-CMAQ**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.BatchFileProcessors.BatchFileCreator**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
- **CSVFileType** – Denotes if the current file in Step 2 is CMAQ/CAMx or monitor.
    - Initialized by:
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
    - Read by:
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
- **EndDateBatchFormat** – End date in format needed by Step 2 batch files.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
- **EndTime** – Ending T value, in date format, as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.Forms.Validation**
- **FullPathOfSurfaceCSV** – Location of air pollution concentration surface file created by the T-SpACE model.
    - Initialized by:
        - **BayesianMelding.frmParams**
    - Read by:
        - **EPAWA41.CSVFileProcessors.RowColToLatLonConv**
        - **EPAWA41.Forms.Validation**
- **InputDirectory** – Directory specified by user in Step 1, where the files created in Step 2 will be saved
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**

        - **EPAWA41.BatchFileProcessors.BatchFileCreator**
        - **EPAWA41.BatchFileProcessors.BatchRunner**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.RowColToLatLonConv**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Main**
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
- **IOAPIFullFilePath –** Location of the CMAQ input file specified in Step 2.
    - Initialized by:
        - **EPAWA41.UserControls.UC-CMAQ**
    - Read by:
        - **EPAWA41.BatchFileProcessors.BatchFileCreatorAddLL**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.RowColToLatLonConv**
        - **EPAWA41.Forms.Main**
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
- **Is64BitVersion –** Boolean used to determine if application is running on a 64-bit machine.
    - Initialized by:
        - **EPAWA41.Forms.Main**
    - Read by:
        - **EPAWA41.Forms.Main**
- **LatLonCMAQPath –** Location of CMAQ/CAMx IOAPI file used as source of latitude and longitude information for CMAQ/CAMx CSV file created in Step 2.
    - Initialized by:
        - **EPAWA41.UserControls.UC-CMAQ**
    - Read by:
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
- **LatLonMonitorPath –** Location of CMAQ/CAMx IOAPI file used as source of latitude and longitude information for monitor CSV file created in Step 2.
    - Initialized by:
        - **EPAWA41.UserControls.UC-Monitor**
    - Read by:
        - **EPAWA41.UserControls.UC-Monitor**
- **MonitorPath –** Location of the monitor file created in Step 2.
    - Initialized by:
        - **EPAWA41.UserControls.UC-Monitor**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.BatchFileProcessors.BatchFileCreator**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
- **ParFilePath –** Location of model parameter file created in Step 3.

    - Initialized by:
        - **BayesianMelding.frmParams**
    - Read by:
        - **EPAWA41.Forms.Main**
- **PartialPathOfSurfaceCSVDirectory –** Folder of surface file created by the T-SpACE model.
    - Initialized by:
        - **BayesianMelding.frmParams**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **PartialPathOfSurfaceCSVFile –** Name of surface file created by the T-SpACE model.
    - Initialized by:
        - **BayesianMelding.frmParams**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **Pollutant –** Pollutant for the current T-SpACE simulation run.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
        - **EPAWA41.Forms.Main**
    - Read by:
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
        - **EPAWA41.Forms.Validation**
- **StartDateBatchFormat –** Start date in format needed by Step 2 batch files.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **EPAWA41.UserControls.UC-CMAQ**
        - **EPAWA41.UserControls.UC-Monitor**
- **StartTime –** Beginning T value, in the appropriate date format, as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.Forms.Validation**
- **TBegin –** Beginning T value as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
- **TEnd –** Ending T value as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
- **ValidationChainFile –** Chain file specified in Step 5.

    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **ValidationKrigeFile** – Krige file specified in Step 5.
    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **ValidationMonitorFile** – Monitor file specified in Step 5.
    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **ValidationReportFile** – Report file specified in Step 5.
    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **ValidationReportTitle** – Report title specified in Step 5.
    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **ValidationSurfaceFile** – Surface file specified in Step 5.
    - Initialized by:
        - **EPAWA41.Forms.Validation**
    - Read by:
        - **EPAWA41.Forms.Validation**
- **XBegin** – Beginning X value as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Validation**
- **XEnd** – Ending X value as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Validation**
- **YBegin** – Beginning Y value as specified in Step 1.

    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Validation**
- **YEnd** – Ending Y value as specified in Step 1.
    - Initialized by:
        - **EPAWA41.UserControls.UC-IntoScreen**
    - Read by:
        - **BayesianMelding.frmParams**
        - **EPAWA41.CSVFileProcessors.IOAPILatLonInserter**
        - **EPAWA41.CSVFileProcessors.TextLatLonInserter**
        - **EPAWA41.Forms.Validation**

2.6 COMMUNICATIONS/MESSAGING

Messaging within the T-SpACE application is minimal, as most of the components reside within the same logical structure or process. There are several instances where the application must spawn (generate) a process so that it can run a user interface console application. This is done using the Process class of the Microsoft.NET framework. Communication to this class is done using 2 methods of that class, Start and WaitForExit. The software spawns a process by invoking the start method. Invoking the '**WaitForExit**' method causes the software to pause until the process has completed.

The **Process** object is used by:

- **EPAWA41.BatchFileProcessors.BatchFileCreator**
    - Run the batch file to create the CMAQ/CAMx and monitor files (ozone and $PM_{2.5}$) in Step 2.
    - Run the batch file to add latitude and longitude to CMAQ/CAMx and monitor files (ozone and $PM_{2.5}$) created in Step 2.
    - Run the batch file to add latitude and longitude to the air pollution concentration surface file created by the model.
- **EPAWA41.Forms.Main**
    - Run the model
    - Run the process to create the validation chain file

3. OPERATIONAL DESIGN

Each subsection details a facet of the T-SpACE operational design. These facets include Use Cases and Typical Processing.

A use case is a description of a system's behavior as it responds to a request that originates from outside of that system.

Typical processing describes the time and memory constraints of the software during a typical process.

3.1 USE CASES

The subsection heading is the name of the use case. A brief description of each use case is provided.

3.1.1 Prepare to Create Simulation Input Use Case

Preparation for creating the T-SpACE simulation input is complete when the user selects the temporal and spatial grids for the simulation, as well as the destination folder for the simulation input.

3.1.2 Create Simulation Input Use Case

Simulation input is created when the user selects a pollutant and creates the CMAQ/CAMx and monitor simulation input files.

3.1.3 Choose Simulation Parameters Use Case

Choosing simulation parameters is complete when the user saves the simulation parameter (*.par) file.

3.1.4 Run Simulation Use Case

Running the simulation is completed when the user launches the T-SpACE model using a simulation parameter file.

3.1.5 Validate Simulation Use Case

Validating a completed simulation is complete when the user launches validation using the T-SpACE simulation output files, monitor data, and kriging data.

3.1.6 Compare AQS Files Use Case

Differences in monitor locations between two AQS data files can be obtained when the user utilizes the AQS File Compare tool.

3.1.7 Compare Airsites Files Use Case

Differences in monitor locations between two air sites data files can be obtained when the user utilizes the Airsites File Compare tool.

3.1.8 Average 4th Highest Surface Files Use Case

A file containing the average of three consecutive annual 4th Highest Surface Files is created when the user launches the Average 4th Highest Surface Files tool.

3.1.9 Create States Grid File Use Case

A file containing the states and regions, along with their corresponding grid cells, is created when the user launches the Apply States to Grid tool.

3.2 TYPICAL PROCESSING

Each subsection details one aspect of a typical processing for the application. These aspects are Sequence and Expected System Behavior.

Sequence lists the steps taken during a typical processing.

Expected system behavior details the software's time and memory constraints of a typical processing.

3.2.1 Sequence

Once the T-SpACE is launched, the sequence for running a complete T-SpACE simulation is as follows:

1. Using the calendars, choose the beginning and end dates for the simulation.
2. Select a folder where the CSV files (created in Step 2) will be saved.
3. Select the X and Y ranges for the simulation. These can either be entered manually, or a state/region can be selected from a list.
4. Select a pollutant for the simulation by clicking "Ozone" or "PM 2.5". The application automatically advances to Step 2.
5. Select a CMAQ/CAMx input file.
6. Optionally change the default output file name.
7. Optionally create an additional version of the output file that also contains latitude and longitude. A source of this information must be provided, either the file selected in sequence 5 above, or from a cells.txt file.
8. Click "Execute"; wait for "Complete" message.
9. Still in Step 2, click the "Monitor" tab.
10. Select a Monitor input file.
11. Optionally change the default output file name.

12. Optionally create an additional version of the output file that also contains latitude and longitude. A source of this information must be provided, either the file selected in sequence 5 above, or a cells file.
13. Click “Execute”; wait for “Complete” message.
14. Click “Step 3: Model Specification” on the tool bar above the Step 2 form.
15. Optionally open an existing parameter file by clicking the open button (the one with folder opening) and selecting the file
16. Review the values on the “Grid” tab, change as desired.
17. Click the first “Priors” tab and review/change the values.
18. Click the second “Priors” tab and review/change the values.
19. Click the “Data” tab and review/change the values.
20. Click the “Boundaries” tab and review/change the values.
21. Click the “Track” tab and review/change the values.
22. Click the “Simulation” tab and review/change the values.
23. Click the “4th Highest” tab and review/change the values.
24. Click the save button to save the parameter (*.par) file to the computer hard disk.
25. Click “Step 4: Launch Model” on the tool bar above the Step 3 form; wait for “Complete” message.
26. Click “Step 5: Launch Validation” on the tool bar above the Step 3 form
27. Review/change the values
28. Click “Run Validation”; wait for “Complete” message.

3.3 EXPECTED SYSTEM BEHAVIOR

The computer that was used to determine the expected behavior of T-SpACE was configured as follows:

- Intel Core 2 Duo E6750 Processor, @ 2.66 GHz
- 3.48 GB of RAM, @ 2.66 GHz
- Windows XP Professional, Service Pack 2

Each subsection details one of the software’s constraints during a typical processing. These constraints are time and memory.

The simulation parameter files used for these runs are provided in Appendix E.

3.3.1 Approximate Time

The subsection name is the process being timed. A description of the process is provided, as well as a breakdown of the time needed to complete the process.

3.3.1.1 *Creating the CMAQ Model Input File (Step 2)*
Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **ozone_2001_12.conc** CMAQ/CAMx input file the time needed was as follows (to the nearest second):

- Creating the CMAQ file – 2 minutes 44 seconds
- Renaming the columns – 12 seconds

- Adding latitude and longitude (optional) – 8 minutes, 20 seconds

Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **pm25_2001_12km.ioapi** CMAQ/CAMx input file the time needed was as follows (to the nearest second):

- Creating the CMAQ file – 2 minutes 49 seconds
- Post-processing – 27 seconds
- Renaming the columns – 14 seconds
- Adding latitude and longitude (optional) – 8 minutes, 23 seconds

3.3.1.2 *Creating the Monitor Model Input File (Step 2)*
Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_44201_2001-0.txt** monitor input file the time needed was as follows (to the nearest second):

- Creating the Monitor file – 2 minutes 44 seconds
- Renaming the columns – 1 seconds
- Adding latitude and longitude (optional) – 23 seconds

Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_88101_2001-0.txt** monitor input file the time needed was as follows (to the nearest second):

- Creating the Monitor file – 1 minutes 41 seconds
- Post-processing – 1 second
- Renaming the columns – 1 second
- Adding latitude and longitude (optional) – 8 seconds

3.3.1.3 *Running the Model (Step 4)*
Using the **<u>ozone files</u>** created as described above, the time needed to run the T-SpACE model was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps**:
    - Read input files – 35 seconds
    - Allocate variables – 1 minute, 15 seconds
    - Burn-in steps – 3 minutes, 27 seconds
    - Allocate variables – 55 seconds
    - Running T-SpACE simulation steps – **17 minutes, 8 seconds**
    - Write files – 8 minutes, 35 seconds
    - Add latitude and longitude – 7 minutes, 22 seconds
    - Write File – 1 minute, 24 seconds
    - Create validation chain file – 25 seconds
- **100 burn-in steps and 2500 simulation steps**:
    - Read input files – 53 seconds
    - Allocate variables – 1 minute, 15 seconds
    - Burn-in steps – 3 minutes, 26 seconds
    - Allocate variables – 1 minute, 16 seconds
    - Running T-SpACE simulation steps – **14 Hours, 21 minutes, 17 seconds**
    - Write files – 9 minutes, 52 seconds

- Add latitude and longitude – 7 minutes, 49 seconds
- Write File – 1 minute, 31 seconds
- Create validation chain file – 34 seconds

Using the **$PM_{2.5}$ files** created as described above, the time needed to run the T-SpACE model was as follows (to the nearest second):

- **10 burn-in steps and 50 simulation steps**:
  - Read input files – 41 seconds
  - Allocate variables – 11 seconds
  - Burn-in steps – 3 minutes, 27 seconds
  - Allocate variables – 13 seconds
  - Running T-SpACE simulation steps – **17 minutes, 7 seconds**
  - Write files – 8 minutes, 9 seconds
  - Add latitude and longitude – 7 minutes, 45 seconds
  - Write File – 29 seconds
  - Create validation chain file – 11 seconds
- **100 burn-in steps and 2500 simulation steps**:
  - Read input files – 41 seconds
  - Allocate variables – 17 seconds
  - Burn-in steps – 3 minutes, 25 seconds
  - Allocate variables – 24 seconds
  - Running T-SpACE simulation steps – **14 Hours, 16 minutes, 50 seconds**
  - Write files – 9 minutes, 34 seconds
  - Add latitude and longitude – 7 minutes, 52 seconds
  - Write File – 32 seconds
  - Create validation chain file – 13 seconds

3.3.1.4 *Validation (Step 5)*

The time needed to validate the model output for the ***ozone*** simulation previously described was as follows (to the nearest second):

- ***10 burn-in steps and 50 simulation steps*** – **14 seconds**
- ***100 burn-in steps and 2500 simulation steps*** – **5 minutes, 44 seconds**

The time needed to validate the model output for the ***$PM_{2.5}$*** simulation previously described was as follows (to the nearest second):

- ***10 burn-in steps and 50 simulation steps*** – **17 seconds**
- ***100 burn-in steps and 2500 simulation steps*** – **6 minutes, 58 seconds**

3.3.2 Memory Usage

The subsection name is the process being evaluated for memory usage. A description of the process is provided, as well as the maximum memory used during the process. When initially loaded, the T-SpACE application uses approximately **19,000 K** (kilobytes) of Random Access Memory (RAM, i.e., computer memory). This subsection only details processes which cause this consumption to increase.

3.3.2.1 *Creating the CMAQ Model Input File (Step 2)*

- Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **ozone_2001_12.conc** CMAQ/CAMx input file used 22,700 K of RAM.
- Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **pm25_2001_12km.ioapi** CMAQ/CAMx input file used 27,900 K of RAM.

3.3.2.2 *Creating the Monitor Model Input File (Step 2)*

- Running a 2001 simulation for ozone using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_44201_2001-0.txt** monitor input file used 23,900 K of RAM.
- Running a 2001 simulation for $PM_{2.5}$ using T = (1, 365), X (1, 213), Y (1, 188) and the **RD_501_88101_2001-0.txt** monitor input file used 27,900 K of RAM.

3.3.2.3 *Running the Model (Step 4)*

When running the simulation steps, the T-SpACE model used approximately **1,828,600 K** (***1.828 MB*** - megabytes) of RAM.

4. DESIGN ISSUES

Each subsection details a facet of T-SpACE's design issues.

4.1 FACTORS AFFECTING DESIGN

Each subsection describes a factor that had an impact on the T-SpACE design.

4.1.1 *Using existing applications for Steps 2 and 3.*

The existing EPA-developed console programs that were used by the T-SpACE model application were run by batch files, because a large number of environment variables required initialization. Processes to create and run these batch files, as well as to wait for their completion, had to be developed.

4.1.2 *Batch files used for Step 2 require short file names.*

The batch files needed to execute the programs for Step 2 require short DOS names for input files, output files, auxiliary files and application name. Routines to translate between long and short names were researched, developed, and implemented so that users had no restrictions on file naming.

4.1.3 *Not having the source code for Step 2 programs.*

The files created in Step 2 had to be altered, either by necessity (i.e., the model expects input values to be natural logarithm transformations of the air pollution concentration values) or by request (i.e., need different column names). Since the source code for the programs that created these files was not available, new routines were developed to read in the files and create new, slightly different, versions of them. These additional steps add to the execution time and significantly increase the amount of disk input/output.

4.2 SUGGESTED FUTURE UPGRADES

Each subsection describes a recommended future upgrade.

4.2.1 *Remove States/Regions outside of grid.*
All states are currently available for selection on the initial screen. When the states are defined by the locational grid, the closest grid cell to the lower left and upper right "corners" of the state are determined. There is always a closest cell, so every state has a corresponding grid region. Not all of the grids cover the continental United States so it is possible for a state's grid region to not actually contain any portion of the state.

4.2.2 *Creation of validation network and kriging files.*
Currently the network and kriging files used in the data validation process must be created outside of the application.

5. SUMMARY

Several different programming languages were used to create the source code for the T-SpACE model. Each computer programming language used in the development of T-SpACE has strengths and unique capabilities, which were leveraged to build a well-engineered software solution.

The current T-SpACE code base is as follows:

- The programs that create the four model input files in Step 2 (CMAQ/Monitor files for Ozone/$PM_{2.5}$) are standalone command-line applications written in FORTRAN (*.f files).
- The program that adds latitude and longitude information is a standalone command-line application written in FORTRAN (*.f files).
- The model is a standalone command-line application written in C++ (*.cpp).
- The Step 2 post-processing programs are standalone command-line applications written in C# (*.cs files).
- The program that creates the chain file for use in data validation is a standalone command-line applications written in C# (*.cs files).
- The program that allows for maintenance to the model simulation files (Step 3) is a standalone Windows-form application written in Visual Basic (*.vb).
- The state tracker object is written in C# (*.cs files).
- The validation program (Step 5) is written in SAS (*.SAS files).

All source code listed above, except for the FORTRAN, C++ and SAS programs, are accessible through the **EPAWA41.sln** file, which opens with Microsoft Visual Studio. Visual Studio allows for the editing, building and debugging of the source code.

Appendix B

Temporal-Spatial Ambient Concentration Estimator: Evaluation For Ozone Data

1.0 Background

To accomplish its mission to protect human health and the environment, EPA has established National Ambient Air Quality Standards (NAAQS) on six selected air pollutants known as criteria pollutants: ozone (O_3); carbon monoxide (CO); lead (Pb); nitrogen dioxide (NO_2); sulfur dioxide (SO_2), and; particulate matter (PM). The states are primarily responsible for maintaining and improving air quality and complying with the NAAQS. Ozone (O_3) and particulate matter (PM) are two of the criteria pollutants whose levels are regulated by NAAQS. Extensive air pollution monitoring networks have been set-up across the US to understand the levels (concentrations) of these air pollutants across the US and over time. Once collected, this information is reported to EPA and is made available publicly after the data is reviewed and analyzed for quality and accuracy.

When assessing air quality, the most direct way is to utilize unbiased ground-level air pollution measurements from the existing surface monitoring network stations located across the US. However, a good portion of the US, especially rural areas, have very sparse or irregularly spaced monitoring stations, resulting in large areas of the country with no information available on air quality. Air quality models estimate the spatial and temporal gradients of air pollution based on emissions inventories and meteorological information. These models, while providing estimates over large regions at relatively low cost, have (statistical) bias and have greater error (variance) than air pollution monitoring networks. These methods include Models-3/Community Multiscale Air Quality (CMAQ) numerical model and the Comprehensive Air Quality Model with Extensions (CAMx). Various statistical methods have been investigated that utilize existing monitoring data and spatial modeling techniques to develop a concentration surface that (retrospectively) estimates the distribution of daily air pollution levels within a specified region of the US. Kriging is one such method. Another technique is the T-SpACE model. The focus of this analysis is how T-SpACE compares to kriging and the CMAQ/CAMx model results for ozone data in the years of 2001 and 2002.

2.0 Model

The T-SpACE model, which uses the ozone monitoring data from EPA's Federal Reference Method (FRM) monitors in the NAMS/SLAMS network and CMAQ/CAMx model output, is used to estimate (retrospectively predict) pollutant levels at daily time scales. T-SpACE assumes that both monitoring data and CMAQ/CAMx data provide good information about the same underlying pollutant surface, but with different measurement error structures. It gives more weight to accurate monitoring data in areas where monitoring data exists, and relies on bias-adjusted CMAQ/CAMx output data in non-monitored areas. The problem is divided into hierarchical components, and each level in the hierarchy is modeled conditionally on its preceding levels. Appendix C provides a detailed description of the hierarchical Bayesian structure of the T-SpACE model.

3.0 Data

Table 1 lists the datasets utilized in the T-SpACE simulation and subsequent data validation.

Table 1. Ozone concentrations (ppm) - Daily 8-hour maximums.

Source	Dates	Resolution	Input file name	File Format
Community Multiscale Air Quality (CMAQ) model	January 1, 2001 – November 29, 2001	12 km	ozone.conc	NetCDF
		36 km	The appropriate format of this file was not available for analysis.	
	January 1, 2002 – November 30, 2002	12 km	2002ac_v4.61_L3b_EUS_12km.combine.hourly.O3.total365	NetCDF
		36 km	2002ac_v4.61_L3b_us36b.combine.hourly.O3.total365	NetCDF
EPA/AQS (FRM NAMS/SLAMS)	January 1, 2001 – December 31, 2001	NA	RD_501_44201_2001-0.txt	Text
	January 1, 2002 – December 31, 2002	NA	RD_501_44201-2002-0.txt	Text

3.1 Models-3/Community Multiscale Air Quality (CMAQ) and Comprehensive Air Quality Model with Extensions (CAMx)

3.1.1 CMAQ

The Models-3/Community Multiscale Air Quality (CMAQ) modeling system goals are to improve: **1)** the environmental management community's ability to evaluate the impact of air quality management practices for multiple pollutants at multiple scales, and; **2)** the scientist's ability to better probe, understand, and simulate chemical and physical interactions in the atmosphere. Traditionally, the CMAQ modeling system has been used to predict air quality across an entire regional or national domain, and then to simulate the effects of various changes in emission levels for policy-making purposes. However, for health studies, it is frequently applied to provide supplemental information on predicted air quality in areas where few or no monitors exist. It also accounts for meteorological conditions to better evaluate health outcomes. Detailed information on the CMAQ modeling system is available at https://www.epa.gov/cmaq and at https://www.cmascenter.org.

Because the CMAQ modeling system makes predictions across an entire two-dimensional domain, the set of model predictions at a given point in time resembles a smooth "surface" when it is graphically displayed, rather than simply a set of points at various locations. This surface, called a *response surface*, appears three-dimensional in that it covers the two-dimensional region and has a third dimension representing the magnitude of the predicted concentrations. The CMAQ model predictions are made relative to a specified rectangular *grid* that covers the entire

region of interest. The grid is composed of a matrix of rectangular *cells.* Within each cell, the CMAQ modeling system predicts average pollution levels at a given point in time. Each cell has equal area which is expressed as *h* x *v*, where *h* and *v* denote each cell's horizontal and vertical dimensions, respectively (in kilometers). Under the CMAQ modeling system, grid cells can either be 36 km x 36 km (which represents the "parent domain" covering the entire continental US) or 12 km x 12 km (which represents primarily the eastern and Midwest regions of the US up to 2007, and the entire US from 2008 and onward). Thus, as the values of *h* and *v* decrease for a fixed area, the CMAQ modeling system generates a greater density of predictions, which increases the resolution of the area's prediction surface.

3.1.2 CAMx

The Comprehensive Air Quality Model with Extensions (CAMx) is an Eulerian (gridded) photochemical dispersion model that allows for integrated "one-atmosphere" assessments of tropospheric air pollution (e.g., ozone, particulates, air toxics) over spatial scales ranging from neighborhoods (local) to continents (global). It is a modern, open-source system that is computationally efficient, flexible, and pubically available. CAMx's Fortran source code is modular and well-documented. The Fortran binary input/output file formats are based on the Urban Airshed Model (UAM) convention, and are compatible with many existing pre- and post-processing tools. Meteorological input fields are supplied to CAMx from separate weather prediction models. CAMx specifically supports the Weather Research and Forecasting (WRF) model, the Mesoscale Model 5 (MM5) and the Regional Atmospheric Modeling System (RAMS). All emission inputs are supplied from external pre-processing systems (e.g., the Sparse Matrix Operator Kernel (SMOKE) system and the Emissions Processor System Version 3 (EPS3).

CAMx simulates the emission, dispersion, chemical reaction, and removal of pollutants by 'marching' the Eulerian continuity equation forward in time (t) for each chemical species (l) on a system of nested three-dimensional grids. The continuity equation specifically describes the time dependency of volume-average species concentration within each grid cell as a sum of all physical and chemical processes operating on that volume. CAMx can perform simulations on four types of Cartesian map projections: Lambert Conic Conformal, Polar Stereographic, Mercator, and Universal Transverse Mercator (UTM). CAMx also offers the option of operating on a geodetic latitude/longitude grid system. The vertical grid structure is defined externally, so layer interface heights may be specified as any arbitrary function of space and/or time. This flexibility in defining the horizontal and vertical grid structures allows CAMx to be configured to match the grid of any meteorological model that is used to provide environmental input fields. Detailed information on the CAMx system can be found at: http://www.camx.com.

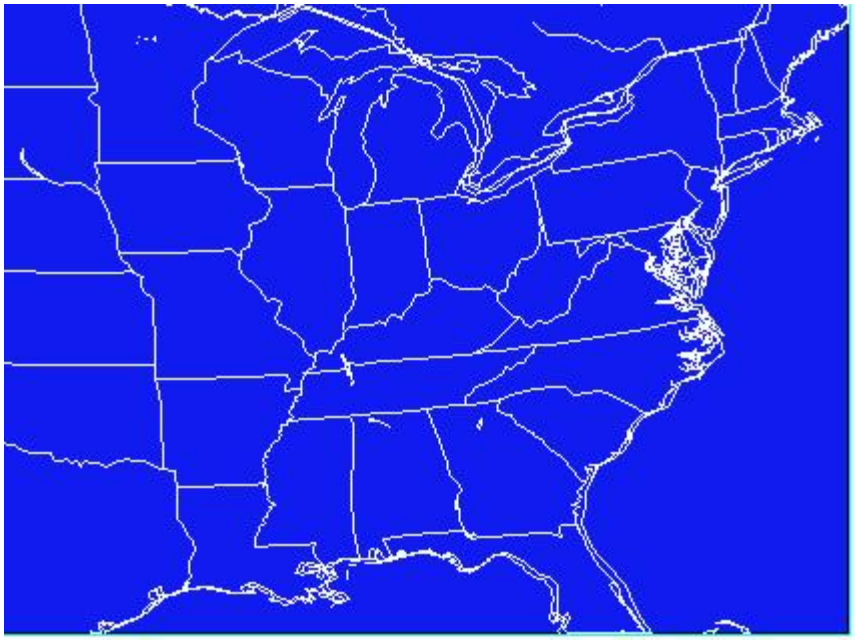


Figure 1. **The maximum area of the United States covered by the 2001 CMAQ 12 km x 12 km Grid.**

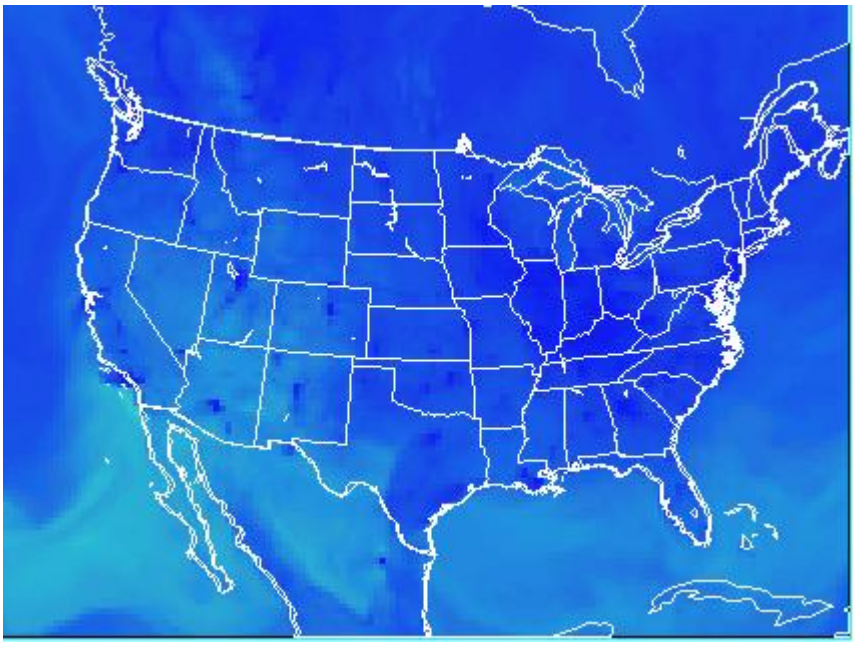


Figure 2. **The maximum area of the United States covered by the 2001 CMAQ 36 km x 36 km grid.**

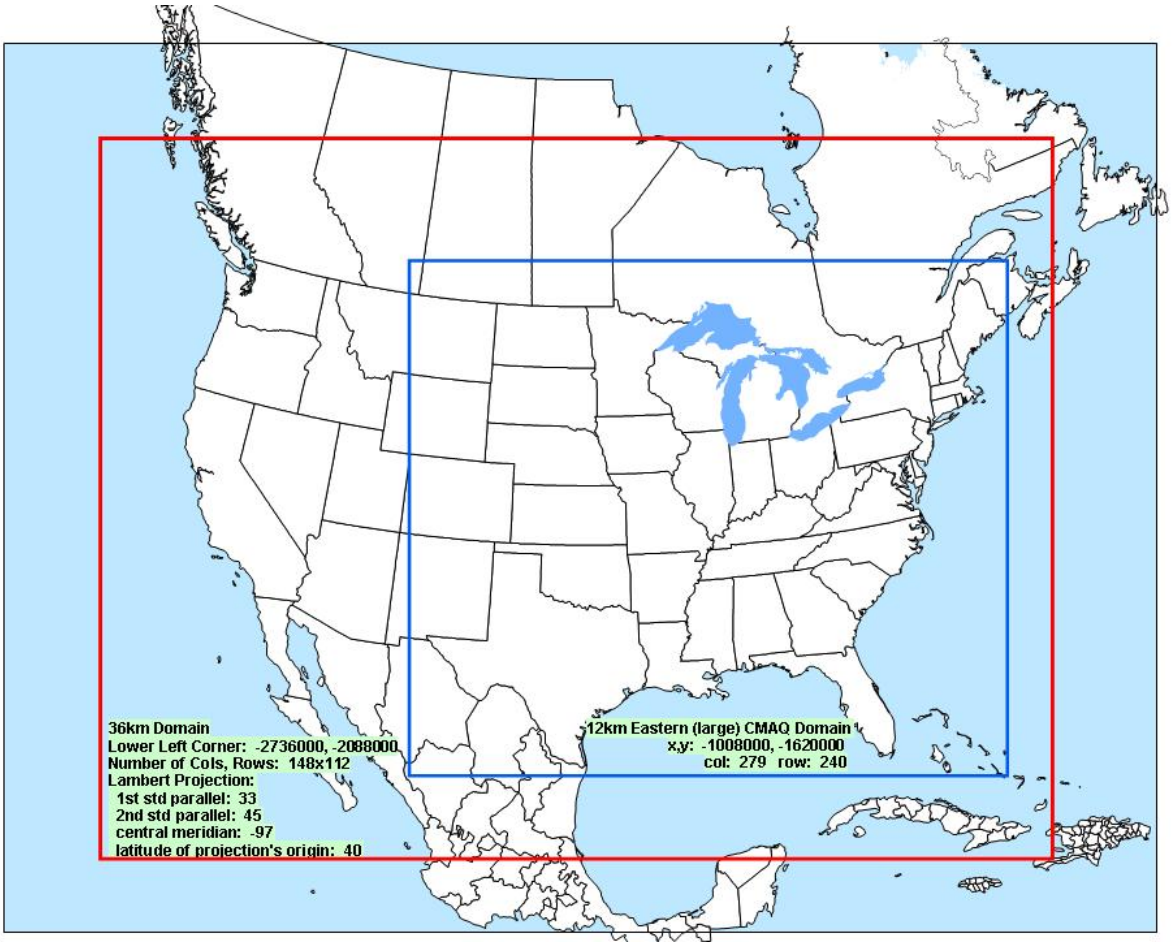


Figure 3. The maximum area of the United States covered by the 2002 CMAQ data for the 12 km x 12 km and 36 km x 36 km grids.

Table 2. The coordinates used in the simulation runs.

Grid	Year	Coverage Used in Simulation Runs		Full Grid Coverage Available	
		Column (x) (Start, End)	Row (y) (Start, End)	Column (x) (Start, End)	Row (y) (Start, End)
12 km	2001	(40, 213)	(30, 188)	(1, 213)	(1, 188)
	2002	(100, 279)	(100, 240)	(1, 279)	(1, 240)
36 km	2001	(30, 170)	(30, 154)	(1, 200)	(1, 184)
	2002	(28, 120)	(30, 82)	(1, 148)	(1,112)

3.2 Network Monitors

Two sets of network monitor data are used in the T-SpACE model simulation. The EPA-validated Federal Reference Method (FRM) monitors that are part of the NAMS/SLAMS network are used to develop the estimated (retrospectively predicted) air pollution concentration surface. For the ozone simulation, concentrations from monitors in the Clean Air Status and Trends Network (CASTNET) are used for comparing predictions from the three types of modeling processes used: T-SpACE, Kriging, and CMAQ/CAMx.

3.2.1. T-SpACE Model Input Monitor File

EPA has a repository of ambient air quality data. This repository is called the Air Quality System (AQS), and is located at: https://www.epa.gov/aqs. The AQS contains measurements of concentrations for the six (6) criteria pollutants (i.e., ozone [O_3]; carbon monoxide [CO]; lead [Pb]; nitrogen dioxide [NO_2]; sulfur dioxide [SO_2], and; particulate matter [PM]) from EPA-developed and funded monitoring networks within the United States. These AQS monitors are FRM monitors. For this evaluation, ozone data available for 2001 and 2002 were extracted from AQS. The exact names of the files are listed in Table 1. Ozone concentrations were collected every day. Daily 8-hour maximum concentrations were used in the analysis. The number of monitors included in the analysis varied with the CMAQ grid that was used in the analysis. There are upwards of 1,200 ozone monitors with data available in AQS. Table 1 lists the original data files downloaded by EPA from EPA's AQS and utilized in the analysis.

The CMAQ files had missing data from November 30, 2001 through December 31, 2001 and December 1, 2002 through December 31, 2002. The data recorded on November 30, 2002 were ten times higher than any daily value in that year and were not included in the analysis. Therefore, the analysis conducted on both the 2001 and 2002 data ran from January 1 through November 29.

3.2.2 Validation

In order to evaluate the predictive ability of T-SpACE, and to compare it to the predictive ability of the Krige model, data from the Clean Air Status and Trends Network (CASTNET) was utilized in the data validation. This network is similar to the FRM network, where measurements are taken on a daily basis. Data from 2001 and 2002 were available. Those monitors that were located on the CMAQ grid were included in the data validation analysis. The red stars in Figure 4 indicate that 47 2001 CASTNET monitors were located within the CMAQ grid boundaries, and were used for the analysis. Four of these 47 monitors were co-located with FRM monitors.

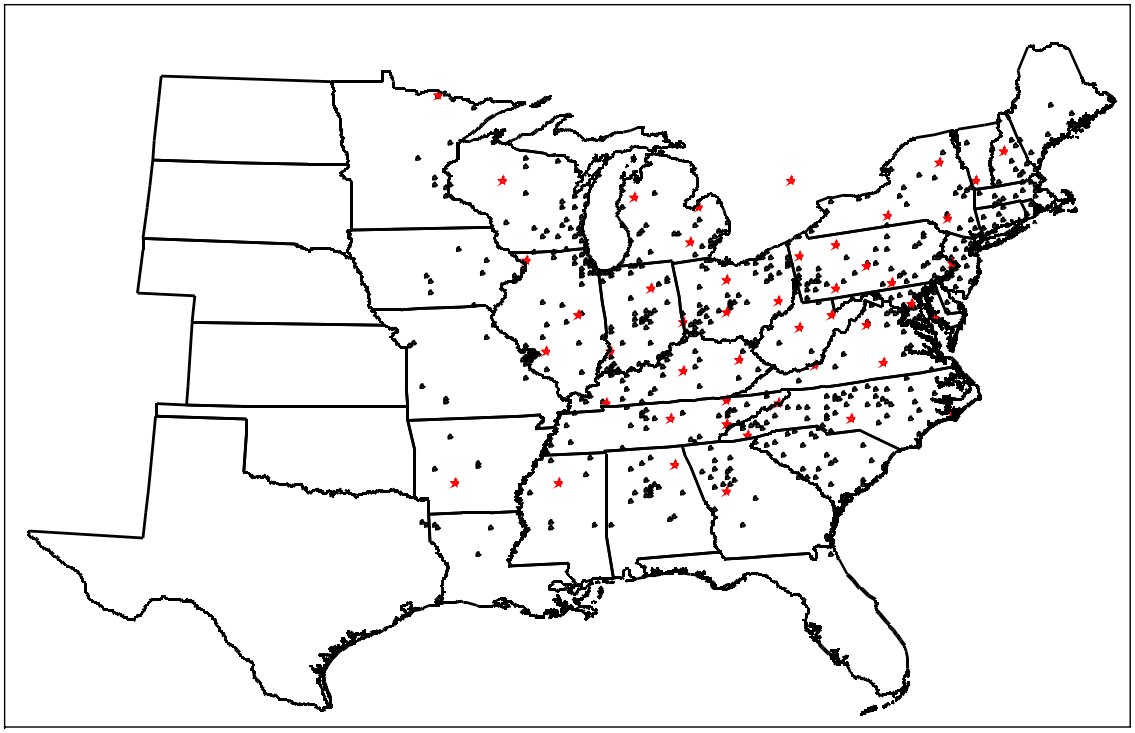


Figure 4. Location of the 2001 CASTNET monitors relative to the FRM monitors included in the analysis.

4.0 Input Parameters

The T-SpACE model requires input from the user. A file with a *.PAR (i.e., parameter) extension is where the user's choices for the model simulation are stored. Appendix E provides an example of the *.PAR file used for Run 1 of the analysis of the 2001, 12 km ozone data. Table 3 provides a brief description of the file and parameter choices. USEPA (2007), and McMillan et al., (2010) provides more details about the model.

Table 2 lists the choices that were made for the x and y grid sizes for each year, and CMAQ/CAMx coordinate system. Note that a simulation was not run for the 36 km x 36 km grid for 2001. The CMAQ data were not available in the appropriate format (NetCDF).

Two time frames for both years were chosen for the 2001 simulations, May 1 (Julian Day 121) to October 31, the ozone season, and January 1 through November 29, a "full" year. Note that the CMAQ data was missing for November 30 through December 31 in 2001 and 2002, and this is why the full year was defined to be January 1 through November 29.

The burn-in steps and the simulation steps were reduced to 10 and 50 for the simulations, to run the simulations relatively quickly. Table 4 illustrates the estimated time to run various scenarios using the T-SpACE software.

Table 3. Summary of the T-SpACE Model Input Parameters.

Parameter name	Parameter	Values	Description
Model version		5-4H	This is the version of the model utilized in the analysis. Currently the MCMC version is 5.1, but 5-4h is used as the current version of T-SpACE. PAR file to designate the MCMC model 5.1 and the addition of the calculation of the 4th highest average concentration (4h).
Grid size	t	start, end	Time frame within a calendar year in days from January 1 (or the date specified below).
	x	start, end	The range for the x-coordinates of the CMAQ grid illustrated in Figures 1, 2, and 3. The choice is dependent on the year and the grid size being used for the analysis. See Table 2 for the ranges used in the analysis and the full range available.
	y	start, end	The range for the y-coordinates of the CMAQ grid illustrated in Figures 1, 2, and 3. The choice is dependent on the year and the grid size being used for the analysis. See Table 2 for the ranges used in the analysis and the full range available.
Year, month, day of time step 1		year, month, day	This is a reference point for the model. It is best to leave this as January 1. The values in "Grid size:t" above, are dependent upon this value. Changing it to anything other than January 1 may result in incorrect information later on.
Bias Spline	t	start, end, number of basis points	Start and end are the same as above. The basis points are used to define the way the model calculates the splines. 4, 8, and 7 are the recommended basis points for t, x, and y, respectively, when the analysis is similar in size to the grid utilized for this validation. These values can be changed depending on the size of the grid chosen for the analysis.
	x	start, end, number of basis points	
	y	start, end, number of basis points	
Prior Parameters	Mu	Mean, variance	The mean for the underlying space-time log transformed ozone process. (Natural log of the concentrations is used in the analysis.) This is initialized to a non-informative normal prior (recommended).
	BetaD	Mean, variance	The covariates for the CMAQ bias structure. This is initialized to a non-informative normal prior (recommended).
	TauX	Mean, precision	The precision of the measurement error in the monitor (TauX) and the CMAQ (TauY) observations, respectively. This is assigned a Gamma distribution. The mean and precision are chosen based on existing knowledge about measurement errors associated with FRM monitoring data and CMAQ results. Various scenarios for the relationship between TauX and TauY are examined in the validation process.
	TauY	Mean, precision	
	TauZ	Mean, precision	The precision of the underlying space-time log transformed ozone process. A non-informative Gamma distribution with mean = 0.001 and precision = 0.001 is assigned. This is recommended.
	RhoZ	Mean, precision	Temporal autocorrelation parameter of the underlying space-time log transformed ozone process. A uniform prior distribution between 0 and 1 is defined for RhoZ. This is recommended.
Directory where the input data are stored		Directory	Both the monitor and the CMAQ input data files must be located in the same directory.
Monitor file		.CSV file	The input monitor data must be a .CSV file of a specified format, which is generated through EPA software.
CMAQ file		.CSV file	The input CMAQ data must be a .CSV file of a specified format that is generated through software developed for EPA. -999 is used to indicate a missing value. Missing value data must be excluded from the analysis by the user choosing the appropriate days and grid cells to include in the analysis.
Order of neighborhood		numeric value	Defines the neighborhood structure of the model. This value can range from first order (1) to fourth order (4). Default is a first order neighborhood. Figure 4 illustrates the neighborhood definitions.

Table 3. (Continued).

Parameter name	Parameter	Values	Description
Track chain flag		numeric value	0 or 1 to not write or write, respectively, the chain of the simulation to the screen. By default, the value is initialized to 0 and the chain is not written to a file. This speeds up the simulation which can be laborious as the grid and time frames become large.
Enable calculation of 4th highest surface		numeric value	0 or 1 to have the simulation calculate an average 4th highest concentration surface for the grid and time frame chosen. The default is 0 (no calculation). A file with _4h.csv as the suffix is developed if this is initialized to 1. (This surface is not evaluated.)
Track chain flag for 4th highest surface		numeric value	0 or 1 to not write or write, respectively, the chain of the simulation for the calculation of the average 4th highest concentration surface. By default, the value is initialized to 0 and the chain is not written to a file. (This surface is not evaluated.)
Directory for where the output data from the model run will be stored		directory	Location to store the series of output files from the model run.
Simulation output file name		.CSV file	Name to be used to store the predicted ozone surface and associated statistics.
Random seeds	1	numeric value	Two seeds are needed for the simulation. By default, these are initialized to 1234 and 5678. These can be changed.
	2	numeric value	
Sampling Period		numeric value	This is how often sampling occurs. The default is 1 (i.e., for each model iteration).
Prompting Period		numeric value	The user can see the simulation progress on the screen. The default is to see each iteration (1). Higher numbers will reduce the amount of information written to the screen as the simulation occurs, thus potentially speeding up the process.
Simulation Flag		numeric value	This indicates whether the simulation results will overwrite a file if a file already exists with the same name.
Number of burn-in steps		numeric value	Number of burn-in steps for the simulation. The default is 1000. For the validation simulations, 10 were used.
Number of simulation steps		numeric value	Number of simulation loops. The default is 5000. For the validation simulations, 50 were used.

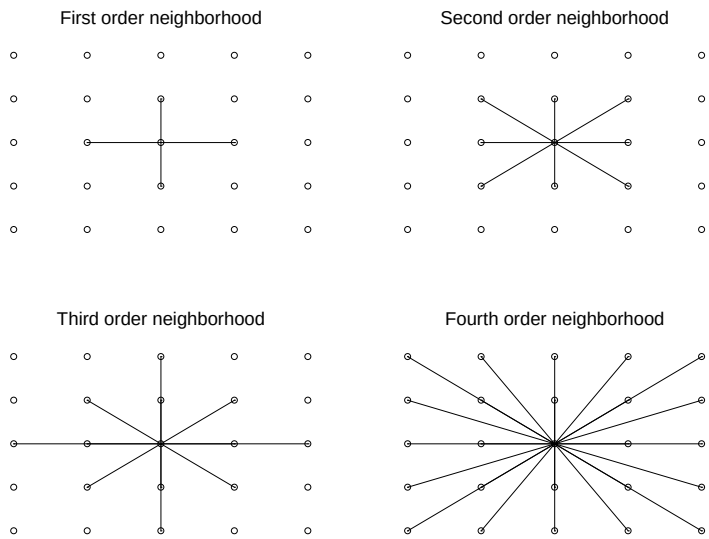


Figure 4. Neighborhood Definitions.

Table 4. Estimated times to run differing number of burn-ins, iterations, and timeframes.

Scenario	Grid size	Estimated Time
10 burn-in, 50 iterations, ozone season	12km x 12 km	9 minutes
	36km x 36 km	2 minutes
50 burn-in, 100 iterations, "full" year	12km x 12 km	50 minutes
	36km x 36 km	9 minutes

The only other parameters that were changed from the default values were the mean and precision for TauX (τ_x) and TauY (τ_y). The choices for the values were provided by EPA and were also based on previous research. Some of the previous research indicated a coefficient of variation (CV) around 20% for τ_x and a CV of 80% for τ_y may be appropriate starting points for choosing the mean and precision (USEPA, 2007). Note that there is an overwhelming amount of CMAQ data available as compared to the air pollution monitor data. It takes a strong prior on τ_y for the prior to have much effect on the posterior. The actual choices for their parameters are illustrated in Tables 5, 6, and 7.

5.0 Model validation results

Using the data and parameter files described above, data validation runs were performed to assess the stability of the T-SpACE model when parameters are changed, and to compare the predictive abilities of T-SpACE with other models, in particular, a Krige model and the CMAQ/CAMx model results.

Table 5 provides a summary of the statistics that are used in the validation analysis.

Table 5. Description of the Validation Summary Statistics.

Statistic	Description
Run number	The number of the run. The choices for the T_X and T_Y are the same for each run number across the year and grid choices.
Prior Assumptions	
Parameter	These are the τ parameters which vary across the runs.
Mean	The assumed prior mean for the Gamma distribution.
Precision	The assumed prior precision for the Gamma distribution.
Marginal Posterior	
Means	These are the average of the estimated posterior mean for the τ's that were produced during the simulation.
Predicted Standard Error (SE)	This is the average of the model-predicted standard error for the mean predicted concentration.
95% Prediction Interval	Using the predicted standard error of the mean predicted concentration from T-SpACE, this represents the % of the validation monitor concentration that fall within the 95% confidence intervals for T-SpACE.
Krige Prediction Interval	Using the predicted standard error of the mean predicted concentration from the Krige model, this represents the % of the validation monitor concentration that fall within the 95% confidence intervals for the Krige model.
Overall Bias	Average of the absolute difference between the monitor concentration and T-SpACE, Krige, and CMAQ/CAMx predicted concentrations, respectively.
Overall MSE	Average of the square of the bias (described above). The average is over location and time.
% improvement by day	For each prediction the comparison represents the % of days which the Krige model and the CMAQ/CAMx model have MSEs that are greater than the MSE for T-SpACE.
% improvement by time	For each prediction the comparison represents the % of time which the Krige model and the CMAQ/CAMx model have MSEs that are greater than the MSE for T-SpACE.

Tables 6, 7, and 8 provide the results of the validation runs using data from May 1 through October 31, the ozone season, while Tables 9, 10, and 11 provide the results for a "full" year of data, January 1 through November 29.

For each time frame, there were five validation runs per year (2001, 2002) and CMAQ grid (12 km x 12 km, 36 km x 36 km). Runs for the 2001, 36 km x 36 km was not possible, since the CMAQ data were not available in the appropriate format (NetCDF), so there are three sets of data validation runs for the ozone data.

5.1 Explanation of the metrics used in the validation

Table 5 describes the information provided in Tables 6 through 11. For reference, the SAS code that generated the summary statistics for the 2001, 12 km, Run 4, is provided in Appendix F (pages F-2 to F-8). The SAS code for developing the Krige model used in the analysis of the 2001, 12 km, Run 4, is provided in Appendix D (pages D-2 to D-5). Similar SAS code was utilized for the other year/grid combinations.

Table 6. Summary Statistics for Model Runs (Model 5_4h) for ozone, 36 kilometers, 2001 (May 1 – October 31).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	75.0E+6	1.0E+6	74.9	0.10	4.84	0.32	94.8	66.5	T-SpACE	0.06	0.04		
	τ_y	0.1E+8	1.0E+8							Krige	0.02	0.03	5.43	25.5
	τ_z	1.0E-3	1.0E-3							CMAQ	0.09	0.06	78.3	78.4
2	τ_x	1.0E+7	1.0E+6	10.0	0.93	7.87	0.29	94.9	66.5	T-SpACE	0.04	0.04		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.03	6.52	23.5
	τ_z	1.0E-3	1.0E-3							CMAQ	0.09	0.06	85.9	82.4
3	τ_x	8.0E+6	1.0E+6	8.04	0.93	7.80	0.30	94.9	66.5	T-SpACE	0.04	0.04		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.03	5.43	17.7
	τ_z	1.0E-3	1.0E-3							CMAQ	0.09	0.06	83.2	82.4
4	τ_x	1.6E+7	1.0E+6	16.0	0.93	7.98	0.28	94.8	66.5	T-SpACE	0.03	0.04		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.03	8.70	25.5
	τ_z	1.0E-3	1.0E-3							CMAQ	0.09	0.06	90.2	84.3
5	τ_x	1.6E+7	1.0E+6	16.0	1.42	9.56	0.26	94.8	66.5	T-SpACE	0.03	0.04		
	τ_y	1.0E+6	1.0E+6							Krige	0.02	0.03	9.24	27.45
	τ_z	1.0E-3	1.0E-3							T-SpACE	0.09	0.06	90.76	84.31

* The other parameters are assumed to be non-informative.

Table 7. Summary Statistics for Model Runs (Model 5_4h) for ozone, 12 kilometers, 2001 (May 1- October 31).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	75.0E+6	1.0E+6	75.0	0.12	3.62	0.44	93.2	84.1	T-SpACE	0.08	0.07		
	τ_y	0.1E+8	1.0E+8							Krige	0.02	0.02	0	8.51
	τ_z	1.0E-3	1.0E-3							CMAQ	0.14	0.08	52.7	51.1
2	τ_x	1.0E+7	1.0E+6	10.0	1.96	10.7	0.26	93.1	84.1	T-SpACE	0.02	0.06		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.02	0.5	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.14	0.08	76.6	78.7
3	τ_x	8.0E+6	1.0E+6	8.03	1.96	10.7	0.27	93.1	84.1	T-SpACE	0.02	0.06		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.02	0.5	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.14	0.08	76.1	78.7
4	τ_x	1.6E+7	1.0E+6	16.0	1.96	10.8	0.26	93.0	84.1	T-SpACE	0.01	0.05		
	τ_y	0.5E+6	1.0E+6							Krige	0.02	0.02	1.1	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.14	0.08	80.4	80.9
5	τ_x	1.6E+7	1.0E+6	16.0	2.32	35.3	0.15	92.1	84.1	T-SpACE	-0.03	0.05		
	τ_y	1.0E+6	1.0E+6							Krige	0.02	0.02	2.72	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.14	0.08	74.5	76.6

* The other parameters are assumed to be non-informative.

Table 8. Summary Statistics for Model Runs (Model 5_4h) for ozone, 12 kilometers, 2002 (May 1 – October 31).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	75.0E+6	1.0E+6	75.0	0.12	3.50	0.43	93.8	80.1	T-SpACE	0.07	0.07		
	τ_y	0.1E+8	1.0E+8							Krige	0.01	0.02	1.09	2.13
	τ_z	1.0E-3	1.0E-3							CMAQ	0.11	0.08	52.2	42.6
2	τ_x	1.0E+7	1.0E+6	10.0	1.83	10.2	0.27	93.7	80.1	T-SpACE	0.01	0.05		
	τ_y	0.5E+6	1.0E+6							Krige	0.01	0.02	2.17	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.11	0.08	76.1	78.7
3	τ_x	8.0E+6	1.0E+6	8.03	1.83	10.2	0.27	93.7	80.1	T-SpACE	0.01	0.06		
	τ_y	0.5E+6	1.0E+6							Krige	0.01	0.02	1.09	2.13
	τ_z	1.0E-3	1.0E-3							CMAQ	0.11	0.08	72.3	78.7
4	τ_x	1.6E+7	1.0E+6	16.0	1.83	10.2	0.27	93.5	80.1	T-SpACE	0.00	0.05		
	τ_y	0.5E+6	1.0E+6							Krige	0.01	0.02	3.26	0
	τ_z	1.0E-3	1.0E-3							CMAQ	0.11	0.08	77.7	78.7
5	τ_x	1.6E+7	1.0E+6	16.0	2.17	10.6	0.26	93.5	80.1	T-SpACE	0.00	0.05		
	τ_y	1.0E+6	1.0E+6							Krige	0.01	0.02	3.26	0
	τ_z	1.0E-3	1.0E-3							T-SpACE	0.11	0.08	77.7	78.7

* The other parameters are assumed to be non-informative.

Table 9. Summary Statistics for Model Runs (Model 5_4h) for ozone, 12 kilometers, 2001 (01/01/2001-11/29/2001).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	7.5E+01	1.3E+04	75.0	0.15	6.76	0.34	92.1	81.6	T-SpACE	0.03	0.06		
	τ_y	1.0E-01	1.0E+09							Krige	0.08	0.04	12.31	21.28
	τ_z	1.0E+00	1.0E-03							CMAQ	0.13	0.10	81.68	76.60
2	τ_x	1.0E+01	1.0E+05	10.0	4.73	38.53	0.14	91.2	81.6	T-SpACE	-0.02	0.07		
	τ_y	5.0E-01	2.0E+06							Krige	0.08	0.04	11.41	17.02
	τ_z	1.0E+00	1.0E-03							CMAQ	0.13	0.10	66.97	76.60
3	τ_x	8.0E+00	1.3E+05	8.03	4.73	38.62	0.14	91.6	81.6	T-SpACE	-0.02	0.07		
	τ_y	5.0E-01	2.0E+06							Krige	0.08	0.04	11.41	17.02
	τ_z	1.0E+00	1.0E-03							CMAQ	0.13	0.10	67.27	76.60
4	τ_x	1.6E+01	6.3E+04	16.0	4.73	38.18	0.14	90.0	81.6	T-SpACE	-0.03	0.07		
	τ_y	5.0E-01	2.0E+06							Krige	0.08	0.04	12.01	17.02
	τ_z	1.0E+00	1.0E-03							CMAQ	0.13	0.10	70.27	78.72
5	τ_x	1.6E+01	6.3E+04	16.0	5.21	39.7	0.14	90.0	81.6	T-SpACE	-0.03	0.07		
	τ_y	1.0E+00	1.0E+06							Krige	0.08	0.04	12.01	17.02
	τ_z	1.0E+00	1.0E-03							CMAQ	0.13	0.10	69.97	78.72

Table 10. Summary Statistics for Model Runs (Model 5_4h) for ozone, 12 kilometers, 2002 (01/01/02-11/29/02).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	7.5E+01	1.3E+04	75.0	0.14	6.48	0.34	93.0	79.4	T-SpACE	0.03	0.06		
	τ_y	1.0E-01	1.0E+09							Krige	0.05	0.03	7.21	12.77
	τ_z	1.0E+00	1.0E-03							CMAQ	0.09	0.08	65.8	70.21
2	τ_x	1.0E+01	1.0E+05	10.0	4.42	38.32	0.14	92.0	79.4	T-SpACE	-0.01	0.07		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	6.61	10.64
	τ_z	1.0E+00	1.0E-03							CMAQ	0.09	0.08	48.35	63.83
3	τ_x	8.0E+00	1.3E+05	8.03	4.42	38.48	0.14	92.4	79.4	T-SpACE	-0.01	0.07		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	4.8	10.64
	τ_z	1.0E+00	1.0E-03							CMAQ	0.09	0.08	46.85	63.83
4	τ_x	1.6E+01	6.3E+04	16.0	4.42	37.77	0.14	91.1	79.4	T-SpACE	-0.01	0.06		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	6.91	8.51
	τ_z	1.0E+00	1.0E-03							CMAQ	0.09	0.08	53.75	65.96
5	τ_x	1.6E+01	6.3E+04	16.0	4.90	39.51	0.14	90.9	79.4	T-SpACE	-0.01	0.06		
	τ_y	1.0E+00	1.0E+06							Krige	0.05	0.03	6.91	8.51
	τ_z	1.0E+00	1.0E-03							CMAQ	0.09	0.08	53.15	65.96

* The other parameters are assumed to be non-informative.

Table 11. Summary Statistics for Model Runs (Model 5_4h) for ozone, 36 kilometers, 2002 (01/01/02-11/29/02).

Run number	Prior Assumptions*			Marginal Posteriors Mean			Predicted SE	95% Prediction Interval	Krige Prediction Interval	Surface	Overall Bias	Overall MSE	% Improvement by Day	% Improvement by Site
	Parameter	mean	precision	τ_x	τ_y	τ_z								
1	τ_x	7.5E+01	1.3E+04	74.99	0.11	7.20	0.27	99.7	69.6	T-SpACE	0.02	0.04		
	τ_y	1.0E-01	1.0E+09							Krige	0.05	0.03	26.13	34.62
	τ_z	1.0E+00	1.0E-03							CMAQ	0.08	0.06	78.08	84.62
2	τ_x	1.0E+01	1.0E+05	10.05	1.28	14.14	0.23	99.7	69.6	T-SpACE	0.01	0.05		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	9.01	30.77
	τ_z	1.0E+00	1.0E-03							CMAQ	0.08	0.06	58.56	80.77
3	τ_x	8.0E+00	1.3E+05	8.05	1.28	14.28	0.23	99.8	69.6	T-SpACE	0.01	0.05		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	7.21	19.23
	τ_z	1.0E+00	1.0E-03							CMAQ	0.08	0.06	54.35	80.77
4	τ_x	1.6E+01	6.3E+04	16.04	1.28	13.75	0.23	99.6	69.6	T-SpACE	0.01	0.04		
	τ_y	5.0E-01	2.0E+06							Krige	0.05	0.03	13.21	34.62
	τ_z	1.0E+00	1.0E-03							CMAQ	0.08	0.06	69.67	84.62
5	τ_x	1.6E+01	6.3E+04	16.04	1.77	15.55	0.21	99.5	69.6	T-SpACE	0.01	0.05		
	τ_y	1.0E+00	1.0E+06							Krige	0.05	0.03	10.81	34.62
	τ_z	1.0E+00	1.0E-03							CMAQ	0.08	0.06	66.37	80.77

* The other parameters are assumed to be non-informative.

5.2 Summary of results

As previously discussed, there were two sets of data validation runs that were performed. Surfaces were developed for the ozone season, defined as May 1 through October 31 and for a "full" year of data. As mentioned earlier, a full year was January 1 through November 29, because data were missing for the CMAQ surface from November 30 through December 31 in both 2001 and 2002. Below are a series of observations from the validation runs. We first compare within the time frame assessed, and then provide overall observations.

5.2.1. Comparisons for the runs during the "Ozone" season of May 1 through October 31

- In general, across years and grid size, Runs 2 through 5 produce very similar results. Run 1 produces similar results across years and grid size, but produces different results than runs 2 through 5.
    - The biggest difference between Runs 2-5 and Run 1 is the choice of the prior mean for τ_x and τ_y. The Run 1 TauX mean prior values are large, relative to the marginal posterior mean, and small relative to the precision priors chosen for the other runs.

    - More variability appears to have been introduced into the predictions in Run 1, yielding a higher Predicted SE, Overall Bias, and Overall MSE. In addition, T-SpACE did not perform as well against the CMAQ model as it did in Runs 2-5.

- Focusing on Runs 2-5,
    - The marginal means are relatively closer to each other, and appear to produce relatively stable results even though the means are changed across the runs.

    - The 95% Prediction Interval for the T-SpACE model in general produces results where nearly 95% of the validation monitors are within the prediction intervals. The Krige Prediction Interval, in general, captures less than 95% of the validation monitors, ranging from 67% for the 36 km, 2001 runs, to 84% for the 12 km, 2001 runs.

    - When comparing the three models, the T-SpACE and Krige models tend to have similar overall bias and overall MSE. The CMAQ model overall bias and overall MSE are relatively larger than both the T-SpACE bias and the krige bias.

    - T-SpACE in general shows a smaller MSE, by day and by site, than the CMAQ data. This indicates that by utilizing the monitor data, the CMAQ variability is “smoothed” out in the T-SpACE model, which produces better predictions.

    - The Krige model in general shows a smaller MSE, by day and by site, as compared to T-SpACE. During the time frame utilized for the simulation runs, the ozone monitors have relatively little variation across time and across the US. Figure 5 provides a view of the data for September 17, 2001. The monitor values are the circles. There is very little variation in the values observed. Kriging will perform very well in situations such as this. Also notice the CMAQ surface in Figure 7 (the yellow surface that is below the orange circles). This indicates that

there might be occasions when the CMAQ results are introducing unnecessary variability into the T-SpACE model.

5.2.2. Comparisons for the runs for a "full" year of data (January 1 through November 29)

In general, the results for the "full" year of data do not differ significantly from the results observed for the ozone season results.

- We do observe that the Krige model has a higher bias than T-SpACE when a full year of data is used. This is consistent across all runs.

- The difference between the MSE for T-SpACE and Krige models is very similar to that observed for the simulations using the ozone season. The Krige model has a slightly smaller MSE than the T-SpACE model across all runs.

- We do see that the T-SpACE percent (%) improvement, by day and by site, is higher for the full year of data. This may indicate that outside the ozone season, the ozone concentrations are more variable, and T-SpACE is able to capture this.

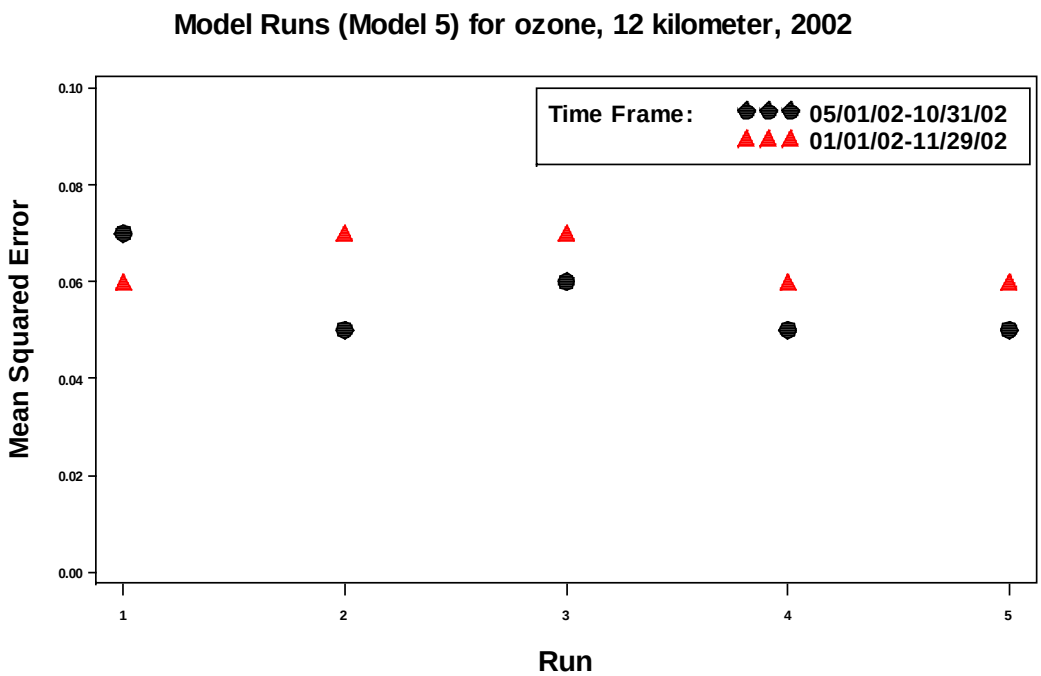


Figure 5. Illustration of the MSE for the ozone season versus the "full" year.

5.2.3. Overall summary

5.2.3.1 Comparison of Ozone Season results to the Full Year results

As seen in Figure 5, the T-SpACE MSE for Runs 2 through 5 is slightly higher for the full year than the ozone season. This is consistent across the Runs. This would seem to be expected as the ozone concentrations observed outside of the ozone season are more varied than those observed during the ozone season. For Run 1, the prior Gamma mean for τ_y is significantly larger (0.1E+8 as compared to 0.5E+6) than those chosen for the other runs. Though the MSE difference is not significantly large, it does produce different results from the other runs.

In general, the T-SpACE and Krige models appear to have their strengths, regardless of the time frame over which the simulation is run. There is an indication with the bias, the % improvement statistics, and prediction interval statistics that the T-SpACE may be able to better handle the variability that may occur outside the ozone season when the concentrations are relatively stable. Though, it should be noted that the indications are slight and it is difficult to make a definitive decision.

5.2.3.2 The T-SpACE model does a better job of capturing the variability as compared to the Krige prediction model

The 95% Prediction Interval summary statistic records the rate at which the prediction intervals included the monitor value (across all days and sites). Ideally the coverage should be close to 95%. As seen across all years and the two time frames, the T-SpACE model is close to or achieves the 95% rate. The Kriging method achieves rates that range from 67% to 84%.

A lower rate, such as the kriging summary statistics, could indicate that the estimated uncertainty is too small or that there is a systematic positive or negative pattern of error (a positive or negative bias) in the estimate of the true value. A rate higher than 95% could indicate that the estimated uncertainty is too large – meaning that the estimated interval should be narrower.

In the case of kriging, performance may be impacted by the choice of covariance structure and the estimates of its parameters. Also, kriging assumes a stationary random process, i.e., the response characteristics such as mean, variance and covariance do not depend on location. Deviations from these assumptions may also impact performance.

T-SpACE can derive posterior intervals using the posterior estimate of response variance, or numerically by taking the sample standard deviation of the MCMC chain outcomes of the MCMC process after the burn-in period. The T-SpACE method assumes that the true response is characterized by a conditional autoregressive process for which the mean values at FRM and CMAQ/CAMx locations are indicated by a combination of the FRM and CMAQ/CAMx values.

5.3.2.3 Uncertainty of the krige prediction model is less than T-SpACE

The bias is the difference between the monitor's reading and the predicted response. The bias is not an absolute value and equidistant high and low values can create a bias that is low when in fact the variability is high. The MSE is the square of these differences. The MSE penalizes large differences more than small differences, thus highlighting the larger differences that may occur between the predicted value and the monitor reading.

The T-SpACE model tends to have a lower bias than the Krige model while the T-SpACE MSE is larger than the Krige model. Figure 6 provides a comparison of the MSE for the three "models", T-SpACE, Kriging, and CMAQ for the 12 km, 2002, full year simulation runs. In general, the differences between these statistics is not very large, indicating that the performance as measured by these statistics is very similar.

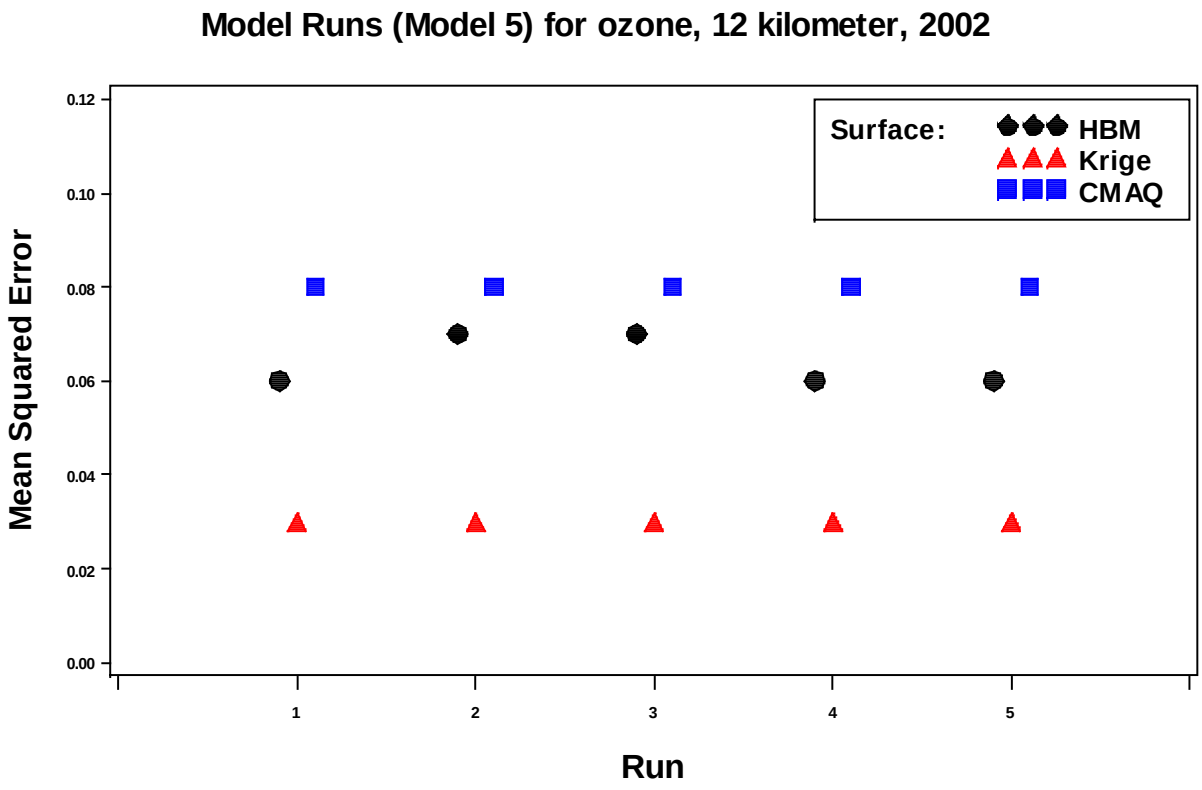


Figure 6. Comparison of the MSE across the three models for the "full" year for 12 km, 2002. Note: HBM = T-SpACE

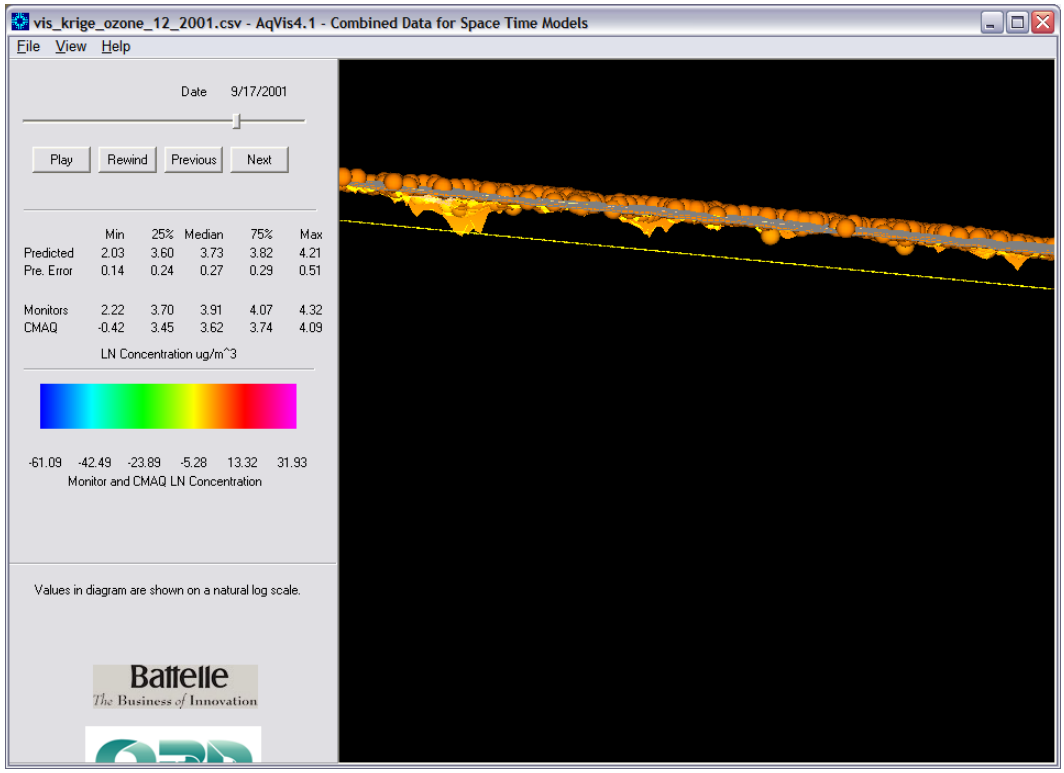


Figure 7. Illustration of the relatively little variation in monitor concentrations during the time frame for the validation runs.

In trying to understand why the krige results had a consistently smaller MSE across time and space than the T-SpACE, an interesting predictive estimate was located over a week in July of 2001. Figures 8 through 14 illustrate the predictions that start off below expected levels on July 3, 2001 and gradually become more pronounced over time with the largest prediction occurring on July 7, 2001. Figures 15 and 16 provide another view of the CMAQ surface and the monitor results. These figures illustrate a “hole” in the CMAQ surface that remains the same on both July 3, 2001 and July 7, 2001 (the brownish surface with the “nick” at the top of the grid to the top, right-side of the grid. The predicted surface (the yellow surface) displays a “hole” a little larger than the “nick” on July 3, 2001, but shows an even larger “hole” on July 7, 2001 where the “nick” in the CMAQ surface is still the same size.

This indicates that the completeness of the CMAQ grid must be examined prior to the simulation, otherwise, there could potentially be a pronounced effect on the T-SpACE predicted results. In our validation analysis, the impact is minimal.

Overall, the Krige model appears to do slightly better than T-SpACE in consistently estimating the concentrations of the individual monitors used in the model validation process, due to the relatively little variability within the monitors. But, the T-SpACE model appears to do a relatively better job of capturing the variability in the model validation with ozone concentrations, allowing the T-SpACE the ability to capture nearly 95% of the validation monitor concentrations within its 95% prediction interval. This is compared to 67% to 84% for the Krige model.

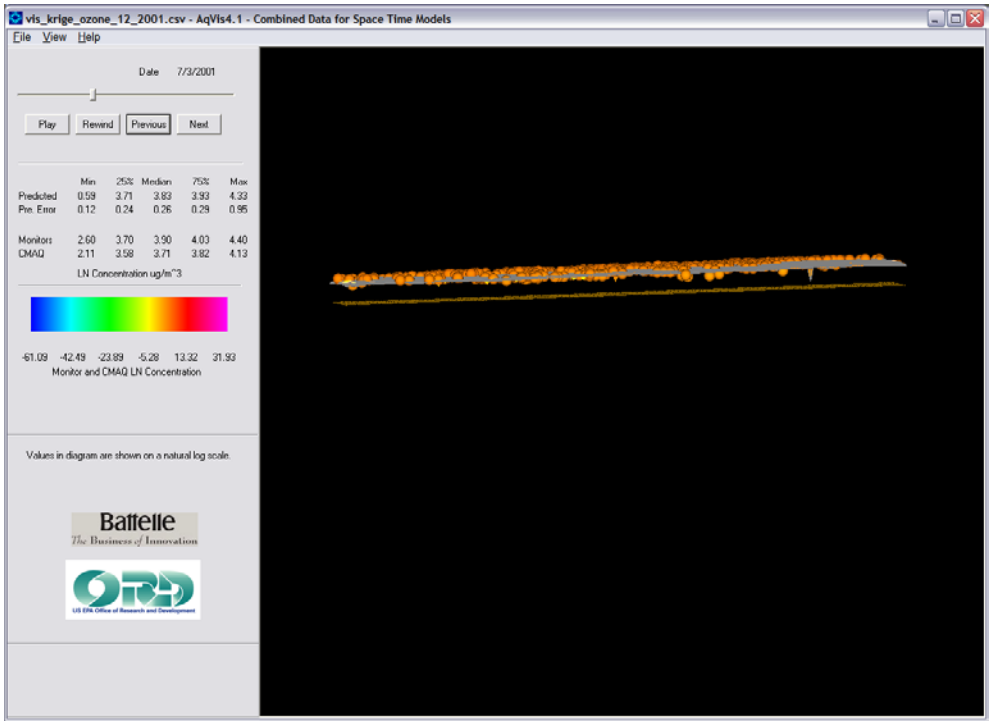


Figure 8. July 3, 2001-Anomolous Prediction.

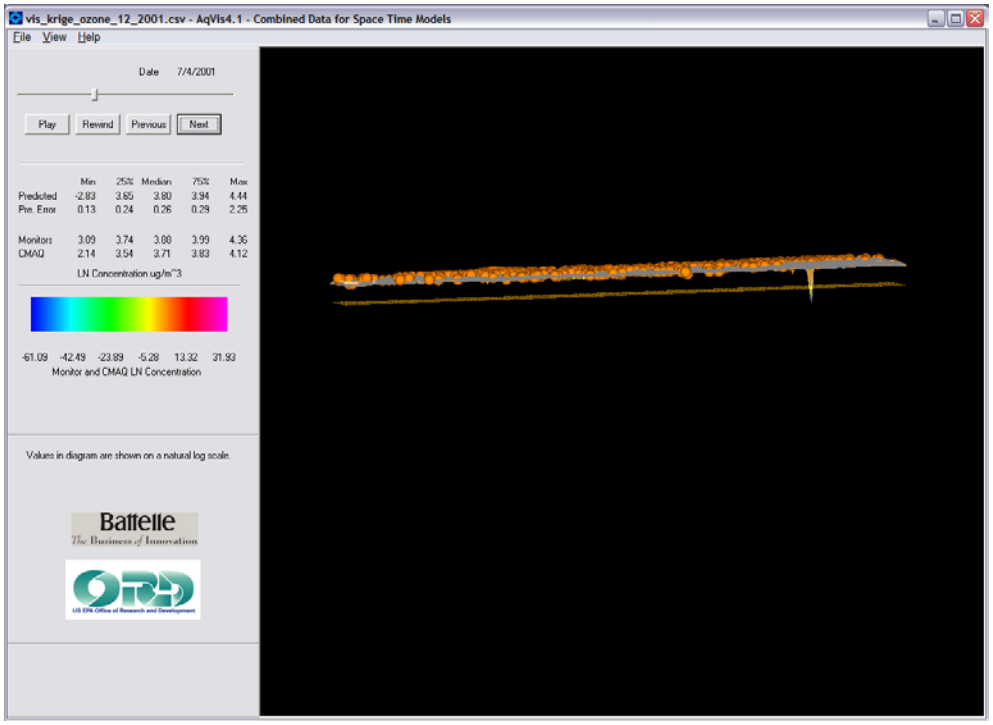


Figure 9. July 4, 2001-Anomolous Prediction.

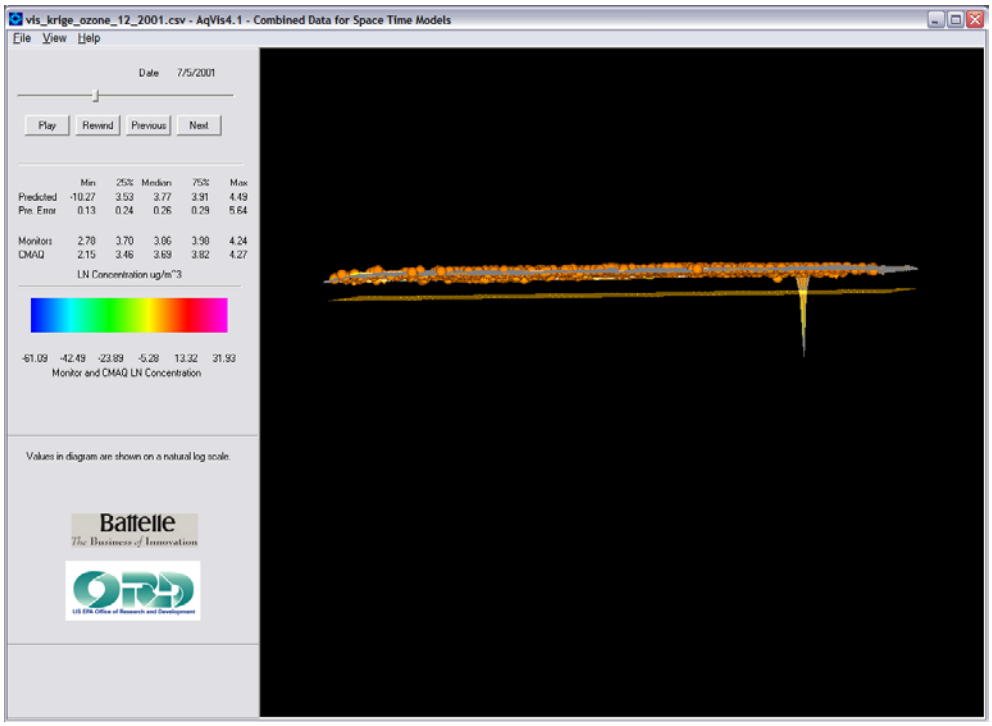


Figure 10. July 5, 2001-Anomolous Prediction.

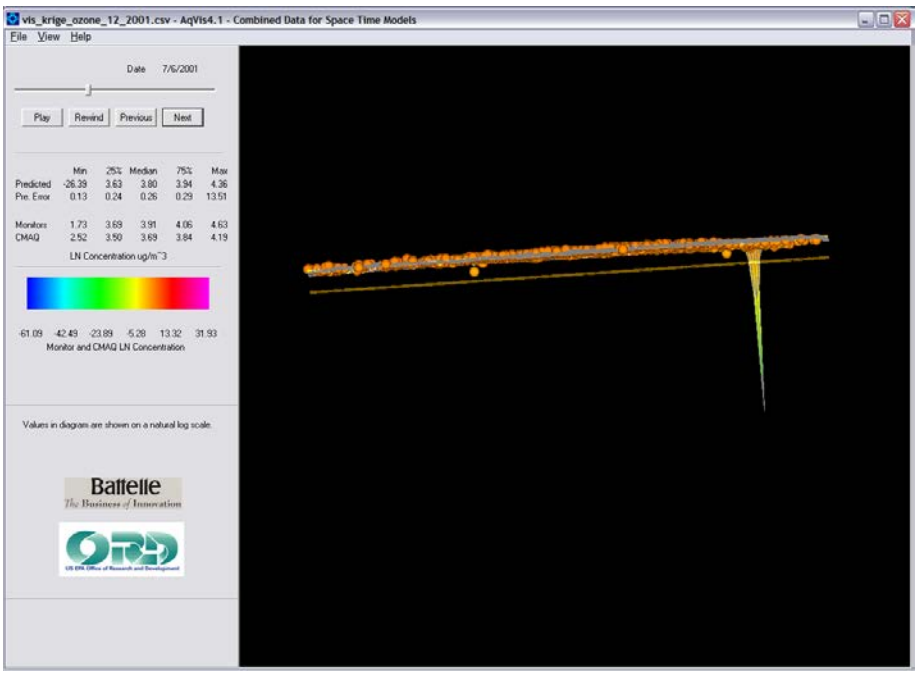


Figure 11. July 6, 2001-Anomolous Prediction.

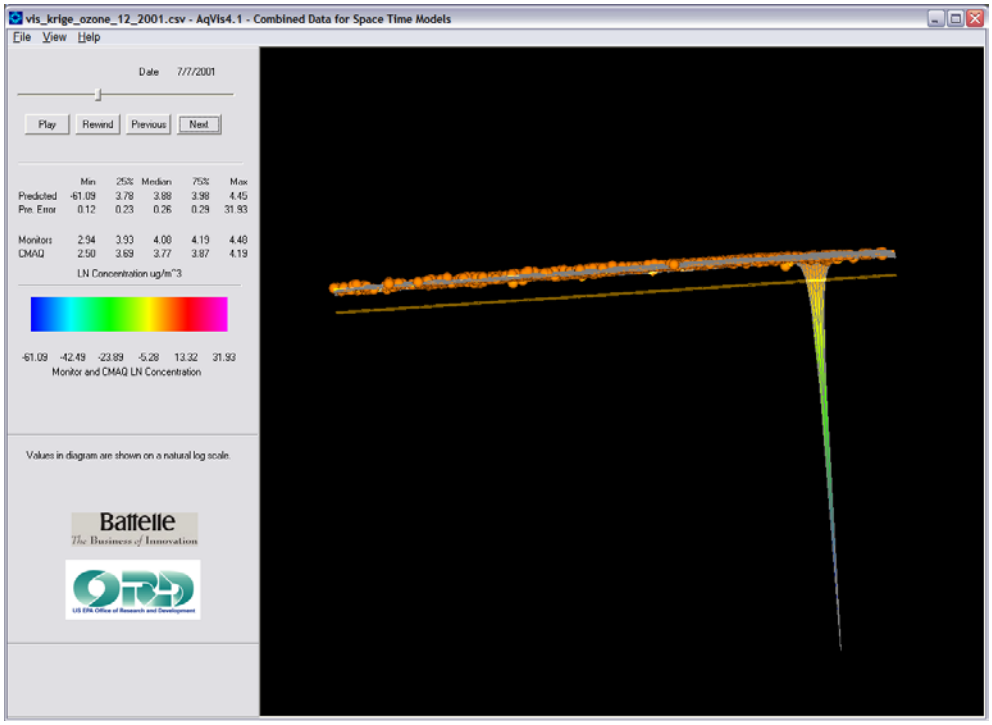


Figure 12. July 7, 2001-Anomolous Prediction.

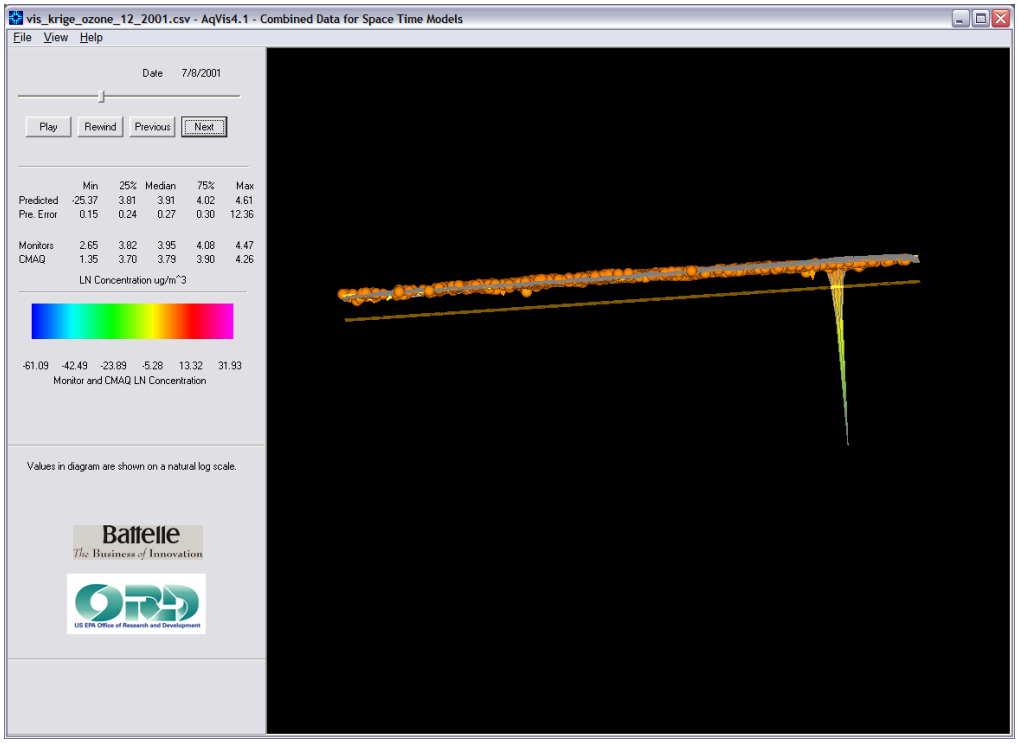


Figure 13. July 8, 2001-Anomolous Prediction.

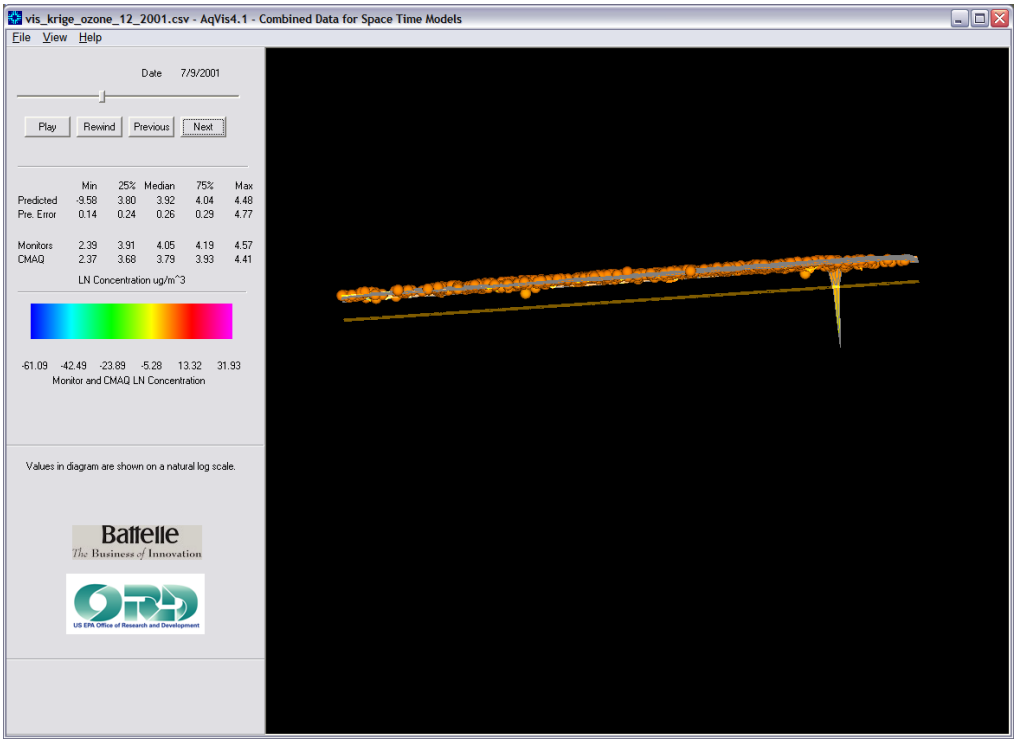


Figure 14. July 9, 2001-Anomolous Prediction.

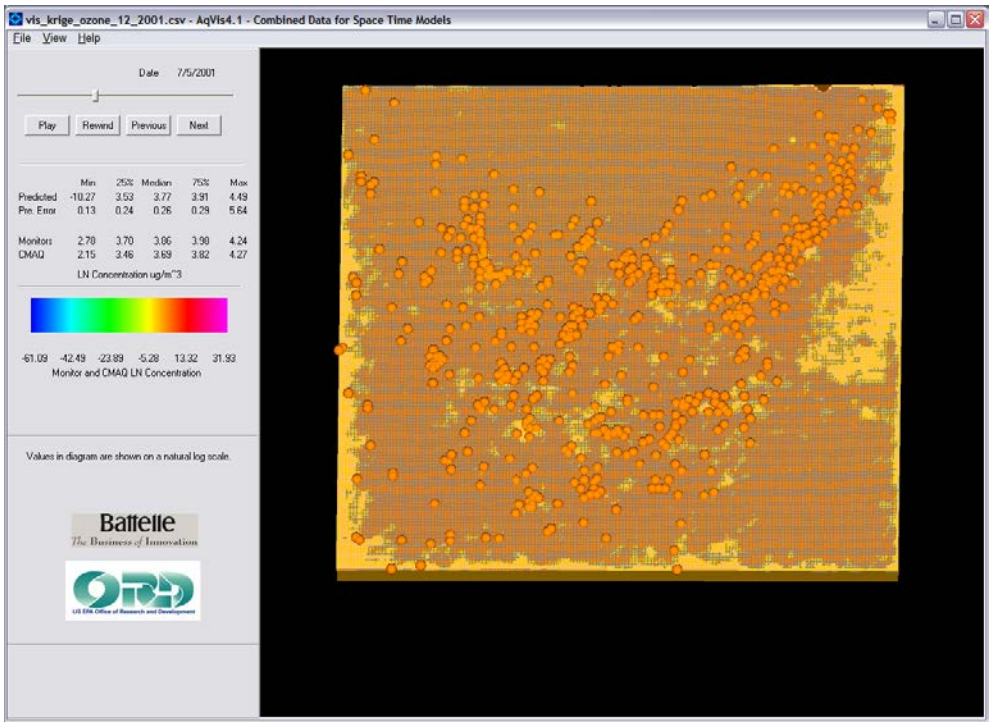


Figure 15. July 5, 2001-Anomolous Prediction – View of CMAQ "hole"

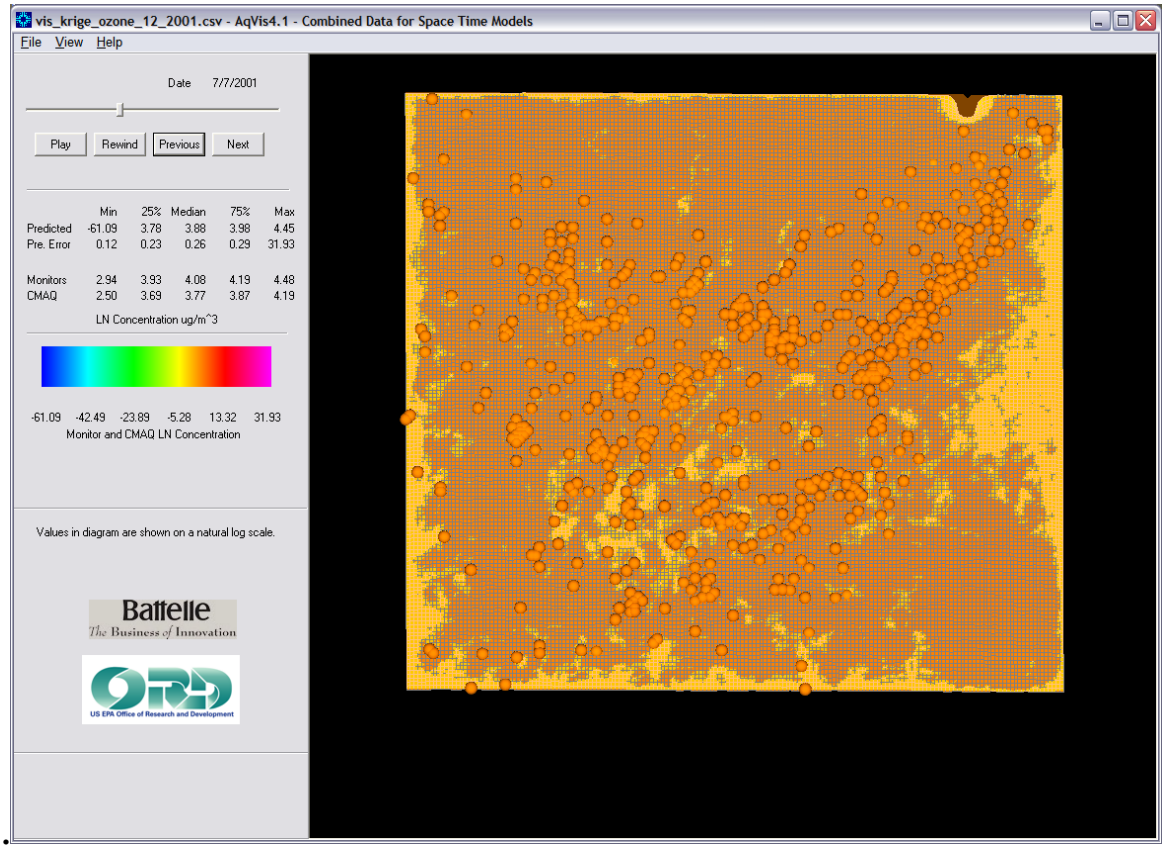


Figure 16. July 7, 2001-Anomolous Prediction-View of CMAQ "hole".

6.0 References

[1] McMillan, N.J., Holland, D.M., Morara, M., and Feng, J., Combining different sources of particulate data using Bayesian space-time modeling, Environmetrics, 2010, Volume 21, pp 48 – 65, DOI: 10.1002/env.984

[2] USEPA (2002) *AQS basics: What data are in AQS?* Presentation by David Lutz at the 2002 AQS Conference. Office of Air Quality Planning and Standards, Office of Air and Radiation, US Environmental Protection Agency.

[3] USEPA (2007) *Draft Report: Overview of EPA's Hierarchical Bayesian Model for Predicting Air Quality Patterns in the United States over Space and Time, for Use with Public Health Tracking Data*, Contract No. EPA-D-04-068, Work Assignment 44, September 28, 2007.

Appendix C

Appendix C: Markov Chain Monte Carlo (MCMC) Description (Model 5.1)

Monte Carlo Markov Chain (MCMC) Description (Model 5.1)

Data:

N^T Number of time points
N^P Number of space points
$N = N^T \times N^P$ Number of events (space-time points)

$x_i \in \mathbf{R}^{N_i^X} \quad i = 1,\ldots,N$ Monitoring data at event i

$X_i \in \mathbf{R} \quad i = 1,\ldots,N$ Sum of the monitoring data at event i (the number of monitoring data for each event *i* can be 0, 1 or more than 1).

$y_i \in \mathbf{R} \quad i = 1,\ldots,N$ CMAQ data at event i (the number of CMAQ data for each event *i* is always and only 1).

$D_{ij} \in \mathbf{R} \quad i = 1,\ldots,N \quad j = 1,\ldots,N^D$ Bias *jth* basis function evaluated at event *i*

The bias is evaluated as a linear combination of 2nd order uniform B-spline functions defined over a 3-dimensional lattice of uniform knots. The coefficients of the linear combination represent unidimensional control points, that is,

$$Bias = \sum_{j=1}^{N^D} D_{ij}\beta_j^D \ .$$

If we indicate with N_1, N_2, N_3 the dimensions of the CMAQ grid (that is, $N_1 = N^T$, $N_2 \cdot N_3 = N^P$ and $N_1 \cdot N_2 \cdot N_3 = N$), and with M_1, M_2, M_3 the dimensions of the control-points grid (this defines the degrees of freedom of the bias, that is, $M_1 \cdot M_2 \cdot M_3 = N^D$), and we decompose the indexes as: $i = i_1 + N_1(i_2 + N_2 i_3)$, $j = j_i + M_1(j_2 + M_2 j_3)$, then the bias matrix is then defined as

$$D_{ij} = b_{j_1}(i_1) b_{j_2}(i_2) b_{j_3}(i_3)$$

where $b_k(u)$ is the 2nd order *k*th B-spline basis function evaluated at the parameter point *u*. Setting $[a_1,b_1]\times[a_2,b_2]\times[a_3,b_3]$ the domain over which the bias is defined (such domain must be bigger than or equal to the CMAQ grid domain), the uniform knots vectors over which the B-spline basis functions are defined are respectively

$$U_r = \{a_r, a_r, a_r, a_r + s_r, a_r + 2s_r \ldots, a_r + (M_r - 3)s_r, b_r, b_r, b_r\} \quad r = 1,2,3$$

where $s_r = \dfrac{b_r - a_r}{M_r - 2} \quad r = 1,2,3$.

Using B-splines as basis functions for the bias allows to control the degrees of freedom of the bias through the number of control points. Furthermore, the piece-wise nature of the B-spline functions respects the principle of locality, that is, local information does not affect regions far from the region where the local information is defined. On the numerical side, B-spline allow a tensor factorization of the bias matrix into three matrixes $B^r_{i_r j_r} = b_{j_r}(i_r), r = 1,2,3$ for a total dimension of $N_1 M_1 + N_2 M_2 + N_3 M_3$, very much less than the total dimension of the full D-matrix, which is $N_1 M_1 \cdot N_2 M_2 \cdot N_3 M_3$.

Parameters

Z_i	$i = 1,\ldots,N$	CAR process at event i.
μ_t	$t = 1,\ldots,N^T$	Mean level of the CAR process.
$\beta^D \in \mathbf{R}^{N^D}$		Vector of coefficient for the bias.
τ^X		Precision of the measurement error in the monitor observations.
τ^Y		Precision of the measurement error in the computer observations.
τ^Z		Precision of the mean process.
ρ^T		Temporal autocorrelation parameter of the mean process.
ρ^P		Spatial autocorrelation parameter of the mean process.

Likelihood:

$$\left[x_{ik} \mid \mu_{t(i)}, Z_i, \tau^X\right] = N\left(w_i, \tau^X\right),$$

$$\left[y_i \mid \mu_{t(i)}, \beta^D, Z_i, \tau^Y\right] = N\left(w_i + \beta^D D_i, \tau^Y\right),$$

where

$$w_i = \mu + Z_i .$$

Priors:

$$\left[Z \mid \tau^Z, \rho^T, \rho^P\right] = N\left(0, \tau^Z\left(\Lambda^T\left(\rho^T\right) \otimes \Lambda^P\left(\rho^P\right)\right)\right)$$

where $\Lambda^T(\rho)$ is the precision matrix corresponding to a time autoregressive model with parameter ρ and Λ^P is the precision matrix corresponding to a space autoregressive model of order r with a zero boundary condition.

Writing

$$n_t^T = \begin{cases} 1 & t = 1, N^T \\ 1 + \left(\rho^T\right)^2 & 1 < t < N^T \end{cases}$$

$$n^P = \text{ number of spatial neighbors}$$

and

$$\mu_i^z = \frac{\rho^T}{n_{t(i)}^T} \sum_{j \in \partial_t i} Z_j + \frac{\rho^P}{n^P} \sum_{j \in \partial_p^r i} Z_j - \frac{\rho^T \rho^P}{n_{t(i)}^T n^P} \sum_{j \in \partial_t i \times \partial_p^r i} Z_j$$

$$\tau_i^z = \tau^Z \frac{n_{t(i)}^T n^P}{\left[1 - \left(\rho^T\right)^2\right]\left[1 - \left(\rho^P\right)^2\right]}$$

where $\partial_t i$ denotes the time first nearest neighbors' events of the event i, $\partial_p^r i$ is the set of space r nearest neighbor's events of the event i, the prior conditional distribution for a single element of Z can be written as

$$\left[Z_i \mid Z_{i-}, \tau^Z, \rho^T, \rho^P\right] = N\left(\mu_i^z, \tau_i^z\right)$$

where the minus after a subscript denotes the set of all subscripts not including the one shown.

The priors corresponding to the other parameters are

$$\left[\mu_t\right] = N\left(\theta^\mu, \tau^\mu\right)$$

$$\left[\beta_i^D\right] = N\left(\theta^D, \tau^D\right)$$

$$\left[\tau^X\right] = G\left(\gamma^X, \delta^X\right)$$

$$\left[\tau^Y\right] = G\left(\gamma^Y, \delta^Y\right)$$

$$\left[\tau^Z\right] = G\left(\gamma^Z, \delta^Z\right)$$

$$\left[\rho^T\right] \sim U\left(a^T, b^T\right)$$

$$\left[\rho^P\right] \sim U\left(a^P, b^P\right)$$

Full Model

$$\left[Z, \mu, \beta^D, \tau^X, \tau^Y, \tau^Z, \rho^T, \rho^P \mid X, Y, S\right] \propto$$

$$\prod_{i=1}^{N} \prod_{k=1}^{N_i^X} \left[x_{ik} \mid Z_i, \mu_{t(i)}, \tau^X\right] \times$$

$$\prod_{i=1}^{N} \left[y_i \mid Z_i, \mu_{t(i)}, \beta^D, \tau^Y\right] \times$$

$$\prod_{i=1}^{N}\left[Z_i \mid Z_{i-}, \tau^Z, \rho^T, \rho^P\right]\times$$

$$[\mu][\beta^D][\tau^X][\tau^Y][\tau^Z][\rho^T][\rho^P]$$

Explicitly:

$$\left[Z, \mu, \beta^D, \tau^X, \tau^Y, \tau^Z, \rho^T, \rho^P \mid X, Y\right]\propto$$

$$\prod_{i=1}^{N}\left(\frac{\tau^X}{2\pi}\right)^{\frac{N_i^X}{2}} \exp\left\{-\frac{\tau^X}{2}\sum_{i=1}^{N}\sum_{k=1}^{N_i^X}\left[x_{ik}-\left(\mu_{t(i)}+Z_i\right)\right]^2\right\}\times$$

$$\prod_{i=1}^{N}\left(\frac{\tau^Y}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\tau^Y}{2}\sum_{i=1}^{N}\left[y_i-\left(\mu_{t(i)}+Z_i+\beta^D D_i\right)\right]^2\right\}\times$$

$$\left(\tau^Z\right)^{\frac{N}{2}}\left|\Lambda^T\left(\rho^T\right)\otimes\Lambda_r^P\left(\rho^P\right)\right|^{\frac{1}{2}} \exp\left\{-\frac{\tau^Z}{2}(Z)^T\left[\Lambda^T\left(\rho^T\right)\otimes\Lambda_r^P\left(\rho^P\right)\right]Z\right\}\times$$

$$[\mu][\beta^D][\tau^X][\tau^Y][\tau^Z][\rho^T][\rho^P]$$

Full Conditional Distributions

Variable: $Z_i \in \mathbf{R}$

$$\left[Z_i \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A = \tau^X N_i^X + \tau^Y + \tau_i^z$$

$$B = \tau^X\left[X_i - N_i^X \mu_{t(i)}\right] + \tau^Y\left[y_i - \left(\mu_{t(i)} + \beta^D D_i\right)\right] + \tau_i^z \mu_i^z$$

Variable: $\mu_t \in \mathbf{R}$

$$\left[\mu_t \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A = \tau^X \sum_{i\in I_t} N_i^X + \tau^Y N^T + \tau^\mu$$

$$B = \tau^X \sum_{i\in I_t}\left[X_i - N_i^X Z_i\right] +$$

$$\tau^Y \sum_{i\in I_t}\left[y_i - \left(Z_i + \beta^D D_i\right)\right] + \tau^\mu \theta^\mu$$

Variable: $\beta^D \in \mathbf{R}^{N^D}$

$$\left[\beta^D \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A_{kl} = \tau^Y \sum_{i=1}^{N} N_i^Y D_{ik} D_{il} + \delta_{kl}\tau^D$$

$$B_k = \tau^Y \sum_{i=1}^{N} D_{ik}\left[y_i - \left(\mu_{t(i)} + Z_i\right)\right] + \tau^D \theta^D$$

Variable: τ^X

$$\left[\tau^X \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}\sum_{i=1}^{N} N_i^X + \gamma^X$$

$$B = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N_i^X}\left[x_{ik} - \left(\mu_{i(t)} + Z_i\right)\right]^2 + \delta^X$$

Variable: τ^Y

$$\left[\tau^Y \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}N + \gamma^Y$$

$$B = \frac{1}{2}\sum_{i=1}^{N}\left[y_i - \left(\mu_{t(i)} + Z_i + \beta^D D_i\right)\right]^2 + \delta^Y$$

Variable: τ^Z

$$\left[\tau^Z \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}N + \gamma^z$$

$$B = \frac{1}{2\tau^Z}\sum_{i=1}^{N}\tau_i^z Z_i\left(Z_i - \mu_i^z\right) + \delta^Z$$

Variable: ρ^T

$$\left[\rho^T \mid -\right] \propto \chi_{\left(a^T, b^T\right)}\left(\rho^T\right)\left(1-\left(\rho^T\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^T,\rho^P\right)Z_i\left(Z_i - \mu_i^z\left(\rho^T,\rho^P\right)\right)\right\}$$

This full conditional is not a recognized form, so it has to be sampled using a Metropolis-Hastings step.

Jump:

$$\rho' \sim J\left(\left|\rho' - \rho^T\right|\right)$$

Acceptance criteria:

$$\min\left\{\frac{\chi_{\left(a^Z, b^Z\right)}\left(\rho'\right)\left(1-\left(\rho'\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho',\rho^P\right)Z_i\left(Z_i - \mu_i^z\left(\rho',\rho^P\right)\right)\right\}}{\left(1-\left(\rho^T\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^T,\rho^P\right)Z_i\left(Z_i - \mu_i^z\left(\rho^T,\rho^P\right)\right)\right\}}, 1\right\}$$

Variable: ρ^P

Let

$$\lambda_i = \frac{1}{2}\left\{\cos\left(\frac{\pi}{N_x+1}i_x\right) + \cos\left(\frac{\pi}{N_y+1}i_y\right)\right\}$$

$$D\left(\rho^P\right) = \prod_{i=1}^{N^P}\left(\frac{1-\rho^P\lambda_i}{1-\left(\rho^P\right)^2}\right)$$

$$\left[\rho^P \mid -\right] \propto \chi_{\left(a^P, b^P\right)}\left(\rho^P\right)D\left(\rho^P\right)^{-\frac{1}{2}N^T} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^T,\rho^P\right)Z_i\left(Z_i - \mu_i^z\left(\rho^T,\rho^P\right)\right)\right\}$$

This full conditional is not a recognized form, so it has to be sampled using a Metropolis-Hastings step.

Jump:

$$\rho' \sim J\left(\left|\rho' - \rho^P\right|\right)$$

Acceptance criteria:

$$\min\left\{\frac{\chi_{\left(a^P,b^P\right)}\left(\rho'\right)D\left(\rho'\right)^{-\frac{1}{2}N^T}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^T,\rho'\right)Z_i\left(Z_i-\mu_i^z\left(\rho^T,\rho'\right)\right)\right\}}{D\left(\rho^P\right)^{-\frac{1}{2}N^T}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^T,\rho^P\right)Z_i\left(Z_i-\mu_i^z\left(\rho^T,\rho^P\right)\right)\right\}},1\right\}$$

Appendix D

Appendix D: SAS Code for Developing the Krige Model Estimates

```
*****************************************************************************
***************
Program Name:          Krige Ozone Surface_2001_12.sas
* Note -
* Getting a useful empirical variogram is an iterative process
* Be sure to plot tempvariogramestiamte predicted vs. distance to verify that
* the distance bins reach a distance at which the variogram is at the sill (where it levels off)
* and with distance bin sizes small enough to adequately represent the variogram
* but not so small that there are too few distance pairs in any particular bin
* for example:
* lagdistance used for PM2.5 2001 emp. variogram = 5;* lagdistance used for O3 2002 emp.
variogram = 10;
* and each year used 20 distance bins
*****************************************************************************
***************;
%let cooperpath=C:\OAQPS\Data;
%let otherpath=C:\OAQPS\Data;

%let path1=&cooperpath.;
%let path2=&cooperpath.;

%let monitor=Monitor_O3_2001_12.csv;
%let castnet=castnet_grid_o3_2001_12km;

%let krigeerror=krigepredozoneerror122001;
%let krigedata=krige_ozone_12_2001;

Libname sd "&path1";
libname sd2 "&otherpath";

*-------------------------
* Importing FRM Monitoring data
*-------------------------;
*--- AQS Ozone Data;
data ozone;
      infile "&path2.\&monitor" missover delimiter=',';
      input time col row logo3;
      if time = . then delete;
proc sort data=ozone(rename=(logo3=value)) out=ozone; * was "Log(ozone)";
   by time col row;
run;


*-------------------------
* Construct variogram
*-------------------------;
```

```
%macro Calcvario(ds,day);
      proc datasets;
        delete tempoz t_vario tempmean tempoz2;

      *--- Get single days data;
      data tempoz;
     set &ds;
     if time=&day.;

      *--- Calculate empirical variogram values;
      proc variogram data = tempoz outvar = t_vario;
     compute lagdistance = 10 maxlags = 20;
     coordinates xcoord = col ycoord = row;
     var value;

      *--- Count number of observations for the day;
      data tempoz;
     set tempoz;
     unify = _n_;
      proc means data=tempoz noprint;
     var unify;
     output out=tempmean;
      data tempmean;
     set tempmean;
     if (_stat_ = "MAX");
      data _NULL_;
     set tempmean;
     call symput('nobsa', trim(left(unify)));

      *--- Repeat dataset nobsa times;
      data tempoz2;
     do unify = 1 to &nobsa.;
        output;
     end;
      data tempoz2;
            set tempoz;
      do i=1 to max;
                  set tempoz2 nobs = max point = i;
            output;
      end;
      proc sort data=tempoz2;
     by unify;

      *--- Add squared differences of values within same cell to variogram dataset;
      data tempoz2;
     set tempoz2;
```

```
      rename col = x;
      rename row = y;
      rename value = val;
       data tempoz;
      merge tempoz tempoz2;
      by unify;
      if ((x = col) & (y = row) & (val ^= value));
      sqdiff = (val - value)**2;
       proc means data=tempoz noprint;
      var sqdiff;
      output out=tempmean mean=sqdiff;
       data tempmean;
      set tempmean(rename=(_FREQ_=count));
         variog=sqdiff/2;
         distance=0;
      keep distance variog count;
       data t_vario;
      set tempmean t_vario;
         if variog^=.;
      keep distance variog count;

       *--- Append this days empirical variogram to rest;
       data allvario;
      set allvario t_vario;
       run;
%mend Calcvario;

%macro Loopvario(ds,DatasetIn);
    data allvario;
      delete;
       run;
       %let DatasetID = %sysfunc(Open(&DatasetIn.));
       %syscall set(DatasetID);
       %let ReturnVal = %sysfunc(fetch(&DatasetID));
       %do %while (&ReturnVal = 0);
              %CalcVario(&ds.,&time.);
              %let ReturnVal = %sysfunc(fetch(&DatasetID));
       %end;
       %let ReturnVal = %sysfunc(Close(&DatasetID));
%mend Loopvario;

*--- Construct a variogram, but subdivide by day;
data datefile;
   do time = 1 to 365 by 1;
      output;
   end;
```

```
run;
quit;
%Loopvario(ozone,datefile);

*--- Fit an Exponential Semivariogram Model with nugget;
ods output ParameterEstimates = tempParameterEstimates;
ods exclude ANOVA ConvergenceStatus CorrB EstSummary IterHistory MissingValues
ParameterEstimates;
proc nlin data=allvario MaxIter=10000;
      parms Nugget = 1, Scale = 200, Range = 2;
      bounds Nugget > 0, Scale > 0, Range > 0;

      model variog = Nugget + (Scale - Nugget) * (1 - exp(-(Distance)/(Range)));
      output Out=tempVariogramEstimate Predicted=PredictedValue;
run;
proc transpose data=tempParameterEstimates out=tempParameterEstimates;
      id Parameter;
      var Estimate;
run;

data _NULL_;
   set tempParameterEstimates;
   call symput('time',trim(left("126")));
   if Nugget^=. then call symput('NuggetIn',trim(left(Nugget)));
   else call symput('NuggetIn',0);
   call symput('RangeIn',trim(left(Range)));
   call symput('ScaleIn',trim(left(Scale)));
run;

*--------------
* Krige macro
*--------------;
%macro krige(ds,day);
   data forkrige;
      set &ds;
      where time =&day.;
   proc krige2d data=forkrige outest=nugget;
      predict var=value;
      model Form=Exponential Nugget=&NuggetIn.  Range=&RangeIn.  Scale=&ScaleIn.;
*      model Form=Gaussian Nugget=&NuggetIn.  Range=&RangeIn.  Scale=&ScaleIn.;
      coord xc=col yc=row;
      grid griddata=predlocs xc=col yc=row;
   data merged;
      set nugget(rename=(estimate=krige_predict));
      time =&day.;
      keep gxc gyc krige_predict stderr time;
```

```
   data krige_predict;
      set krige_predict merged;
   run;
%mend krige;

%macro Loopkrige(ds,DatasetIn);
   data krige_predict;
      delete;
   run;
   %let DatasetID = %sysfunc(Open(&DatasetIn.));
   %syscall set(DatasetID);
   %let ReturnVal = %sysfunc(fetch(&DatasetID));
   %do %while (&ReturnVal = 0);
%put Loopkrige time = &time.;
      %krige(&ds.,&time.);
      %let ReturnVal = %sysfunc(fetch(&DatasetID));
   %end;
   %let ReturnVal = %sysfunc(Close(&DatasetID));
%mend Loopkrige;

*--------------------
* Bring in the CASTNET data for ozone
*--------------------;
data ozone_validation;
       set sd2.&castnet;
proc sort data=ozone_validation out=unique_site nodupkey;
       by col row;
data predlocs;
    set unique_site;
       keep row col ;
run;

%Loopkrige(ozone,datefile);

data sd2.&krigeerror;
       set krige_predict;
       rename gxc=col gyc=row;
run;

proc export data=krige_predict outfile="&otherpath.\&krigedata..csv" replace; run;
```

Appendix E

Appendix E: Simulation.PAR (2001, 12 km, Ozone-Run 1)

```
# Simulation parameter file

5-4H         # Model version

1, 333         # Grid Size t
40, 213        # Grid Size x
30, 180        # Grid Size y

2001, 01, 01  # Year, month, day of time step 1

1,     333,        4         # Bias Spline in t
40,      213,      8         # Bias Spline in x
30,      188,      7         # Bias Spline in y

0, 1.0E-3     # Prior Parameters for Mu ~    Normal(mean, prec)
0, 1.0E-3     # Prior Parameters for BetaD ~ Normal(mean, prec)

7.5E+7, 1.0E+6              # Prior Parameters for TauX  ~ Gamma(mean, prec)
1.0E+7, 1.0E+8              # Prior Parameters for TauY  ~ Gamma(mean, prec)
1.0E-3, 1.0E-3              # Prior Parameters for TauZ  ~ Gamma(mean, prec)

0, 1          # Prior Parameters for RhoZ  ~ Gamma (shape, rate)

C:\                         # Directory where data are stored
aqs_12km_o3_2001.csv        # File name where monitor data X are stored
cctm_12km_2001.csv          # File name where CMAQ csv is stored

1             # Order of neighborhood for Z (between 1 and 4)
0             # Track chain flag for Z

1             # Enable calculation of 4th highest statistics
0             # Track chain flag for 4th highest statistics

C:\                         # Directory where the simulation outputs are

written
Surface_03_12_2001.csv      # File name where mean, covariate, bias,

monitoring, and computer average surfaces are written
Surface_03_12_2001_4h.csv   # File name where 4th highest surface is written

1234          # Random seed 1
5678          # Random seed 2

1             # Sampling period for thinning
1             # Prompting period for diagnostic
1             # New/Load simulation flag. 1 = New, 0 = Load
50            # Number of burn-in steps
100           # Number of simulation steps
```

Appendix F

Appendix F: SAS Code for Generating the Summary Statistics

```
*-----------------------------------------------------*
Program Name:     Summarize model parameters-Ozone_12_2001.SAS
*====> Please change the following to match your system;
%macro CallVal(RunNum, SimFile, ValMonFile, KrigePredErrFile);
      %let LocModelOutFile=C:\OAQPS\Ozone_2001_12\&RunNum;
      %let LocMonitorValDat=C:\OAQPS\Data;
      * HBM data*;
      %let inchainfile=Chain.csv;
      %let modelsurfacefile=&SimFile..csv;

      * Edit the date specifying the day before the start date of the calendar year
      * 'afterDate' is the day before the start date;
      %let afterDate = '31dec2000'd;

      %let ttl=Ozone 2001 12 km - &RunNum;
      %let minx=40;
      %let maxx=213;
      %let miny=30;
      %let maxy=188;

      * Kriged data*;
      %let krigedsurface=&KrigePredErrFile;

      * Monitor data:  improve and stn for PM2.5 and CASTNet for O3*;
      %let castnet=&ValMonFile;

      %let monrdg = o3_8hrmax;
      %let logrdg = lno3_8hrmax;
      %let o3Valcols=   &monrdg. &logrdg.;
      %let Valcols = &o3Valcols.;

      libname sd "&LocModelOutFile";
      libname sd2 "&LocMonitorValDat";

      *-------------------------------------------------------------
      * Calculate the summary statistics on the HBM model parameters
      *-------------------------------------------------------------;
      proc import out    = chain
                    datafile = "&LocModelOutFile.\&inchainfile"
            dbms = csv replace;

      * tau is the precision parameter in HBM
        x denotes FRM monitor precision
        y denotes CMAQ 'precision'
        z denotes precision for Conditional autoregressive (CAR) process
      These are estimated from the characteristics of the HBM Markov chain
```

```
	setaux is se_taux; ;
ods rtf file= "&LocModelOutFile.\PerformanceOutput_&runnum..rtf";
	proc univariate data = chain plot;
		var taux tauy tauz;
		output out = sumparms
			mean = mntaux mntauy mntauz
			stderr = setaux setauy setauz
			n = ntaux ntauy ntauz;

	proc print data = sumparms;
		title "Summary Statistics for &ttl";

	* posterior variance to be computed from x and z precision ******;
	data _null_;
	  set sumparms;
	  call symput("sigmax",1/mntaux);
	  call symput("sigmaz",1/mntauz);
	run;

	%put &sigmax;
	%put &sigmaz;

	*--- Comparison summaries ---*;
	* Read in HBM predicted surface, referred to as modelsurface;
	proc import out    = hbm
			datafile = "&LocModelOutFile.\&modelsurfacefile"
		dbms = csv replace;
	data monhbm;
		set hbm;
		where monitordata > 0;
	run; quit;

	data hbm_surfacepreds(keep   = time xcoord ycoord predavg predstd computerdata test
monitordata
				rename = (xcoord=col ycoord=row predAvg=predsurface
ComputerData=CMAQSurface));
	  set hbm;
	  *--- Remove the data where a prediction was not made ---*;
	  time=time-&afterDate.;
	*   time=time-14975-365;
	    if predAvg ^= -999;
	  test = predstd*predstd;
	  *--- This step exponentiates the results back to normal units ---*;
	  array vars [4] predAvg biasAvg monitorData computerData;
	  do i =1 to 4;
	    if vars[i] ne -999 then vars[i]=exp(vars[i]);
```

```
      end;
   *--- The bias is transformed ---*;
   if vars[2] ne -999 then vars[2]=(vars[2]-1)*vars[1];
   drop i;
   *--- For the validation results, any monitor result greater than 600 truncated to 600 ---*;
   if vars[3]>600 then vars[3]=-999;
   if index(var12,'0D'x)=0 then covardata=input(var12,8.);
   else covardata=input(substr(var12,1,index(var12,'0D'x)-1),8.);
run; quit;

proc sort data = hbm_surfacepreds;
   by time col row;
data check;
      set hbm_surfacepreds;
      diffpred_CMAQ = predsurface-cmaqsurface;
      if monitordata > 0 then diffpred_mon = predsurface-monitordata;
                              else diffpred_mon = .;
proc sort data = check;
      by time;
proc means data = check noprint;
      by time;
      var diffpred_CMAQ diffpred_mon;
      output out = sumt(drop = _type_ _freq_)
               mean = mnCMAQdiff mnMondiff
               stderr = seCMAQdiff seMondiff
               n = nCMAQdiff nMondiff;
run; quit;

proc sort data = check;
      by col row;
proc means data = check noprint;
      by col row;
      var diffpred_CMAQ diffpred_mon;
      output out = sumcr(drop = _type_ _freq_)
               mean = mnCMAQdiff mnMondiff
               stderr = seCMAQdiff seMondiff
               n = nCMAQdiff nMondiff;
data moncr;
      set sumcr;
      where mnmondiff > 0;
run; quit;

*--- Kriging data for comparison ---*;
data krige_predict;
  set sd2.&krigedsurface.;
  krigepred = exp(krige_predict);
```

```
run; quit;

proc sort data=krige_predict;
   by time col row;
run;

*--- IMPROVE and STN, or CASTNet, data for comparison ---*;
data surf_validation (keep = time col row site &Valcols.);
  set sd2.&castnet. (rename=(siteID= site));

  where (col ge &minx.) and (row ge &miny.) and (row le &maxx.);
run; quit;

proc sort data = surf_validation;
 by time col row;
run;
*-----------------------------------------------------------------------------
*     Combine the model prediction, Kriging results, and the IMPROVE/STN data
*   NOTE:  The bias is calculated as bias = validation value - predicted value
*-----------------------------------------------------------------------------;
data combined;
merge     krige_predict
          surf_validation (in = invalidation)
        hbm_surfacepreds (in = inpred);
   by time col row;

   *--- Data to keep in the file ---*;
   if invalidation;
   if predsurface^=.;

   if CMAQsurface = -999 then CMAQsurface = .;
   if predsurface = -999 then predsurface = .;

   *--- Calculate the summary stats ---*;
   *--- regular units ---*;
   mse_model = (predsurface-&monrdg.)**2;
   mse_krige = (krigepred-&monrdg.)**2;
   mse_CMAQ = (CMAQsurface-&monrdg.)**2;

   mse_diff = mse_krige-mse_model;
   mse_diff_2 = mse_CMAQ-mse_model;

   bias_model = &monrdg.-predsurface;
   bias_krige = &monrdg.-krigepred;
   bias_CMAQ = &monrdg.-CMAQsurface;
```

```
      *--- log units ---*;
   mse_lmodel = (log(predsurface)-&logrdg.)**2;
   mse_lkrige = (krige_predict-&logrdg.)**2;
   mse_lCMAQ  = (log(CMAQsurface)-&logrdg.)**2;

   * compare MSE of Krige to HBM and of CMAQ to HBM;
   mse_ldiffKrige = mse_lkrige - mse_lmodel;
   mse_ldiffCMAQ  = mse_lcmaq - mse_lmodel;

   bias_lmodel = &logrdg. - log(predsurface);
   bias_lkrige = &logrdg. - krige_predict;
   bias_lCMAQ  = &logrdg. - log(CMAQsurface);
      *--  - Summarize performance by time ---*;
proc sort data = combined out=combined_oz;
   by time;
run;

*--- Average daily MSE for HBM, Kriging, and CMAQ ---*;
* avg mse_ldiff... > 0 implies HBM error is smaller on average;
* se_mse... provides measure of variation in the magnitude of square errors;


proc means data = combined_oz noprint;
   by time;
   var mse_lmodel mse_lkrige mse_lcmaq mse_ldiffkrige mse_ldiffcmaq
         bias_lmodel bias_lkrige bias_lCMAQ;
   output out = mse_oz (drop = _type_ _freq_)
         mean = mse_lmodel mse_lkrige mse_lcmaq
               mse_ldiffkrige mse_ldiffcmaq
                     bias_lmodel bias_lkrige bias_lCMAQ
         stderr = se_mse_lmodel se_mse_lkrige se_mse_lcmaq
                        se_mse_ldiffkrige se_mse_ldiffcmaq
                        se_bias_lmodel se_bias_lkrige se_bias_lCMAQ
            n = n_mse_lmodel n_mse_lkrige n_mse_lcmaq
             n_mse_ldiffkrige n_mse_ldiffcmaq
                   n_bias_lmodel n_bias_lkrige n_bias_lCMAQ;

proc print data = mse_oz;
         var mse_lmodel mse_lkrige mse_lcmaq
                   mse_ldiffkrige mse_ldiffcmaq
                   bias_lmodel bias_lkrige bias_lCMAQ
            se_mse_lmodel se_mse_lkrige se_mse_lcmaq
               se_bias_lmodel se_bias_lkrige se_bias_lCMAQ
            n_mse_lmodel;
         format mse_lmodel mse_lkrige mse_lcmaq
                     bias_lmodel bias_lkrige bias_lCMAQ
```

```
                    se_mse_lmodel se_mse_lkrige se_mse_lcmaq
                 se_bias_lmodel se_bias_lkrige se_bias_lCMAQ
                 6.4;
                title1 'Average Daily MSE, Bias, and number of observations';
    run; quit;

*--- Annual MSE for HBM, Kriging, and CMAQ.  Bias summary for HBM, kriging, and CMAQ
---*;
proc means data=combined_oz noprint;
  var mse_lmodel mse_lkrige mse_lcmaq mse_ldiffkrige mse_ldiffcmaq
      bias_lmodel bias_lkrige bias_lCMAQ;
  output out = overall_mse_oz (drop = _type_ _freq_)
      mean = mse_lmodel mse_lkrige mse_lcmaq
            mse_ldiffkrige mse_ldiffcmaq
                  bias_lmodel bias_lkrige bias_lCMAQ;
proc print data = overall_mse_oz label;
    var mse_lmodel mse_lkrige mse_lcmaq
       bias_lmodel bias_lkrige bias_lCMAQ;
    label mse_lmodel = 'Overall MSE HBM'
          mse_lkrige = 'Overall MSE Krige'
          mse_lcmaq = 'Overall MSE CMAQ'
          bias_lmodel = 'Overall Bias HBM'
          bias_lkrige = 'Overall Bias Krige'
          bias_lcmaq = 'Overall Bias CMAQ';
    format mse_lmodel mse_lkrige mse_lcmaq
          bias_lmodel bias_lkrige bias_lCMAQ 5.2;
run; quit;

*--- Summarize performance by site ---*;
proc sort data = combined_oz;
  by site;
proc means data = combined_oz noprint;
  by site;
  var mse_lmodel mse_lkrige mse_lcmaq
      mse_ldiffkrige mse_ldiffcmaq
    bias_lmodel bias_lkrige bias_lCMAQ;
  output out = mse_oz2
      mean = mse_lmodel mse_lkrige mse_lcmaq
                  mse_ldiffkrige mse_ldiffcmaq
            bias_lmodel bias_lkrige bias_lCMAQ;
proc print data = mse_oz2;
    var mse_lmodel mse_lkrige mse_lcmaq
     bias_lmodel bias_lkrige bias_lCMAQ;
    title1 'Average MSE by site for O3';

*--- Comparing HBM to CMAQ and Krige based on MSE ---*;
```

```
* mse_oz is daily statistics (by 'time’);
data mse_oz1;
  set mse_oz;
  retain sum_1 sum_2;

  if mse_ldiffkrige > 0 then do;
    krige_larger=1;
       sum_1+1;
  end;

  if mse_ldiffcmaq > 0 then do;
   cmaq_larger=1;
   sum_2+1;
  end;
proc means data = mse_oz1 noprint;
       var sum_1 sum_2 time;
       output out = Improve_Time(drop = _type_ _freq_ t1 t2)
                max = krigeworse CMAQworse
                n   = t1 t2 ndays;
data Improve_time;
       set Improve_time;
       Pct_Time_Krige_worse = (krigeworse / ndays) * 100;
       Pct_Time_CMAQ_worse = (cmaqworse / ndays) * 100;
proc print data = improve_time;
       title1 "&ttl: % Improvement over Kriging and CMAQ";
       var ndays pct_time_krige_worse krigeworse pct_time_CMAQ_worse cmaqworse;
       format pct_time_krige_worse pct_time_CMAQ_worse 7.2;
run; quit;

* mse_oz2 is statistics by site;
data mse_oz3;
  set mse_oz2;
  retain sum_1 sum_2;
  if mse_ldiffkrige > 0 then do;
    krige_larger=1;
       sum_1+1;
  end;
  if mse_ldiffcmaq>0 then do;
   cmaq_larger=1;
   sum_2+1;
  end;
  run; quit;
proc means data = mse_oz3 noprint;
       var sum_1 sum_2 _freq_;
       output out = Improve_site(drop = _type_ _freq_ t1 t2)
                max = krigeworse CMAQworse
```

```
                n   = t1 t2 nsites;
data Improve_site;
      set Improve_site;
      Pct_site_Krige_worse = (krigeworse / nsites) * 100;
      Pct_site_CMAQ_worse = (cmaqworse / nsites) * 100;
proc print data = improve_site;
      title1 "&ttl: % Improvement over Kriging and CMAQ";
      var nsites pct_site_krige_worse krigeworse pct_site_CMAQ_worse cmaqworse;
      format pct_site_krige_worse pct_site_CMAQ_worse 7.2;
run; quit;
*-----------------------------------------------------------------------------------------
* 95% Prediction Interval: this interval is calculated using modelstd=sqrt(sigmaz+sigmax)
* MCMC Prediction Interval: this interval is calculated using
modelstd=sqrt(predstd*predstd+sigmax)
* here begin to calculate the prediction interval
*-----------------------------------------------------------------------------------------;
* pm_surfacepreds is HBM predicted surface;
* sigmax is posterior variance of FRM data;
* sigmaz is posterior variance of CAR process;
data compare;
   merge krige_predict(keep=time col row krige_predict stderr)
        hbm_surfacepreds(keep=time col row predsurface cmaqsurface predstd)
        surf_validation(in=Validation);
   by time col row;
 if Validation;
 if predsurface ^=.;
     sigmax=&sigmax;
       sigmaZ=&sigmaz;
  *  modelstd=sqrt(predstd*predstd+sigmax);
    modelstd=sqrt(sigmaz+sigmax);

* keep track of confidence interval coverage of prediction vs. monitor measurement;
 if ((&logrdg.>log(predsurface)-2*modelstd) and (&logrdg.<log(predsurface)+2*modelstd)) then
modelCI=1;
 else modelCI=0;
 if ((&logrdg.>krige_predict-2*stderr) and (&logrdg.<krige_predict+2*stderr)) then krigeCI=1;
 else krigeCI=0;
proc means data = compare noprint;
  var modelCI krigeCI predstd;
  output out = PredictionInterval(drop=_TYPE_ _FREQ_)
        mean = modelCI krigeCI predstd;
      proc print data = predictioninterval;
            var modelCI krigeCI predstd;
            title1 'Average Prediction Interval for O3';
      run; quit;
ods rtf close;
```

```
*proc datasets library=work kill;
run; quit;
%mend CallVal;

*%CallVal(Run1, simulation_o3_2001_12_run1, castnet_grid_o3_2001_12km,
krigepredozoneerror122001);
*%CallVal(Run2, simulation_o3_2001_12_run2, castnet_grid_o3_2001_12km,
krigepredozoneerror122001);
*%CallVal(Run3, simulation_o3_2001_12_run3, castnet_grid_o3_2001_12km,
krigepredozoneerror122001);
*%CallVal(Run4, simulation_o3_2001_12_run4, castnet_grid_o3_2001_12km,
krigepredozoneerror122001);
%CallVal(Run5, simulation_o3_2001_12_run5, castnet_grid_o3_2001_12km,
krigepredozoneerror122001);
```

Appendix G

T-SpACE User's Guide

Hierarchical Bayesian Space-Time Modeling of Air Pollution Data

January, 2018
Version 5_4h.0.1.21

Table of Contents

1.0 INTRODUCTION

EPA has developed a method to combine air pollution monitoring data with air quality model output to maximize the advantages offered by both data sources, while minimizing the disadvantages of each, when predicting fine particulate matter ($PM_{2.5}$) and ozone (O_3) concentrations. Air quality models estimate the spatial and temporal gradients of air pollution based on emissions inventories and meteorological information. These models, while providing estimates over large regions at relatively low cost, have been found to be biased and to have greater error than air pollution monitoring networks. Air pollution monitoring networks generally provide relatively accurate, unbiased results. However, because air pollution monitoring networks are sparsely and irregularly spaced over large spatial domains, they are not able to produce large-scale predictions of air pollutant concentrations.

EPA has developed a Temporal-Spatial Ambient Concentrator Estimator (T-SpACE) model that combines air pollution monitoring data from the 2,545 National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) with estimates from the Models-3/Community Multiscale Air Quality (CMAQ) and the Comprehensive Air Quality Model with Extensions (CAMx) air quality models to maximize the advantages offered by both data sources, while minimizing the disadvantages of each. T-SpACE requires data preparation and model parameter input from the user to run a simulation to produce an estimated air pollution concentration surface. To streamline the set-up of the T-SpACE model, EPA developed the software and the associated guidance described in this User's Guide to allow the user to prepare the input data for the model run, choose the model specifications, run the model, receive output files containing the estimated air pollution concentration surface, and run a set of validation procedures to compare the T-SpACE run with a standard kriging method, and the CMAQ/CAMx results.

2.0 MODEL DATA

Currently, the T-SpACE model requires two input data sources, air pollution estimates from the Models-3/Community Multiscale Air Quality (CMAQ) or the Comprehensive Air Quality Model with Extensions (CAMx) air quality models, and air pollution monitoring data from the NAMS/SLAMS air pollution monitoring network. Each of these two data sources offers unique advantages and disadvantages. One strong advantage of air pollution monitoring data is that it provides information at specific spatial locations. Another advantage of air pollution monitoring data is that the data tend to be more accurate and exhibit less bias than comparable CMAQ/CAMx output. However, air pollution monitoring data do not provide information about every location within a spatial domain, since it is impossible to know exactly what an air pollution monitor would have recorded in a currently unmonitored location. In contrast to air pollution monitoring data, CMAQ/CAMx data does provide information across an entire domain of interest, but CMAQ/CAMx data can provide estimates of air pollutant concentrations. However, CMAQ/CAMx data are averaged over square grid cells, so the spatial information is not as detailed as that obtained with air pollution monitors. Also, CMAQ/CAMx data can be biased in some cases, due to the fact that the accuracy of the results can be strongly influenced by the input emissions inventories and meteorological data.

Under the CMAQ modeling system, grid cells can either be 36 km x 36 km (which represents the "parent domain" covering the entire continental US) or 12 km x 12 km (which represents primarily the eastern and Midwest regions of the US up to 2007, and the entire US from 2008 and onward). The 2001 and 2002 $PM_{2.5}$ and O_3 CMAQ data have been assessed with the CMAQ model. Figures 1, 2, and 3 illustrate the area of coverage for the 2001 12 km CMAQ data, 2001 36 km CMAQ data, and the 2002 12 km and 36 km CMAQ data, respectively.

Figure 1. The maximum area of the United States covered by the 2001 CMAQ data 12 km grid.

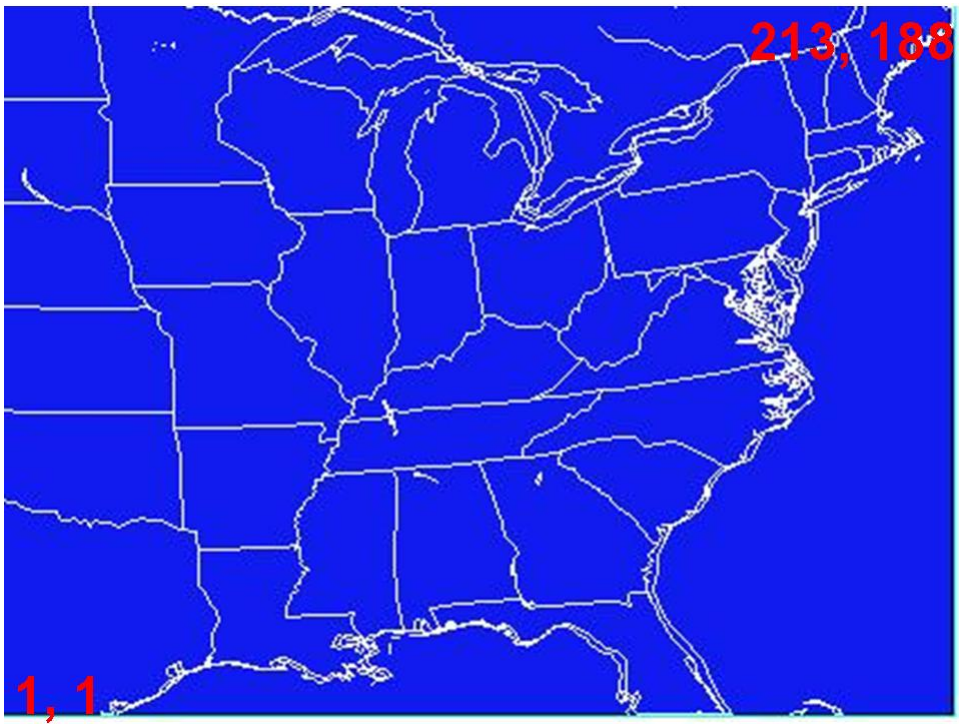


Figure 2. The maximum area of the United States covered by the 2001 CMAQ data 36 km grid.

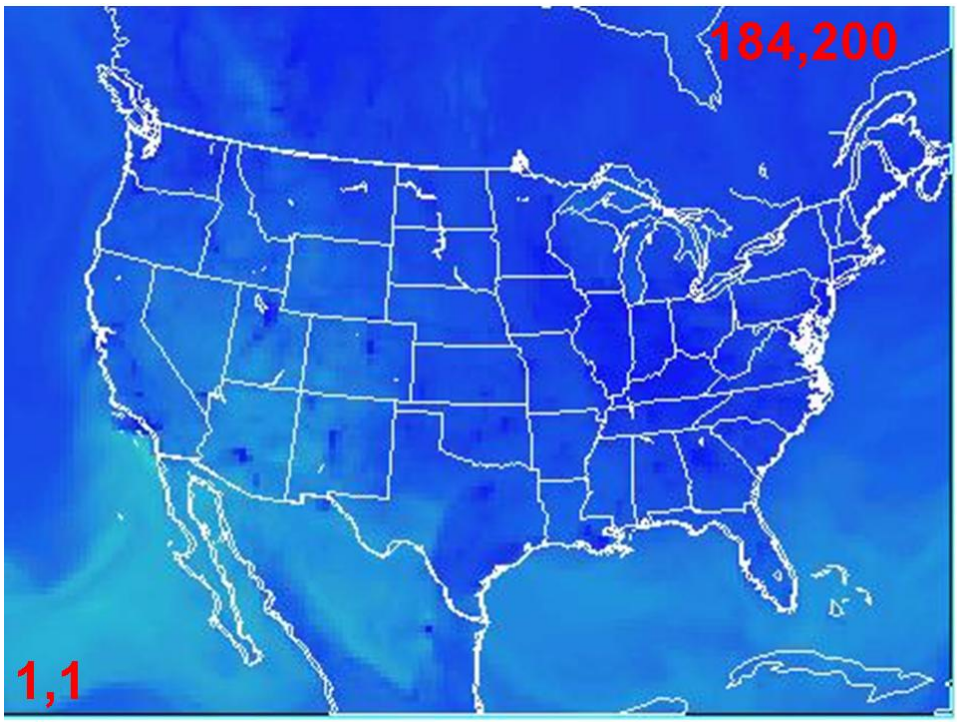
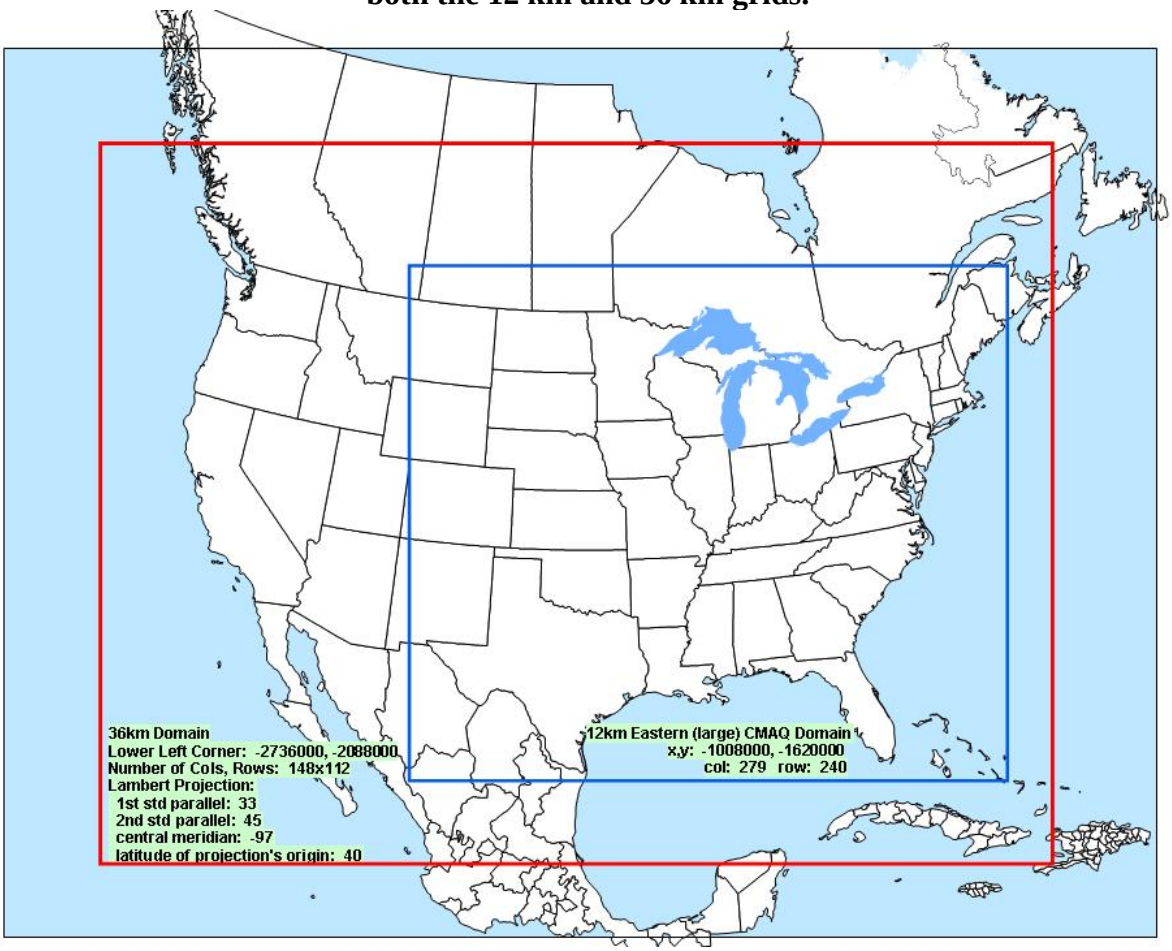


Figure 3. The maximum area of the United States covered by the 2002 CMAQ data for both the 12 km and 36 km grids.



The CMAQ computer model uses emissions information and meteorological data as inputs and calculates, via dispersion modeling, predicted 24-hour integrated concentrations. The grid cells are approximately 36 km × 36 km for the 36 km file and 12 km x 12 km grids for the 12 km file. Detailed information on the CMAQ modeling system is available at https://www.epa.gov/cmaq and at https://www.cmascenter.org.

The monitor data consists of daily concentrations that are collected from Federal Reference Method (FRM) samplers within the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) networks. Only those concentrations with a sample duration of one day (24-hour integrated sample) are included in the analysis. The data was extracted from the EPA Air Quality System (AQS). Detailed information AQS can be found on the AQS web page at https://www3.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm.

Table 1 lists the data sets that have been analyzed using the T-SpACE software.

Table 1. Input data to the T-SpACE.

Data Source	Air Pollutant (units)	Dates	Resolution	Input file name	File Format	Available? *
Community Multiscale Air Quality (CMAQ) model data	$PM_{2.5}$ ($\mu g/M^3$)	January 1, 2001 – December 31, 2001	12 km	pm25_12km.ioapi	NetCDF	Yes
			36 km	cmaq_36km.ioapi	NetCDF	Yes
		January 1, 2002 – December 31, 2002	12 km	2002ac_v4.61_l3b_eus_12km.dailyavg.pm25	NetCDF	Yes
			36 km	2002ac_v4.61_l3b_us36b.dailyavg.pm25	NetCDF	Yes
	O_3 (ppm)	January 1, 2001 – December 31, 2001	12 km	ozone.conc	NetCDF	Yes
			36 km	Not available		No
		January 1, 2002 – December 31, 2002	12 km	2002ac_v4.61_L3b_EUS_12km.combine.hourly.O3.total365	NetCDF	Yes
			36 km	2002ac_v4.61_L3b_us36b.combine.hourly.O3.total365	NetCDF	Yes
Daily concentrations from the Federal Reference Method (FRM) samplers of the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) networks	24-hour integrated $PM_{2.5}$ concentration ($\mu g/M^3$)	January 1, 2001 – December 31, 2001	NA	AQSPROD_373180-1.txt	Text	Yes
		January 1, 2002 – December 31, 2002	NA	AQSPROD_38771-1.txt	Text	Yes
	O_3 (ppm)	January 1, 2001 – December 31, 2001	NA	RD_501_44201_2001-0.txt	Text	Yes
		January 1, 2002 – December 31, 2002	NA	RD_501_44201-2002-0.txt	Text	Yes

* The data is available from EPA directly. The AQS monitor data is directly available from https://www3.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm.

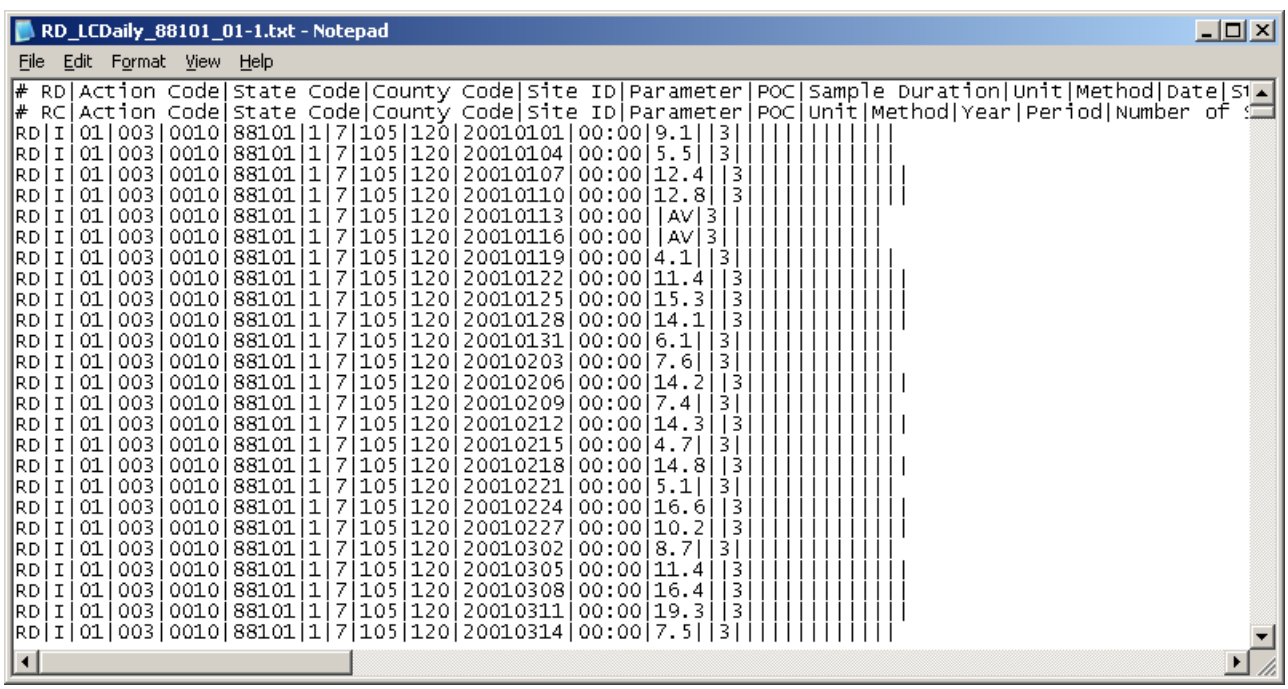
2.1 Format of initial monitor data from AQS

Currently the input file must be the same format as the AQS *.txt files that are available on EPA's web site. The two lines that describe the contents of each data row are reproduced here and Figure 4 illustrates the required structure of the data.

RD | Action Code | State Code | County Code | Site ID | Parameter | POC | Sample Duration | Unit | Method | Date | Start Time | Sample Value | Null Data Code | Sampling Frequency | Monitor Protocol (MP) ID | Qualifier - 1 | Qualifier - 2 | Qualifier - 3 | Qualifier - 4 | Qualifier - 5 |
Qualifier - 6 | Qualifier - 7 | Qualifier - 8 | Qualifier - 9 | Qualifier - 10 |
Alternate Method Detectable Limit | Uncertainty

RC | Action Code | State Code | County Code | Site ID | Parameter | POC | Unit | Method | Year |
Period | Number of Samples | Composite Type | Sample Value | Monitor Protocol (MP) ID | Qualifier - 1 | Qualifier - 2 | Qualifier - 3 | Qualifier - 4 | Qualifier - 5 | Qualifier - 6 | Qualifier - 7 | Qualifier - 8 | Qualifier - 9 | Qualifier - 10 | Alternate Method Detectable Limit | Uncertainty

Figure 4. Screen shot of the required AQS Monitor data file structure.


RD_LCDaily_88101_01-1.txt - Notepad

File Edit Format View Help

```
# RD|Action Code|State Code|County Code|Site ID|Parameter|POC|Sample Duration|Unit|Method|Date|S
# RC|Action Code|State Code|County Code|Site ID|Parameter|POC|Unit|Method|Year|Period|Number of
RD|I|01|003|0010|88101|1|7|105|120|20010101|00:00|9.1||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010104|00:00|5.5||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010107|00:00|12.4||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010110|00:00|12.8||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010113|00:00||AV|3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010116|00:00||AV|3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010119|00:00|4.1||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010122|00:00|11.4||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010125|00:00|15.3||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010128|00:00|14.1||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010131|00:00|6.1||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010203|00:00|7.6||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010206|00:00|14.2||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010209|00:00|7.4||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010212|00:00|14.3||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010215|00:00|4.7||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010218|00:00|14.8||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010221|00:00|5.1||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010224|00:00|16.6||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010227|00:00|10.2||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010302|00:00|8.7||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010305|00:00|11.4||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010308|00:00|16.4||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010311|00:00|19.3||3||||||||||||
RD|I|01|003|0010|88101|1|7|105|120|20010314|00:00|7.5||3||||||||||||
```

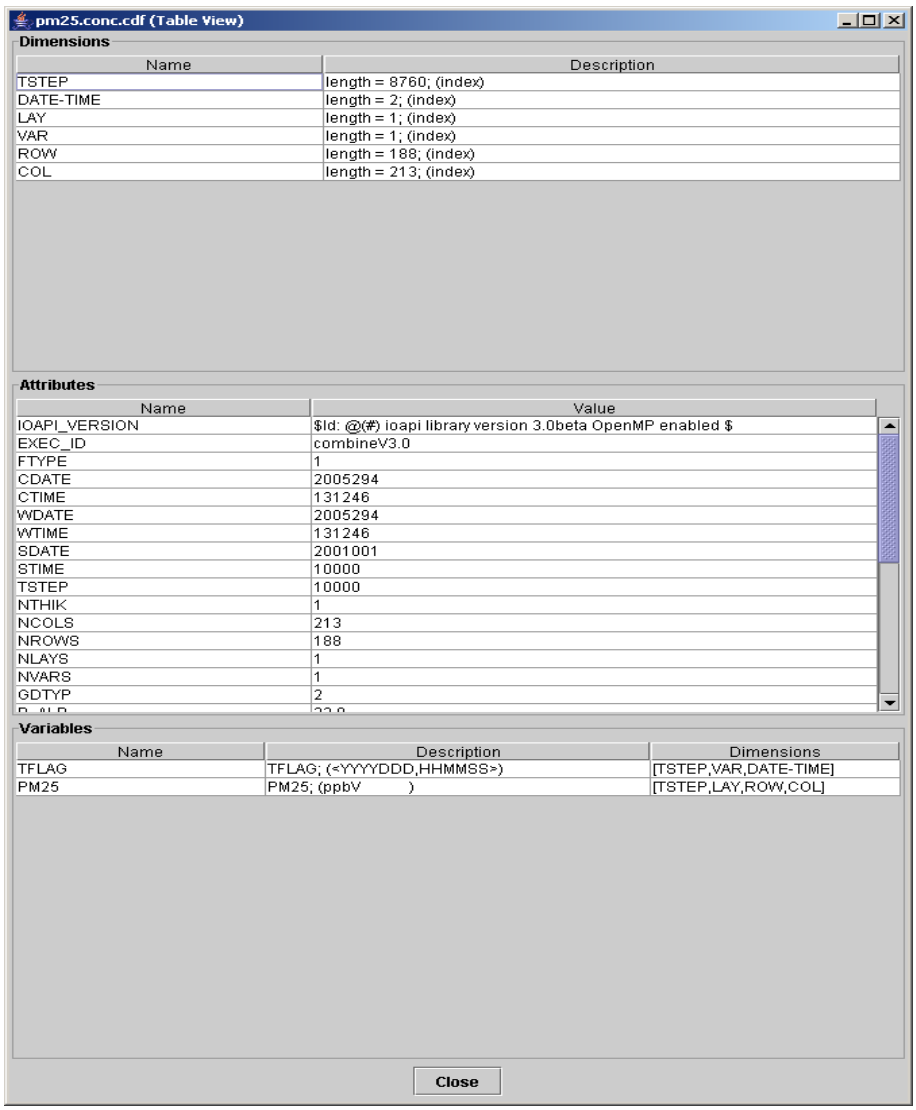
2.2 Format of the initial simulated surface data (CMAQ)

The format of the initial simulated surface data file must be the same as the CMAQ data files.

The CMAQ data files are NetCDF format data files with the following variables. Figure 5 illustrates the structure of the data file. Note that the illustration is specific to $PM_{2.5}$. Ozone would replace $PM_{2.5}$ for the ozone data.

```
netcdf file: pm25.conc
variables:
   TFLAG (Integer)            Timestep, format is YYYYDDD (where DDD is 1-366) or HHMMSS
   PM25  (Float)              In ppbV
   LambertConformal (Char)    Lambert Conformal Conic
   X (Double)                 Synthesized x coordinate, in km
   Y (Double)                 Synthesized y coordinate, in km
   Z (Double)                 Synthesized z coordinate, in km
   Time  (Integer)            Seconds since 2001-01-01 01:00:00 UTC
```

Figure 5. Structure of the simulated surface input data file.



3.0 SIMULATION MODEL

A component of the T-SpACE software contains the model executable software which processes the input data. In this version of the T-SpACE software, the model executable software component is **Model5-4H.exe**. This executable file is transparent to the user and is only listed here for documentation purposes only. The executable file included in this software has been developed for a 32-bit, Windows XP or higher machine. A 64-bit, Windows version of the executable is also available. The executable file implements a custom-designed Markov Chain Monte Carlo (MCMC) algorithm and can calculate an average, three consecutive year (annual) 4th highest concentration surface for $PM_{2.5}$. Section A provides a detailed description of the MCMC algorithm that has been developed for T-SpACE.

4.0 SOFTWARE TO RUN THE MODEL

The software that runs the T-SpACE model, using the data described above, can be grouped under five major components (steps). The last step visualizes the resultant air pollutant concentration surface. The steps are as follows:

- Step 1: Choose Time/Grid
- Step 2: Prepare Model Input Data
- Step 3: Model Specification
- Step 4: Launch Model
- Step 5: Launch Validation
- Step 6: Visualize Surface

Each of these steps is described below. A user must start with Step 1 and follow each step sequentially. Even if the user has run the T-SpACE software previously and has an existing parameter (*.PAR) file that they would like to use to run a simulation, the user must run Step 1 first, proceed to Step 2 to specify the location of the CMAQ/CAMx surface data file that contains the grid information, and then proceed to Step 3 to locate the user-generated *.PAR file.

*Please note that all *.CSV files generated by the Temporal-Spatial Ambient Concentrator Estimator (T-SpACE) are accessed (opened and closed) multiple times during the simulation session. During development and testing of the Temporal-Spatial Ambient Concentrator Estimator, it was discovered that real-time virus scanning software interfered with writing to the *.CSV files, especially when large files are generated. When real-time virus scanning software operates in the background, whenever a request is made to open a file, the real-time virus scanning software scans the file, thus locking it from use by any other program. If the file is large, this can take a relatively large amount of time. Because the T-SpACE software must open and close the *.CSV files several times, when the files are locked by real-time virus scanning software, this interferes with the ability of T-SpACE to write the results of the T-SpACE simulation to the appropriate files. Because of this interaction between virus scanning software and T-SpACE, it is recommended that the model be run on a computer without virus scanning software, or on a computer where the virus scanning software has been disabled.*

4.1 Installation

When the T-SpACE installation CD is inserted in the CD drive of the computer, the installation process should begin immediately as shown in Figure 6. The user must click the “Next” button to continue the installation.

Figure 6. First screen during the installation process.

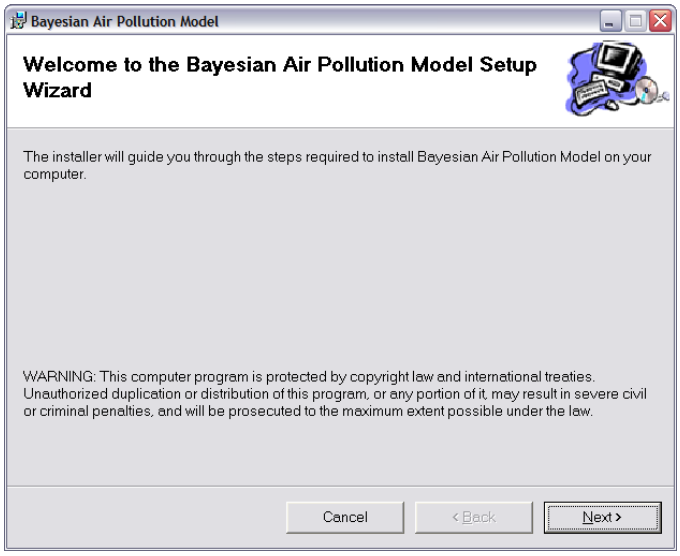


Figure 7. Location for the Hierarchical Bayesian Air Pollution Model Software (T-SpACE)

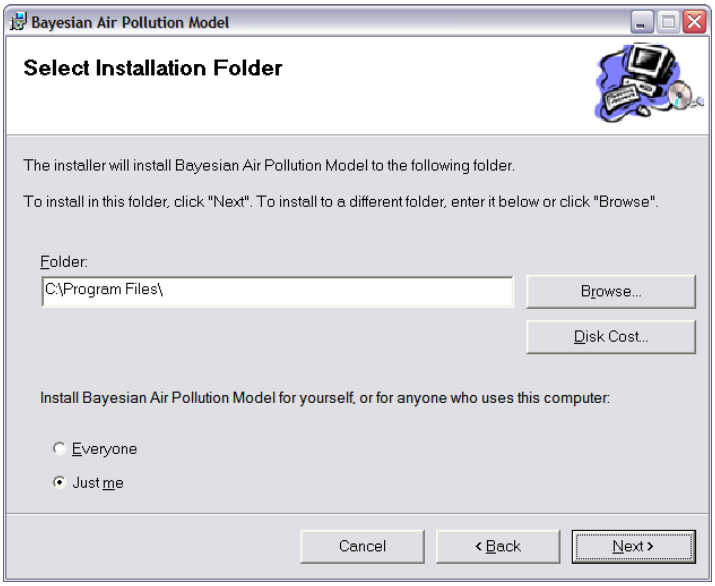
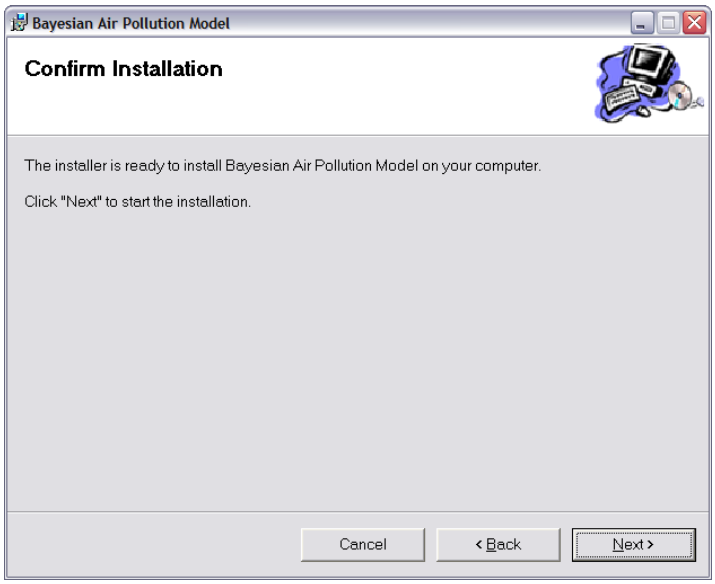


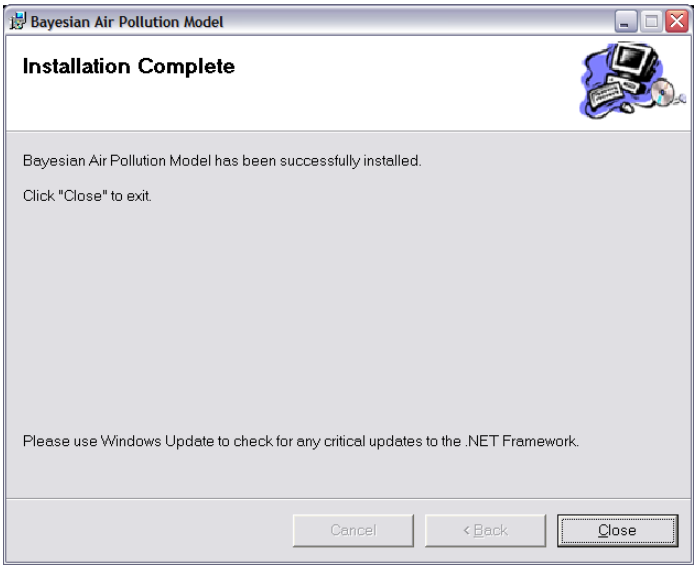
Figure 7 prompts the user to select the directory/folder where the T-SpACE software is installed. (Please note that this is not where results are stored. Only system files are stored here.) The user must specify a directory/folder other than the default folder (C:\Program Files) to store results. The user must also select whether T-SpACE will be available only to the assigned user/owner of the computer or available to anyone logging into the computer where T-SpACE is installed.

Figure 8. Screen Indicating that Installation of the Software has completed.



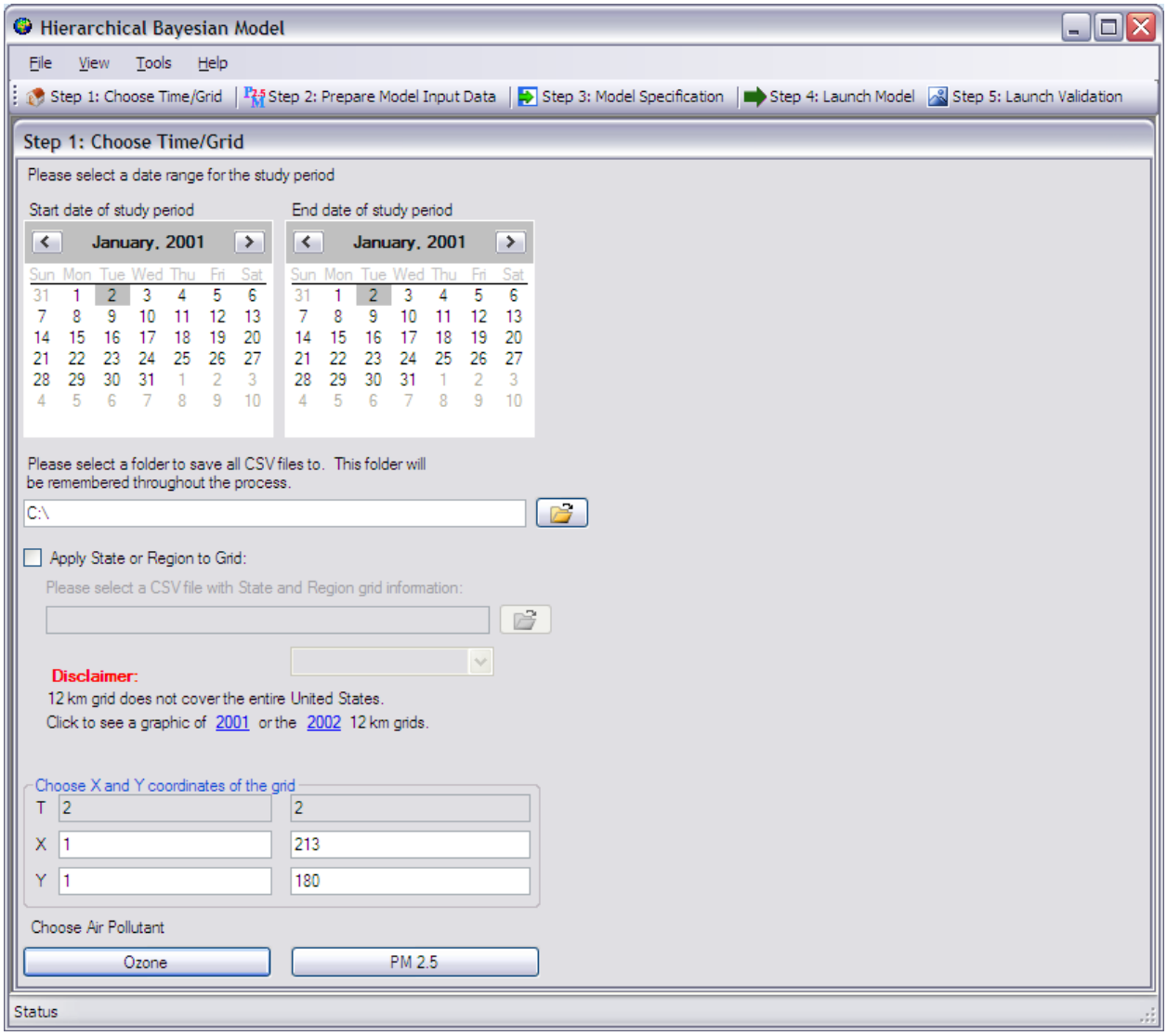
User must choose “Next” to continue the installation (Figure 8).

Figure 9. Screen indicating that installation of the software has completed.



The user must click “Close” to complete the installation (Figure 9). Please note that T-SpACE requires the Microsoft .NET 2.0 framework. During the installation .NET 2.0 has been installed on your computer. If you encounter problems during the installation it may be that you do not have privileges for installing software on your machine. If problems do arise, please see your system administrator to assist you in the installation.

Figure 10. Opening screen for the T-SpACE (Hierarchical Bayesian Model) software.



4.2 Menu Bar

There is a menu bar which is available to users throughout the simulation process. The menu bar is at the top of the screen as illustrated in Figure 10. The menu bar includes the following:

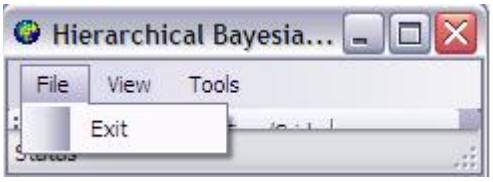
- File
- View
- Tools
- Step 1: Choose Time/Grid
- Step 2: Prepare Model Input Data
- Step 3: Model Specification
- Step 4: Launch Model
- Step 5: Launch Validation

File, view, and tools will be described here, while the following sections will provide details for each of the steps.

4.2.1 File

Figure 11 illustrates what is available when the user clicks on File.

Figure 11. Main Menu: File.



The only option available to a user is to Exit the T-SpACE software. Clicking on exit will close the software.

4.2.2 View

Figure 12 illustrates that there are two options for the user under 'View'. The user can choose to display the T-SpACE steps (Step 1 through Step 5) on the toolbar by checking

Figure 12. Main Menu: View.

'Toolbar', or choose not display the model steps on the toolbar, by un-checking 'Toolbar'.

4.2.3 Tools

Seven tools have been developed for the user. Three are shown below (and the entire list will be provided in Section I):

- Compare AQS Data Files
- Compare Airsites Files
- Average 4th-Highest Surface Files

Figure 13. Main Menu: Tools.

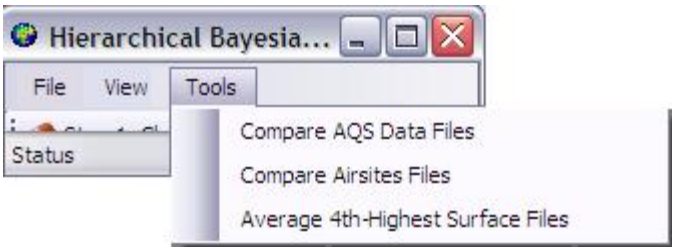
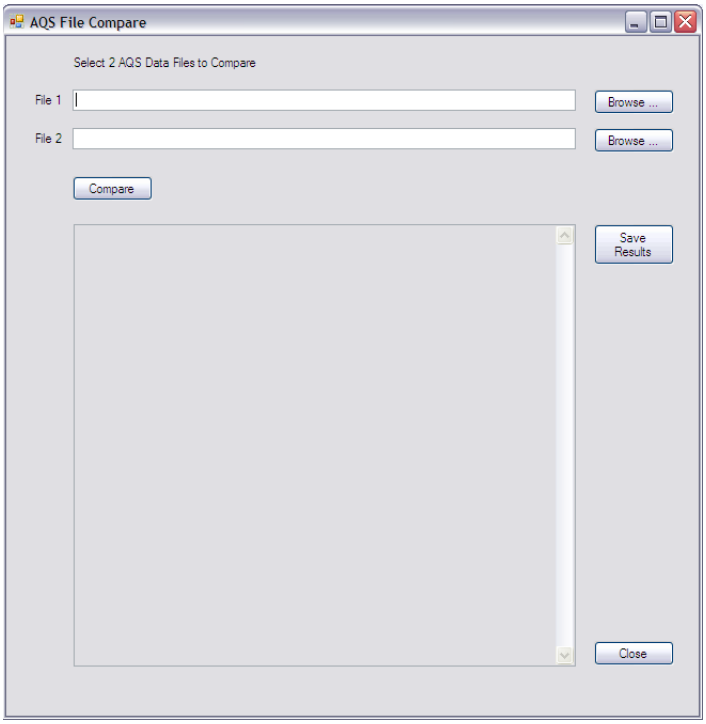


Figure 13 illustrates the three options available under Tools.

4.2.3.1 Compare AQS Data Files

Figure 14. AQS File Compare.



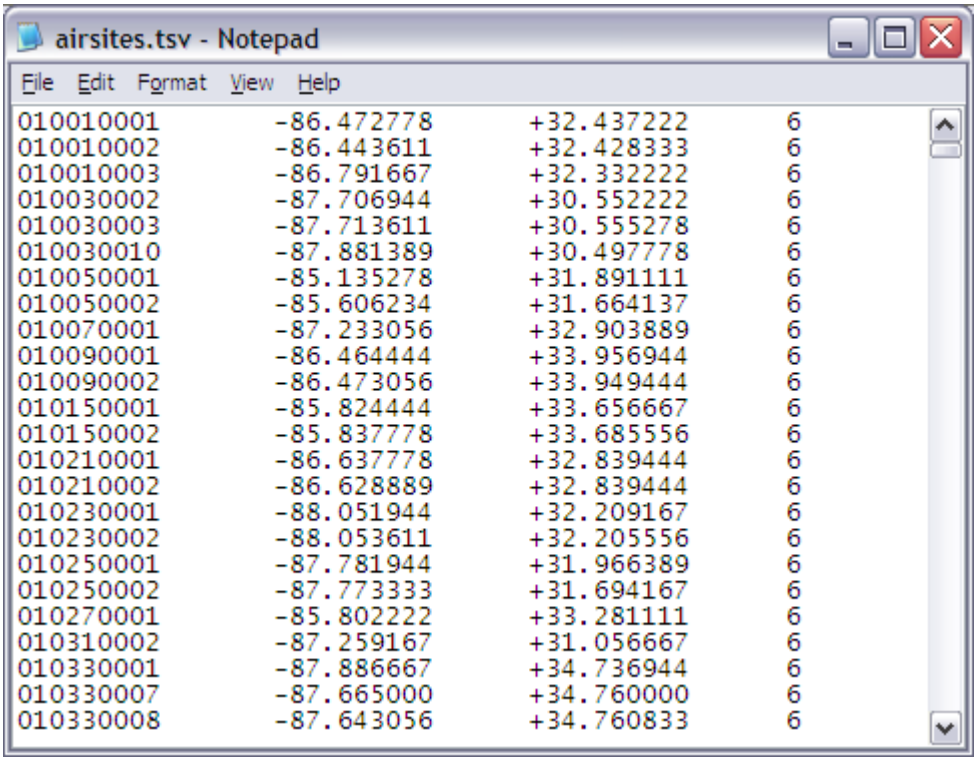
This tool provides the user with the ability to compare *.CSV model input monitor data files to understand the differences between the two files. The *.CSV file is the file developed in Step 2 (discussed later). Figure 14 shows that the user chooses two files to compare, clicks on the "Compare" button, and the results of the comparison are provided in the window below the "Compare" button. Note that this is a line-by-line file comparison.

This tool is useful for users who want to know if the input files they have created in the past are the same or not.

4.2.3.2 Compare Airsites Files

Airsites is a file that is used to provide the latitude and longitude for each of the air pollution monitoring sites. This tool allows the user to compare two airsites files to determine if they are the same or if there are differences. If there are differences, the differences are listed. Figure 15 illustrates the structure of the Airsites file, while Figure 16 illustrates the comparison screen.

Figure 15. Airsites.TSV file.

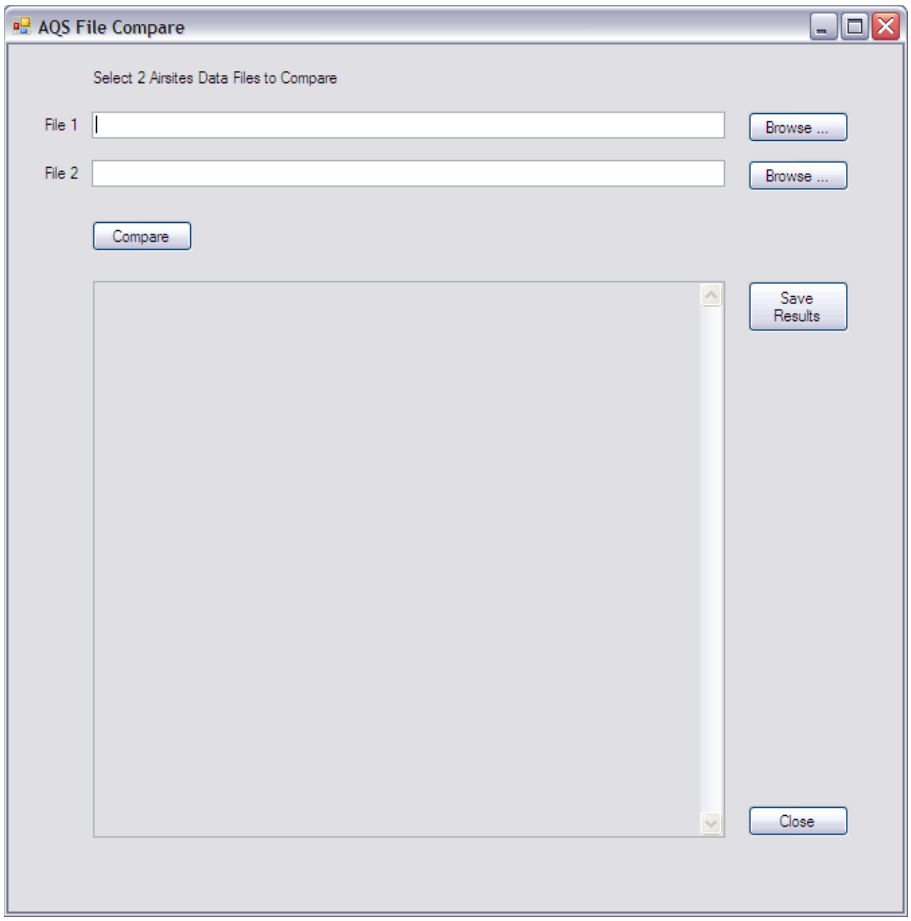


airsites.tsv - Notepad

File Edit Format View Help

010010001	-86.472778	+32.437222	6
010010002	-86.443611	+32.428333	6
010010003	-86.791667	+32.332222	6
010030002	-87.706944	+30.552222	6
010030003	-87.713611	+30.555278	6
010030010	-87.881389	+30.497778	6
010050001	-85.135278	+31.891111	6
010050002	-85.606234	+31.664137	6
010070001	-87.233056	+32.903889	6
010090001	-86.464444	+33.956944	6
010090002	-86.473056	+33.949444	6
010150001	-85.824444	+33.656667	6
010150002	-85.837778	+33.685556	6
010210001	-86.637778	+32.839444	6
010210002	-86.628889	+32.839444	6
010230001	-88.051944	+32.209167	6
010230002	-88.053611	+32.205556	6
010250001	-87.781944	+31.966389	6
010250002	-87.773333	+31.694167	6
010270001	-85.802222	+33.281111	6
010310002	-87.259167	+31.056667	6
010330001	-87.886667	+34.736944	6
010330007	-87.665000	+34.760000	6
010330008	-87.643056	+34.760833	6

Figure 16. Airsites.TSV file comparison screen.

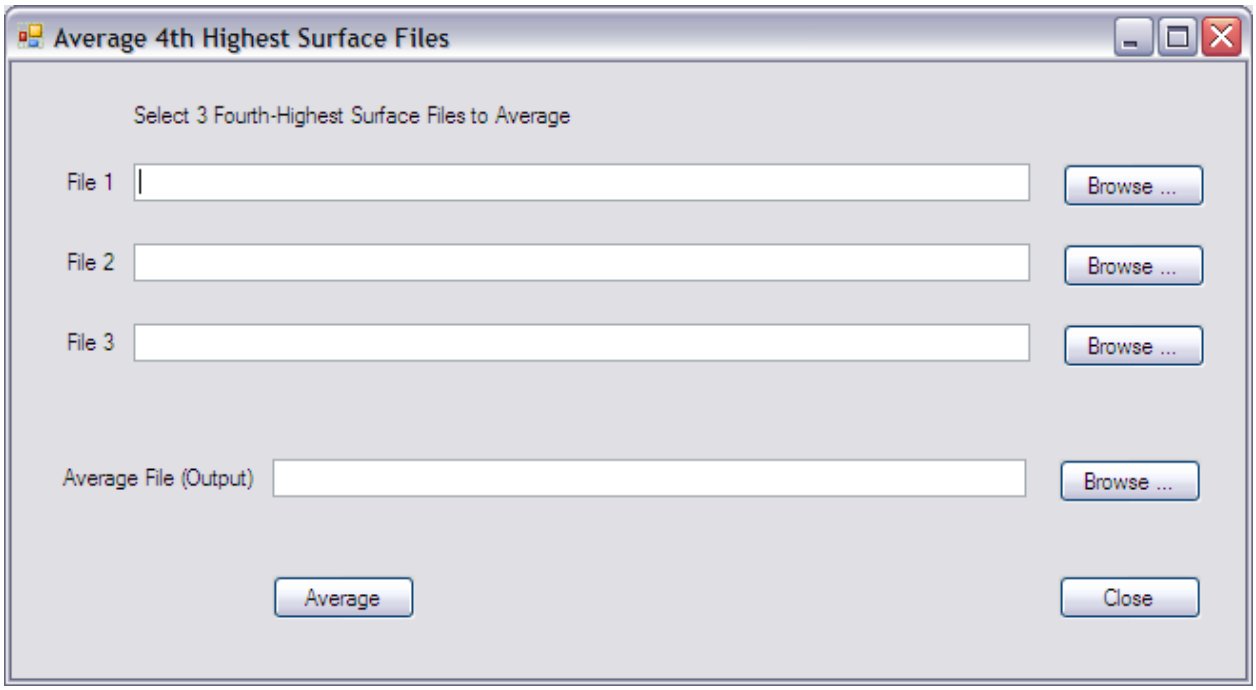


4.2.3.3 Average 4th-Highest Surface Files

This tool allows the user to calculate a three consecutive year average 4th highest surface file. The calculation of a single 4th highest surface file will be discussed in Section 4. See [2] for an explanation of the calculation of the three-year average.

In order to calculate the three-year average, a user must specify three 4th highest surface files. It is recommended that these files be calculated on similar time frames and with a similar number of burn-in and simulation steps. Figure 17 illustrates the input screen.

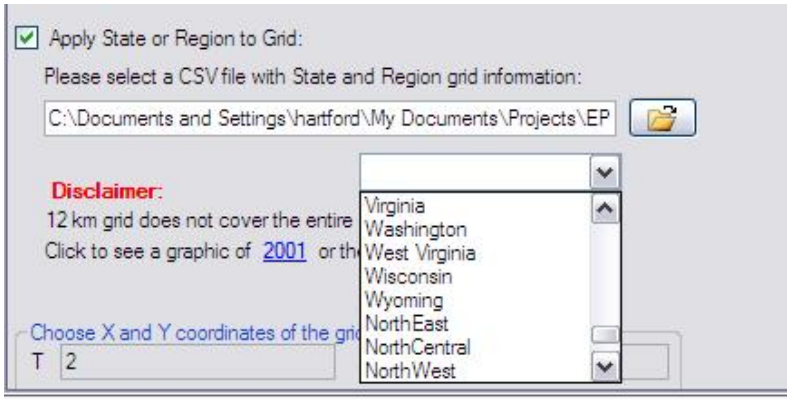
Figure 17. Average 4th highest surface calculation input screen.



4.2.4 Choosing a Region or State for the Simulation

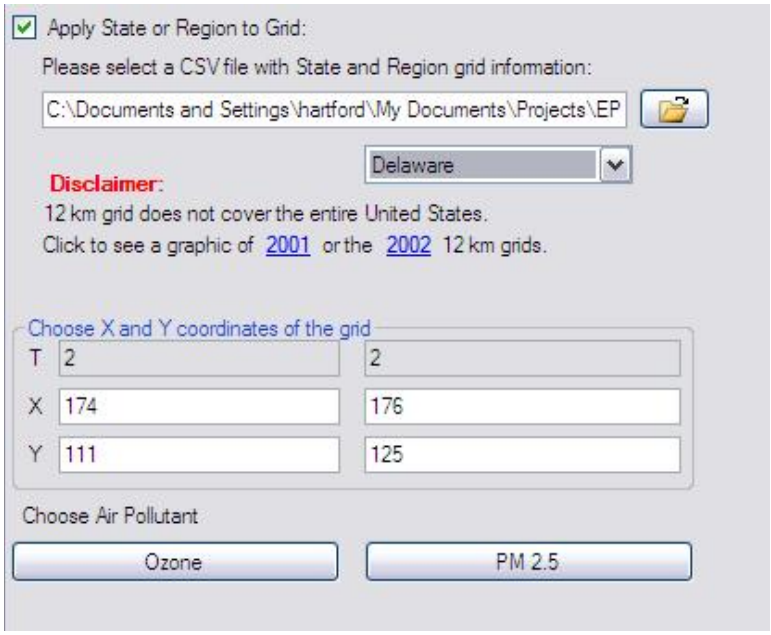
T-SpACE has a feature which allows users the ability to identify the appropriate grid coordinates for a particular region or state. On the screen for Step 1, the user checks the box next to "Apply State or Region to Grid". The user is then provided four default CSV files to choose the appropriate coordinate system. The default files are found in the directory \StateFiles where the software was loaded and are named:

StateGrid2001-12km.csv
StateGrid2001-36km.csv
StateGrid2002-12km.csv
StateGrid2002-36km.csv



The user must first load this file into the system by choosing the file. Once the

user has chosen the file, a dropdown menu will be provided to the user. This dropdown menu shows the states that are available and provides six regions. Note the disclaimer (The 12 km^2 grid does not cover the entire United States), which is true for all years before 2008. Two hyperlinks are provided to allow the user the opportunity to see the 12 km^2 state coverage for the years 2001 and 2002.



Once the user makes their choice, the grid coordinates are placed in the X and Y rows on the screen. Using the latitude and longitude of the lowest left corner and highest right corner, the state is transformed to grid coordinates. The latitude and longitude of the centroid of the grid cell is used to define whether the grid appears in the state or not. Note that this essentially defines each state as a rectangle, therefore portions of surrounding states may be 'connected' to the chosen state.

4.3 Step 1: Choose Time/Grid

Step 1 consists of the following actions:

- Choosing the time frame over which the simulation will be run
- Establishing the location of all output files
- Choosing the area of the United States over which the simulation will be run

As shown above, Figure 10 illustrates the input screen for collecting this information.

4.3.1 Please select a date range for the study period

The study period is limited to at most one calendar year of data. The user cannot choose a time frame that spans two calendar years. For instance, the software will not allow a user to choose a date in one year and choose a second date in another year. i.e., December 15, 2006 to January 5, 2007 would not be allowed. Please note that all testing to date has utilized only data from the years 2001 and 2002.

In order to choose the time frame, the user can click on the month. They will see a drop down box that will allow them to choose the month of interest. Or, the user can use the arrows at the top of each calendar to navigate between months. The date within the month is chosen by clicking on the number on the calendar.

As the user chooses the start and end date for the period, they will see the T: row under "Choose X and Y coordinates of the grid" change. The value in T represents the number of days from January 1 for the value that is chosen, where January 1 is day 1. As can be seen in Figure 10, since January 2 is both the start and end date, the value of T is 2, 2.

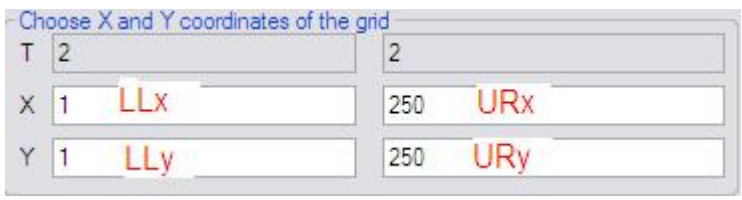
4.3.2 Please select a folder to save all CSV files

Next, the user is asked to choose a location where all files created during the simulation will be written and stored. (Note that this is different from the system file location discussed earlier.) Input data files are created in Step 2, the model parameter specification file (*.PAR) is created in Step 3, and the model simulation output files are created in Step 4. All these files will be placed in the specified folder. If no folder is specified, the default location of the files created with this software will be "C:\", the root folder.

4.3.3 Choose X and Y coordinates of the grid

As mentioned in 4.3.1, the T coordinates are automatically updated as the user navigates the start and end date calendars. The user cannot make changes to the T row manually. All changes must be made using the calendar described above. The X and Y coordinates that are referred to in this area of the input screen, are the coordinates of the square grid over which the user would like to run T-SpACE model. As shown in Figure 18, the coordinates are read as follows:

Figure 18. Assigning grid coordinates.



(LLx, LLy) → (URx, URy)

where

LLx Lower Left x-coordinate of the square grid. This is the first column of the input screen.
LLy Lower Left y-coordinate of the square grid. This is the first column of the input screen.
URx Upper Right x-coordinate of the square grid. This is the second column of the input screen.
URy Upper Right y-coordinate of the square grid. This is the second column of the input screen.

If the user would like to use the CMAQ grid, as illustrated in Section 2, there are the 12 kilometer CMAQ grid or the 36 kilometer CMAQ grid. For the CMAQ data, Figures 1, 2, and 3 illustrate the maximum coordinate range for each grid size and year and the area of coverage for

the respective grids. Note that the software does not provide default ranges, but after each T-SpACE model run, the coordinates from the previous run are listed.

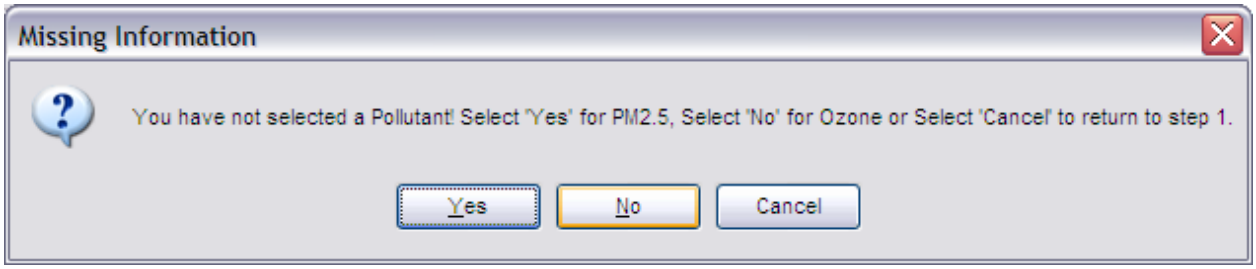
Figures 1, 2, and 3 illustrate the maximum ranges available. Recommendations for CMAQ grid sizes, based on areas of study conducted by EPA, are as follows:

2001, 12 km (41, 49) → (213, 180) (Eastern half of the US coverage and illustrated in Figure 19)
2001, 36 km (13, 15) → (142, 94) (US only coverage)

4.3.4 Choose Air Pollutant

The user must choose the air pollutant for which the T-SpACE model simulation will be run. Currently, the two choices are $PM_{2.5}$ or ozone.

If the user does not choose an air pollutant and attempts to run another step in the simulation, they will see the following error screen.



4.4 Step 2: Prepare Model Input Data

As discussed in Section 2, there are two sets of input data that are required to run the simulation. The data are the CMAQ/CAMx estimated air quality model concentrations, and the actual NAMS/SLAMS air pollutant concentrations obtained at monitoring stations. This step takes the information from Step 1, gathers additional information including the name of the raw data file and the name the user would like to call the analysis data set created in this step, and performs the appropriate transformations to create .CSV files that are used as input to the simulation run.

Figure 20 illustrates the initial screen when Step 2 is chosen. There are two tabs on this screen

Figure 19. The area of the United States over which 2001, 12 Km CMAQ published results of the model have been run. (The NAMS/SLAMS network sites are red circles).

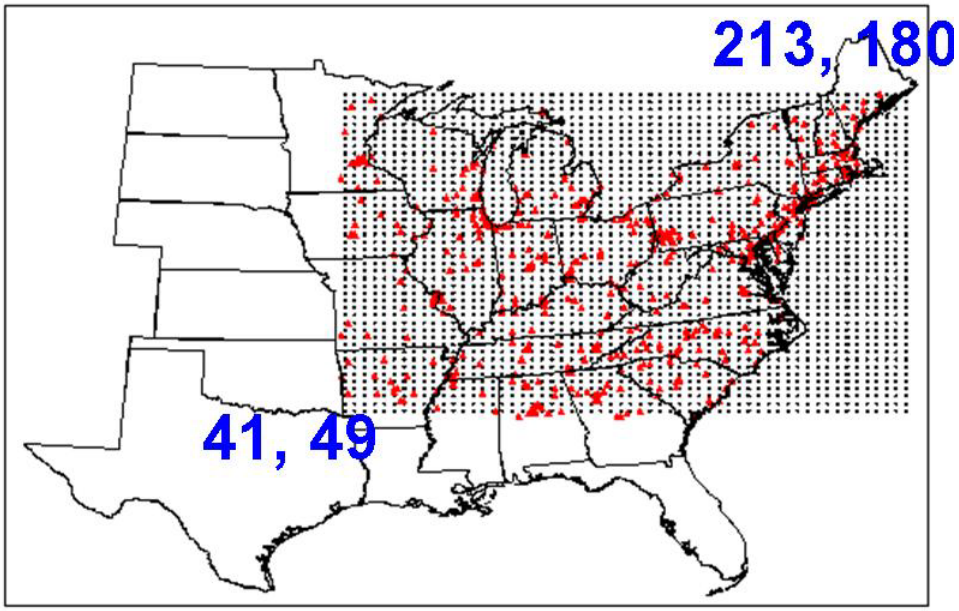
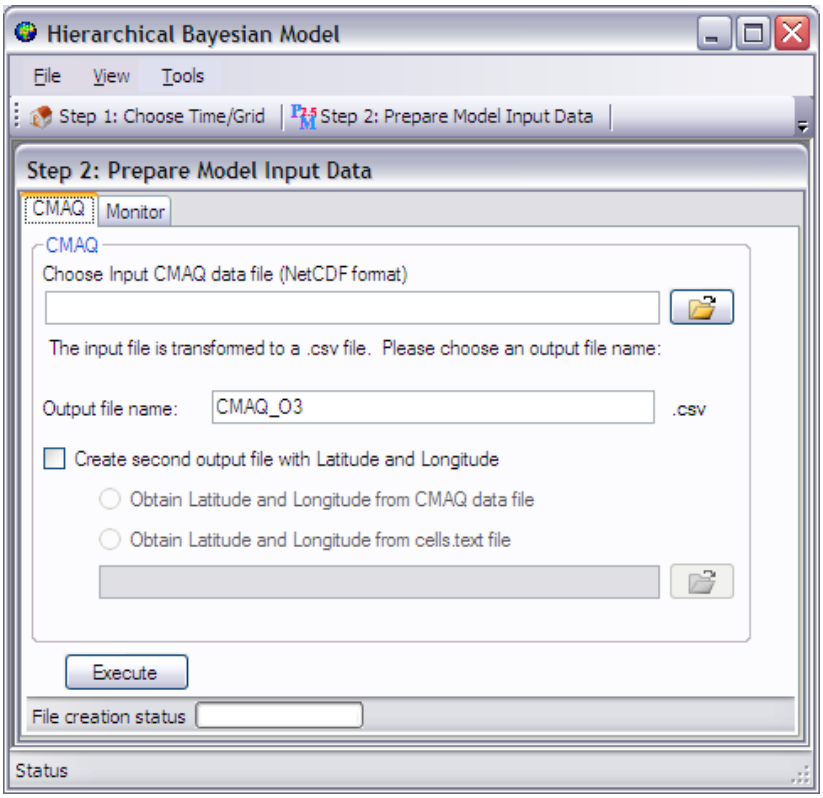


Figure 20. Step 2: Prepare Model Input Data – CMAQ (or CAMx).



4.4.1 CMAQ: Choose Input CMAQ data file

In Step 1, the user has already made a decision on the area of coverage. What the user is doing here is choosing the appropriate raw data file to be processed. The user can enter the full path, ending with the name of the file, or the user can hit the folder button and browse their computer to locate the file. If a user utilizes the browse function, when they have selected a file, the full pathname and filename will be entered in this box. A post-processing step requires this file, so it is absolutely necessary that this file be specified, even if the development of the input data files is going to be skipped. Note that if an attempt to run Step 4 is made without a CMAQ/CAMx data file, the user will be asked to supply one. The format of this file was discussed earlier in Section 2. This is a file that is in NetCDF format and has the structure listed in Section 2.

4.4.2 CMAQ: Output file name

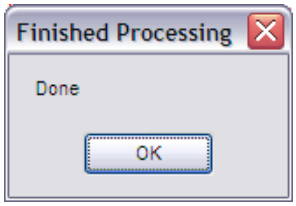
The user can specify any file name in this area of the input screen. The created file will automatically have a .CSV extension. The user should **NOT** specify a pathname here, since the pathname has already been entered in **Step 1: Please select a folder for saving .CSV files**. If the user does specify a pathname, an error will occur, and the analysis data set will not be created.

4.4.3 CMAQ: Create a second output file with latitude and longitude

This is a utility that was developed to provide the user with a file that looks like the input CMAQ/CAMx file, but also includes the latitude and longitude for the center of the CMAQ/CAMx locational grid cell. The user can choose to use the CMAQ/CAMx NetCDF file, which contains the latitude and longitude for each grid cell, or may input a file that is called Cells.txt. Section D illustrates the structure of Cells.txt. This file is available from EPA, and is year-specific, with a specified grid size.

4.4.4 CMAQ: Execute

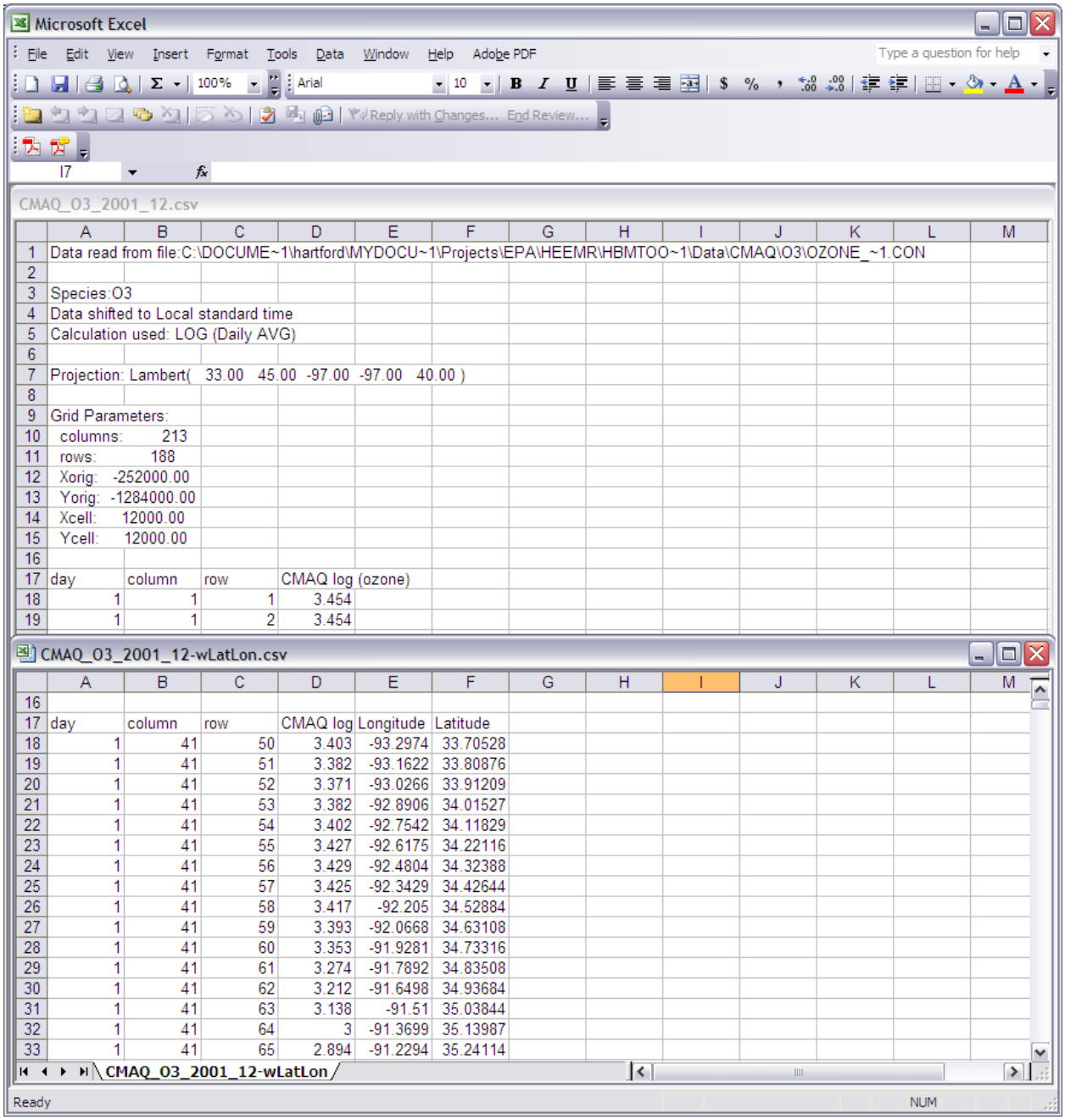
Once the input raw data file is generated, and the name of the analysis data set is created, and when the decision to create a second file with latitude and longitude has been made, and the file created, and all the information in Step 1 is the way that the user wishes, the user should click on the 'Execute' button.



The completion of the T-SpACE model execution is indicated by a box that pops up with "Done". There are three files created during the model execution. The CMAQ/CAMx input file, a copy of the CMAQ/CAMx input file with latitude and longitude (not to be used as input), and a log file containing the steps that were completed. Figure 21 shows the CMAQ/CAMx input file compared to the CMAQ/CAMx input file with latitude and longitude.

The file labeled "**CMAQ_O3_2001_12km.csv**" is the CMAQ/CAMx input file created during the example run. The file has a series of information in the header of the file. Row 18 is where the actual data begins. The header information is as follows:

Figure 21. CMAQ simulation input file created in Step 2 compared to the CMAQ input file with latitude and longitude.



CMAQ_O3_2001_12.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data read from file:C:\DOCUME~1\hartford\MYDOCU~1\Projects\EPA\HEEMR\HBMTOO~1\Data\CMAQ\O3\OZONE_~1.CON												
2													
3	Species:O3												
4	Data shifted to Local standard time												
5	Calculation used: LOG (Daily AVG)												
6													
7	Projection: Lambert(33.00 45.00 -97.00 -97.00 40.00)												
8													
9	Grid Parameters:												
10	columns: 213												
11	rows: 188												
12	Xorig: -252000.00												
13	Yorig: -1284000.00												
14	Xcell: 12000.00												
15	Ycell: 12000.00												
16													
17	day	column	row	CMAQ log (ozone)									
18	1	1	1	3.454									
19	1	1	2	3.454									

CMAQ_O3_2001_12-wLatLon.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
16													
17	day	column	row	CMAQ log	Longitude	Latitude							
18	1	41	50	3.403	-93.2974	33.70528							
19	1	41	51	3.382	-93.1622	33.80876							
20	1	41	52	3.371	-93.0266	33.91209							
21	1	41	53	3.382	-92.8906	34.01527							
22	1	41	54	3.402	-92.7542	34.11829							
23	1	41	55	3.427	-92.6175	34.22116							
24	1	41	56	3.429	-92.4804	34.32388							
25	1	41	57	3.425	-92.3429	34.42644							
26	1	41	58	3.417	-92.205	34.52884							
27	1	41	59	3.393	-92.0668	34.63108							
28	1	41	60	3.353	-91.9281	34.73316							
29	1	41	61	3.274	-91.7892	34.83508							
30	1	41	62	3.212	-91.6498	34.93684							
31	1	41	63	3.138	-91.51	35.03844							
32	1	41	64	3	-91.3699	35.13987							
33	1	41	65	2.894	-91.2294	35.24114							

Relevant CMAQ/CAMx header information:

Line 1: Provides the full pathname and the filename of the CMAQ/CAMx NetCDF file used to create the file.
Line 3: Lists the species. In this case, it is ozone (O_3).
Line 4: Indicates if daylight saving shift was employed in the appropriate time frame.
Line 5: Indicates the type of value the concentration is. In this case, it is the log of the daily average concentration.
Line 7: This provides the projection (Lambert) used to transform the grid to the contour of the earth.
Line 10: Indicates the number of columns of the grid available in the NetCDF file.
Line 11: Indicates the number of rows of the grid available in the NetCDF file.
Line 17: These are the names of the columns.

Day is the number of the day (Julian Day) from January 1
Column is the column number of the grid
Row is the row number of the grid
CMAQ/CAMx log (ozone) is the log ozone concentration.

The file named "**CMAQ_O3_2001_12-wLatLon.csv**" has the exact same structure as "**CMAQ_O3_2001_12km.csv**", except there are two more columns of data (starting at line 18).

Longitude is the longitude of the center of the indicated CMAQ/CAMx grid cell
Latitude is the latitude of the center of the indicated CMAQ/CAMx grid cell

Comparing the two files in Figure 21 shows the following:

1. The structure of the two files is exactly the same except there are two additional columns of data in the "**wLatLon**" file.

2. The input CMAQ/CAMx data file includes all available data from the NetCDF file, while the "**wLatLon**" file only includes the CMAQ/CAMx grid specified in Step 1.

Finally, to ensure that the conversion from the NetCDF to the *.CSV input file has been performed correctly, the user is provided a log file of the conversion. The log file is a text file that is written to the directory specified in Step 1. The name of the log file is

RecordedOutput_{MM-DD-YYYY}{H-M}_{AM/PM}S.txt

where

MM	= Month in which file was created
DD	= Day on which file was created
YYYY	= Year in which file was created
H	= Hour of creation
M	= Minute of creation
AM/PM	= AM or PM creation time
S	= Seconds of the minute in which the file was created

Figure 22 presents an example log file for creating a CMAQ/CAMx analysis input data file.

Figure 22. Log file for the conversion of the CMAQ NetCDF file to the *.CSV format.

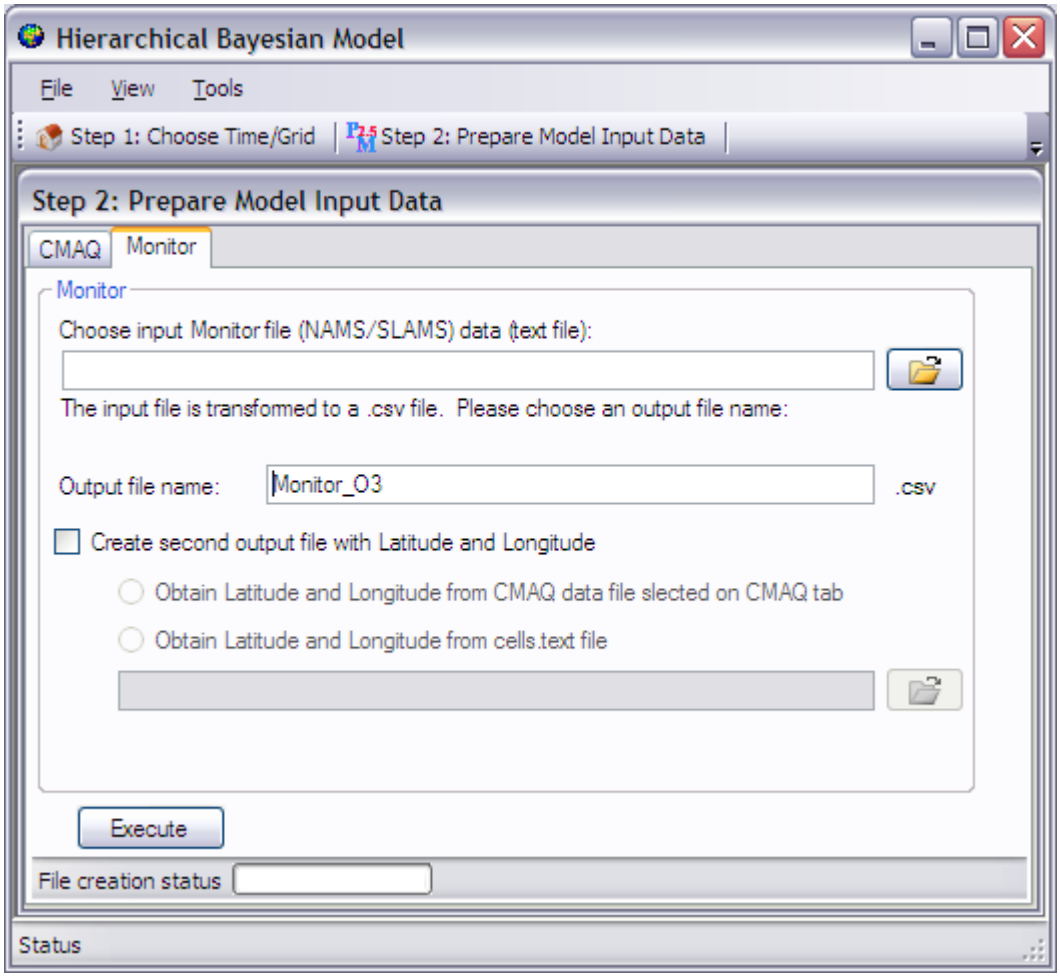
RecordedOutput_3-16-200911-41_AM216.txt - Notepad

File Edit Format View Help

```
     Value for IOAPI_NO_STDOUT not defined;returning default:  TRUE

     This program uses the EPA-AREAL/MCNC-Environmental Programs
     Models-3 I/O Applications Programming Interface, which is
     built on top of the netCDF I/O library (Copyright 1993, 1996
     University Corporation for Atmospheric Research/Unidata
     Program).  Portions copyright (C) 1992-2002 MCNC and
     Carlie J. Coats, Jr., and 2003 Baron Advanced Meteorological
     Systems, LLC.  See URL
     "http://www.emc.mcnc.org/products/ioapi/NOTICES.html" for
     conditions of use.

     Library release tag: 2003280

     $Id: @(#) ioapi library version 2.2 $
     netCDF version VERSION of Jan 30 2001 08:38:56 $


     Value for EXECUTION_ID:  'cmaqpm25O3.exe'
     Value for INFILE:
'C:\DOCUME~1\hartford\MYDOCU~1\Projects\EPA\HEEMR\HBMTOO~1\Data\CMAQ\O3\OZONE_~1.CON'
     Value for IOAPI_CHECK_HEADERS not defined;returning default:  FALSE

     "INFILE" opened as OLD:READ-ONLY
     File name
"C:\DOCUME~1\hartford\MYDOCU~1\Projects\EPA\HEEMR\HBMTOO~1\Data\CMAQ\O3\OZONE_~1.CON"
     File type GRDDED3
     Execution ID "????????????????"
     Grid name "M_12_OAQPS"
     Dimensions: 188 rows, 213 cols, 1 lays, 1 vbles
     NetCDF ID:         3  opened as READONLY
     Starting date and time  2001001:010000 (1:00:00   Jan. 1, 2001)
     Timestep                        010000 (1:00:00 hh:mm:ss)
     Maximum current record number    8760
     Value for CALTYPE:  'AVG'
     Value for SPECIES:  'O3'
     Value for ISLOCAL:  T returning TRUE
     Value for FACTOR not defined; returning default:  1.000000
     Value for USELOG:  T returning TRUE
     Value for OUTFILE:
'C:\DOCUME~1\hartford\MYDOCU~1\Projects\EPA\HEEMR\WA54~1\TASK5-~1\CMAQ_O3_2001_12.csv'
     Value for START_DATE:  2001001
     Value for END_DATE:  2001031
     Value for IOAPI_ISPH:  '19'
     INITSPHERES:  input sphere Normal Sphere
     Value for TZFILE:
'C:\DOCUME~1\hartford\MYDOCU~1\Projects\EPA\HEEMR\WA54~1\SOFTWA~1\SOFTWA~1\tz.csv'
     Time step error reading file:  INFILE
     Requested date & time:      2001001:000000
     File starting date & time:  2001001:010000
     File time step:           010000
     >>--->> WARNING in subroutine READ3
     Time step error reading file:  INFILE
     M3WARN:  DTBUF 0:00:00   Jan. 1, 2001
     >>--->> WARNING in subroutine READ3
     Time step not available for file:  INFILE
     M3WARN:  DTBUF 0:00:00   Jan. 1, 2001
End of run
CMAQ column renamer started at 3/16/2009 11:42:01 AM
Column Renaming complete in 00:00:01.5000288
```

The log file indicates if any problems occurred during the conversion process. Figure 22 illustrates a conversion that executed properly.

Figure 23. Step 2: Prepare Model Input Data – CMAQ/CAMx.



4.4.5 Monitor

The steps for creating the Monitor input data file are exactly the same as the steps for creating the input CMAQ/CAMx data file.

- To create an input monitor data file, the CMAQ/CAMx NetCDF file MUST be specified on the CMAQ tab, even if you are not developing a CMAQ/CAMx input file.

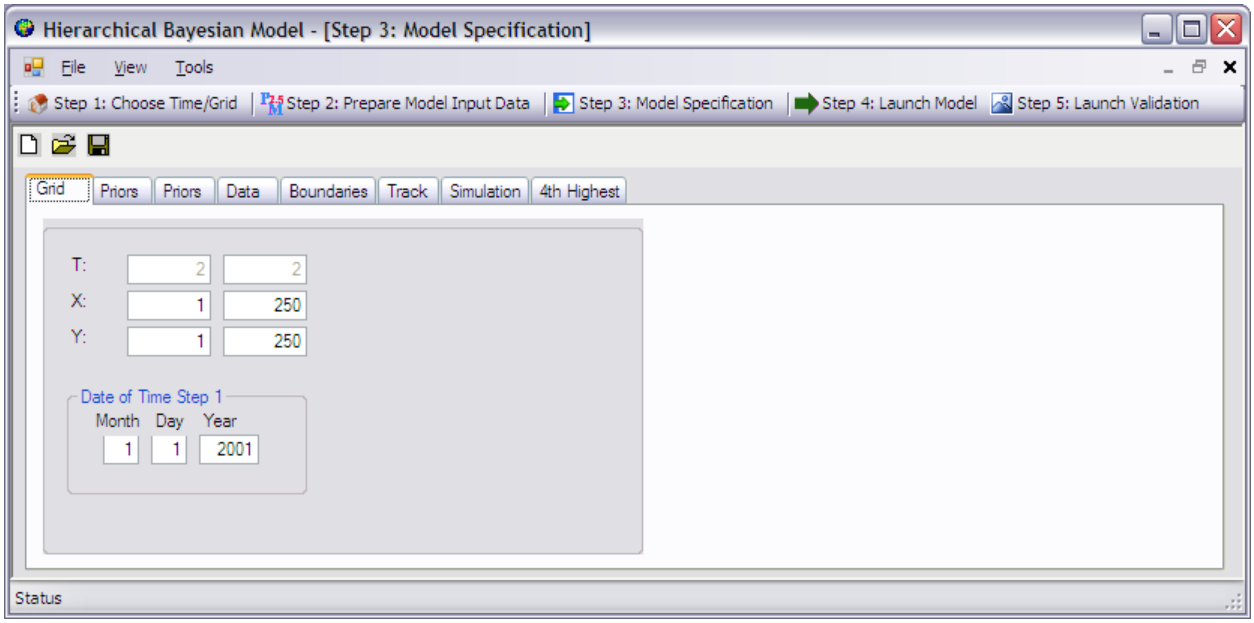
- The input monitor file must be a NAMS/SLAMS data file. This file is available through AQS in the appropriate text format. The exact structure of this file is presented and discussed in Section 2.

4.5 Step 3: Model Specification

Step 3 is where the user specifies the parameters for the T-SpACE simulation.

Figure 24 illustrates the opening screen for Step 3. The purpose of Step 3 is to specify the parameters of the model for the T-SpACE simulation, identify the input data, name the output data, and determine whether the average 4th highest surface will be calculated. Once the user has made all these decisions, a *.PAR (parameter) file that stores all the choices made, is saved and is then available for later use. Below is a description of the choices available to the user.

Figure 24. Step 3: Opening Screen.



4.5.1 New, Open, Save

Just above the tabs, the user will see three icons.  . These icons represent "New", "Open", and "Save" for the *.PAR (parameter) file.


New — Clicking on this icon clears all information that has been entered in the various choices available to the user. No files will be deleted when this button is selected. Only the choices made on-screen during this session in Step 3 will be changed. Note that the T, X, and Y values chosen in Step 1 will be changed to reflect any changes made here.


Open — If a user would like to utilize a previously developed *.PAR file, the user can click on this icon. Note that prior to opening a *.PAR file, the user must have specified the CMAQ/CAMx NetCDF file in Step 2.

Save When a user completes their choices, clicking this icon allows the user to save their choices to a *.PAR file.

4.5.2 Grid

On this tab, the user is specifying the grid coordinates over which the T-SpACE simulation will be run. Below is a description of each of the inputs available to the user.

T Represents the range of time in days (see "**Date of Time Step 1**" described below). The user cannot change the time frame on this tab and can only make changes in Step 1. The changes made in Step 1 will be reflected here.

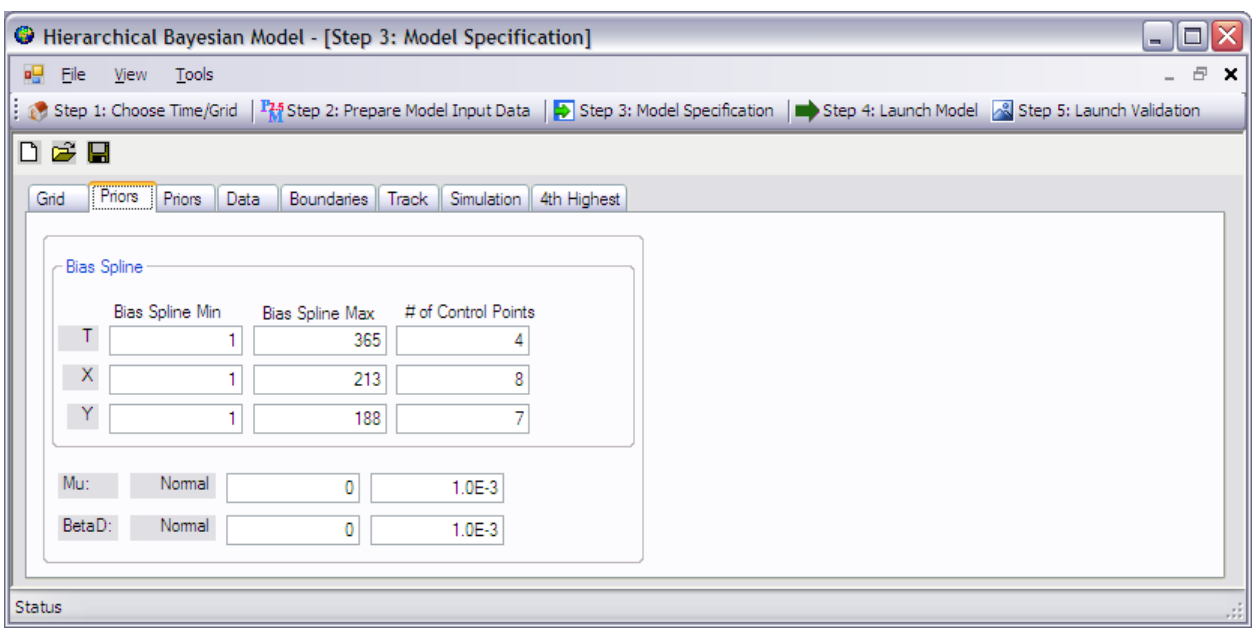
X Represents the x-coordinates of the locational grid over which the T-SpACE simulation will be run. By default, these are the x-coordinates chosen in Step 1. The user may make changes here to the x-coordinates if another grid is desired for the analysis. The user should ensure that the input data files are specified.

Y Represents the Y-coordinates of the grid over which the T-SpACE simulation will be run. By default, these are the y-coordinates chosen in Step 1.

Date of Time Step 1

The first day from which to start counting the day range. The default is January 1, YYYY and this is the recommended date. If a user chooses, they can enter any date here. For instance, if a user chooses to put in May 1, 2001, and for example, the T range is 2 to 4, then T=2 is May 2, 2001 (and not January 2, 2001).

Figure 25. Priors-First tab.



4.5.3. Priors-First Tab

When clicking on the first 'Priors' tab, the user sees the window illustrated in Figure 25. The purpose of this tab is to allow the user to specify the information required to calculate the Bias Splines. Note that the selection of prior information is up to the user. For the T-SpACE model, the priors are intentionally initialized to 'non-informative' values, to ensure that they have no effect on the posterior distribution. If the user has some understanding of the statistical distribution of the data, then the user can change the priors to reflect their knowledge.

A detailed explanation of the simulation model calculations is provided in Section A. Below, in Table 2, is a description of the input a user should provide.

Table 2. Definition of Prior Parameters – Prior Screen 1.

Variable	Bias Spline Min	Bias Spline Max	number of Control Points
T	First day of the time range over which the bias will be calculated. By default, this is the first day of the year.	Last day of the time range over which the bias will be calculated. By default, this is the last day of the year.	This defines the degrees of freedom in the model for time. By default, 4 is chosen representing a control point for each season of the year.
X	Represents the min x-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the min x-coordinate is 1, for the 12 km grid. If the user is using the 36 km grid, then the default should be 13.	Represents the maximum x-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the maximum possible x-coordinate for the 12 km is given, 213. If the user is working with the 36 km grid, then this should be 142.	This is the degrees of freedom for the x-coordinate in the bias function. 8 is the default entry. Please note that if this number gets too high the performance of the simulation may be slow.
Y	Represents the min y-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the min x-coordinate is 1, for the 12 km grid. If the user is using the 36 km grid, then the default should be 15.	Represents the maximum y-coordinate of the grid over which the bias spline will be calculated. By default, and recommended, the maximum possible x-coordinate for the 12 km is given, 188. If the user is working with the 36 km grid, then this should be 94.	This is the degrees of freedom for the x-coordinate in the bias function. 7 is the default entry. Please note that if this number gets too high the performance of the simulation may be slow.

Consideration of the following should occur when choosing the number of control points:

- Since the number of control points is roughly equal to the number of degrees of freedom of the T-SpACE simulation, the user wants to be conscious of not over- or under-fitting the model by choosing too many or too few control points

- For T, time, the number of control points has been chosen to represent season. If the user believes there to be a monthly bias in the concentrations, then they may choose this value to be 12.

- For X and Y, the number of control points has been chosen to represent an estimated rate on which bias changes in each direction. A metric that may be used to choose the number of control points for X and Y could be the length of the grid over which the simulation is to run divided by the average scale of the spatial bias (every 25 grid cells).

It is recommended that there should be less than 100 control points in the x-y coordinate system.

- It is also recommended that a user of the system run sensitivity analyses to understand how sensitive the surface predictions are to the number of control points.

In addition, there are two normal regression priors which are specified on this screen.

$\mu_t \quad t = 1, \ldots, N^T$	Mean level of the Conditional Auto Regressive (CAR) process (Mu)
$\beta^D \in \mathbf{R}^{N^D}$	Vector of coefficient for the bias (BetaD)

By default, each of these is defined by a non-informative normal distribution with μ equal to 0 and the variance equal to 1.0E-3 (0.001). If the user has information about the distribution of Mu or BetaD, the user can enter the information here.

4.5.4. Priors-Second Tab

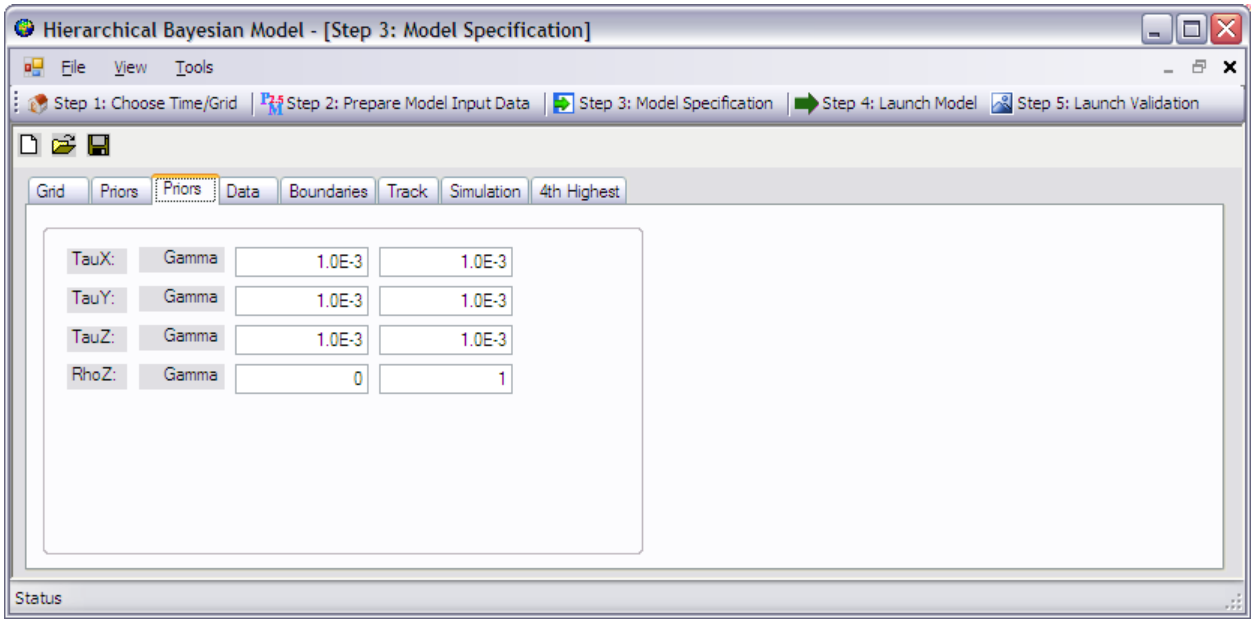
The priors being entered on this second 'Priors' tab are used to specify prior (statistical) information about the unknown, but true, variances of the air pollution monitor observations, the air quality model simulated air pollution concentration surface, and random process included in the simulation to account for presence of correlation among air pollution measurements collected successively in both space and time. In addition, the user can specify prior information about the correlation of the random process on this screen. Figure 26 illustrates the input screen for these priors. Note that if the user is interested in the details of these parameters, they should see [1] and [2].

Specifically, the τ's represent the reciprocal of the unknown variance of the air pollution monitor observations (measured concentrations), the air quality model generated air pollution concentration surface, and the random process that is a part of T-SpACE. The parameters that the user is specifying on this screen are

τ^X	Precision of the measurement error in the monitor observations (TauX)
τ^Y	Precision of the measurement error in the computer observations (TauY)
τ^Z	Precision of the mean process (TauZ)
ρ^Z	Temporal autocorrelation parameter of the mean process (RhoZ)

The τ's of the model are defined by a gamma distribution. For all three τ's, the first column represents the shape of the gamma distribution and the second column represents the rate for the distribution. By default, the τ's are uninformative (shape = 0.001, and rate = 0.001) because each prior needs to be assigned by the user with prior knowledge of the data and the model. These default choices squeeze the gamma distribution to zero. This is a non-informative choice since the sampling equation for the τ's in the posterior distributions is equal to sample size of data set plus the shape of the prior. If shape is small (i.e., the non-informative prior), adding that value does not affect the posterior. Thus, without prior knowledge of the data and the model, the non-informative priors have no effect on the posterior distribution.

Figure 26. Priors-Second Tab.



ρ (Rho) is the temporal autocorrelation parameter that is modeled by a uniform distribution to constrain the correlation between the minimum, 0, and the maximum, 1. By default, the constraints are initialized to non-informative with the first column representing the minimum range for the uniform distribution and the second column representing the maximum range. If a user has an understanding of the possible range for the correlation, they can change the default values to numbers between 0 and 1.

4.5.5 Data

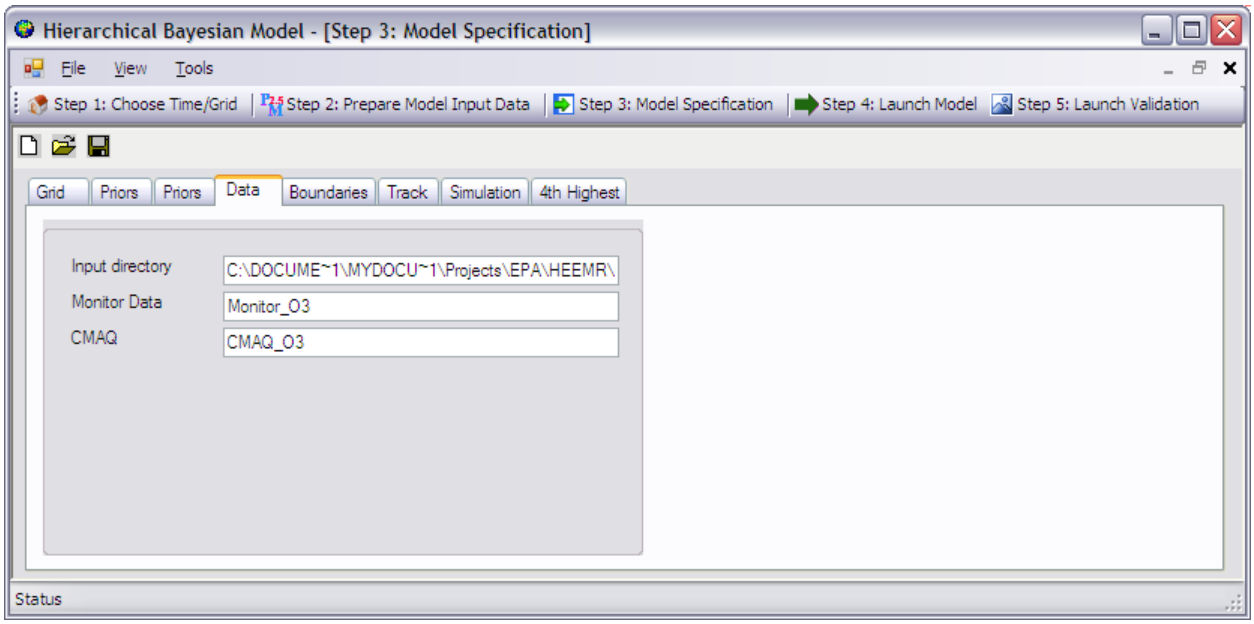
As illustrated in Figure 27, the data tab is where the user specifies the input air pollution monitor and simulated air pollution concentration surface files for the T-SpACE simulation. By default, the input directory is specified in Step 1, and the air pollution monitor data, and the CMAQ/CAMx data are the **.CSV** files created in Step 2. If the user does not want to use the default files, they can make the changes here.

Please note that if the user chooses to change the inputs, they should be as follows:

Input Directory	The path name for the location of the monitor and CMAQ/CAMx data. Please note that both sets of data need to reside in the same directory.
Monitor Data	This needs to be a **.CSV** file in the format that is created in Step 2. If the file is not of this format, then the user will receive an error.

CMAQ Data	This needs to be either a 12 km or 36 km CMAQ/CAMx **.CSV** file in the format created in Step 2. If the file is not of this format, then the user will receive an error.

Figure 27. Input Data.



4.5.6. Boundaries

As illustrated in Figure 28, the boundaries tab is where the user defines the neighborhood structure of the model. Figure 29 illustrates the four possible choices for neighborhoods in the model. By default, the simulation uses a first order neighborhood. Detailed information about the boundaries is provided in Section A.

Figure 28. Boundaries.

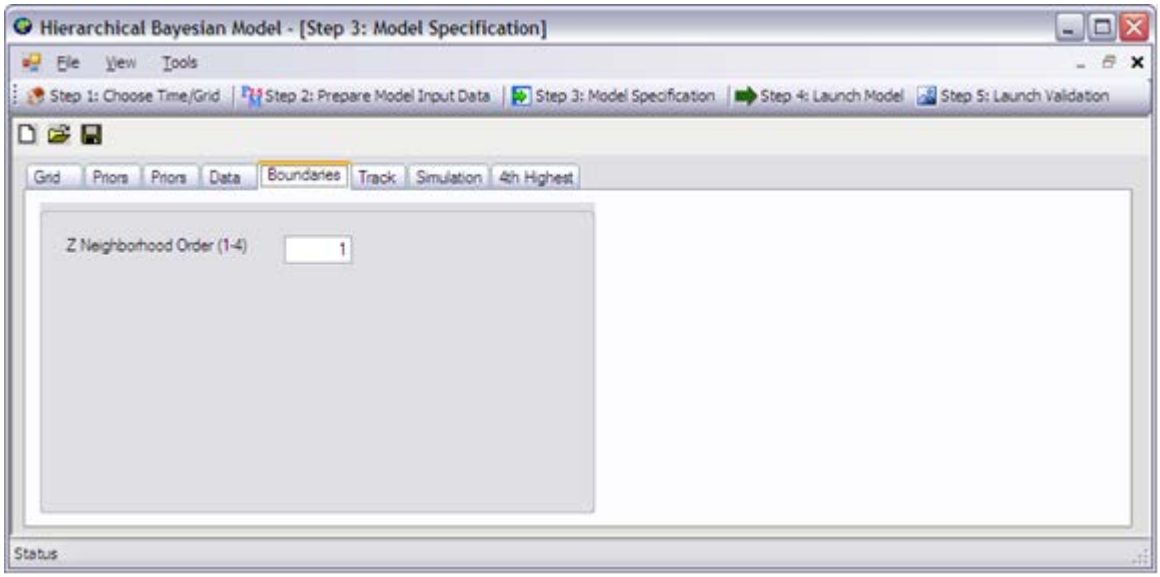
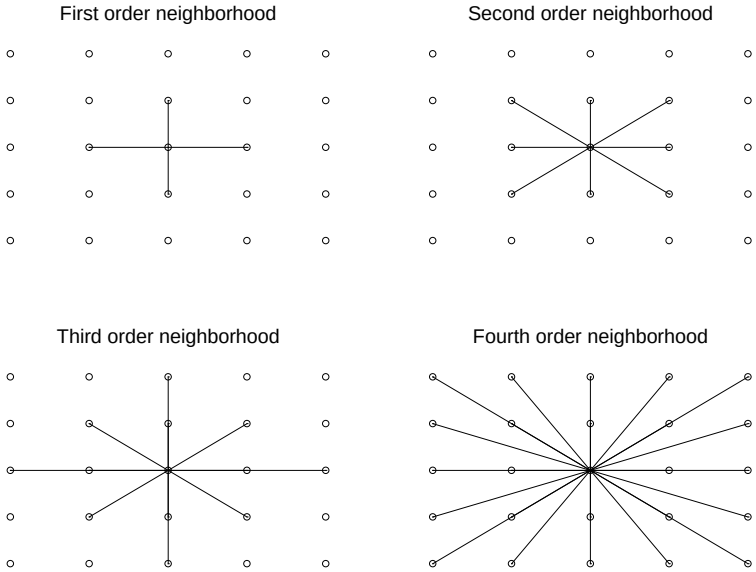


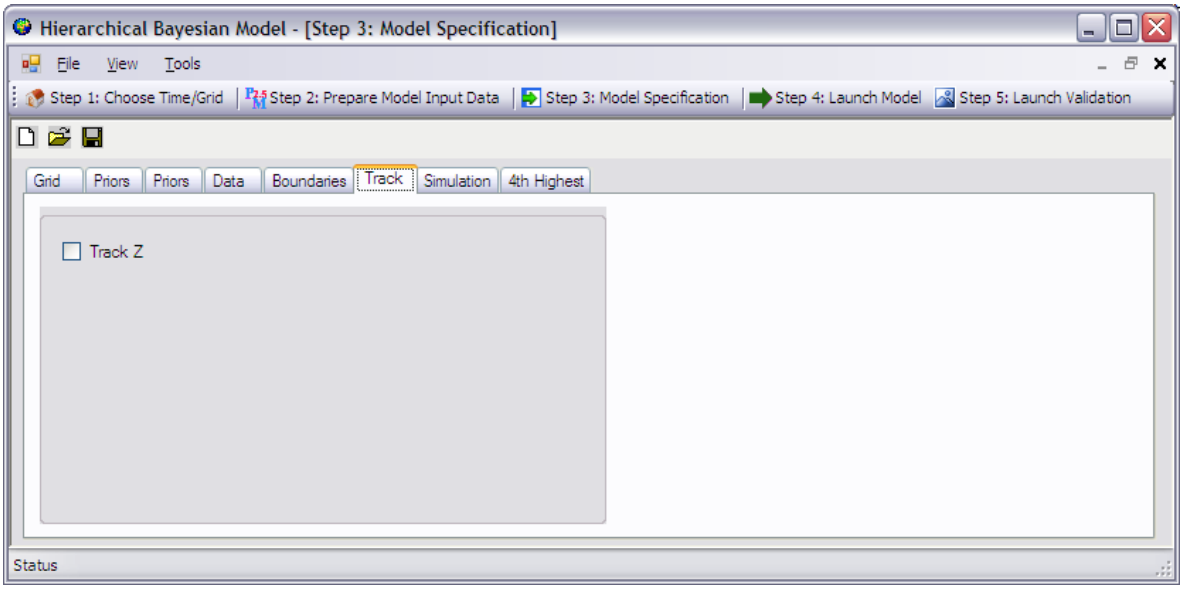
Figure 29. Boundaries-Neighborhood Definitions.



4.5.7 Track

Figure 30 illustrates where the user can track the chain flag for z during the simulation. This chain is written to a file located in the output directory specified on the Simulation screen. By default, this remains unchecked as it slows down the processing of the simulation. For diagnostic purposes, the user may want to turn this on, by marking a 'check' the 'Track Z' box.

Figure 30. Track.



4.5.8 Simulation

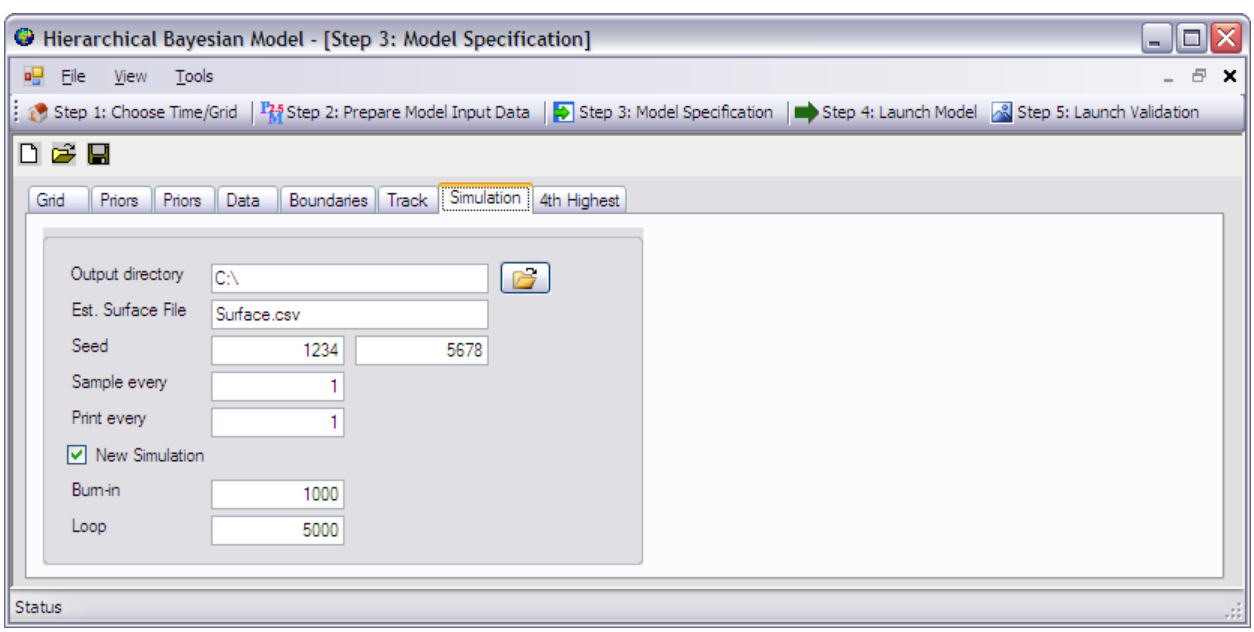
This tab is where the user specifies how the simulation will be run and where the simulation results will be saved. As shown in Figure 31, there are several inputs the user can specify.

Output Directory	By default, this is the location provided in Step 1. This directory where the user would like the generated simulation files placed.
Est. Surface File	This is the name of the file that will contain the estimated air pollution concentration surface. By default, this is named '**surface.csv**'. The user can name the file any name they want provided the extension is .CSV.
Seed	Two random seeds are need to seed the simulation. By default, the two seeds utilized are 1234 and 5678. The user can specify the seeds they would like if they prefer something other than the default values.
Sample every	During the simulation, this is how often sampling occurs. By default, this is initialized to 1 (i.e., sample each model iteration).

Print every	During the simulation, the user can see the progress on screen. This determines at what step the results are printed to the screen. By default, this is 1 (i.e., print each step on the screen). Please note that choosing a lower number may slow down the processing.
New Simulation	Choosing this indicates whether to overwrite current files or not.
Burn-in	Number of burn-ins loops for the simulation. The default is initialized to 1000.
Loop	Number of simulation loops to complete. The default is initialized to 5000.

For "Burn-in" and "Loop", the default values were chosen heuristically based on the size of the default grid. The user can utilize the chain file to help understand if their choice for the number of burn-ins and simulations loops is appropriate. In particular, TauX, TauY, and TauZ are in the chain file. Plot these parameters by the iteration. If the parameter value keeps increasing or decreasing, then those runs should be in burn-in iterations. If the parameter jumps between a range (i.e., bouncing back and forth), we consider this to be the point at which the simulation loop should begin. In terms of the total number of iterative loops, there is not a magic number. It is best if the user plots the parameter values and determines where the parameters appear to converge.

Figure 31. Simulation.



4.5.9 4th Highest

As discussed earlier in this section, under "Tools" on the top menu bar, the user can choose to calculate the average 4th-highest concentration surface using three consecutive annual (individual) 4th highest air pollution concentration surfaces. This tab is where the user can choose to create an individual 4th-highest concentration surface. As shown in Figure 32, by default, the 4th highest concentration surface is not calculated. In order to calculate the surface during the simulation, the user must check "Enable 4th Highest Concentration" and provide the name of the file for the surface. It is recommended that the user use the same name as the surface file, but add "_4h" to the filename so that the user knows which surface run the 4th highest concentration surface is related to. The note on this input screen reminds users that if they are looking to create a 3-year average 4th highest concentration surface, each of the individual runs should utilize the same time frame, the same grid size, and roughly, the same order of magnitude of burn-in and loop sizes.

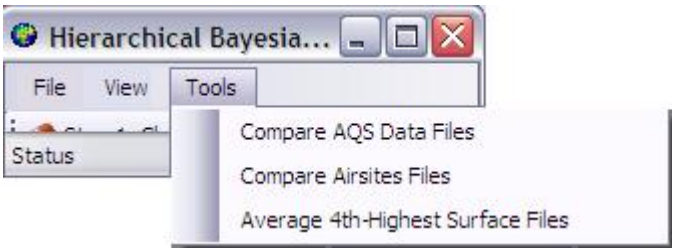
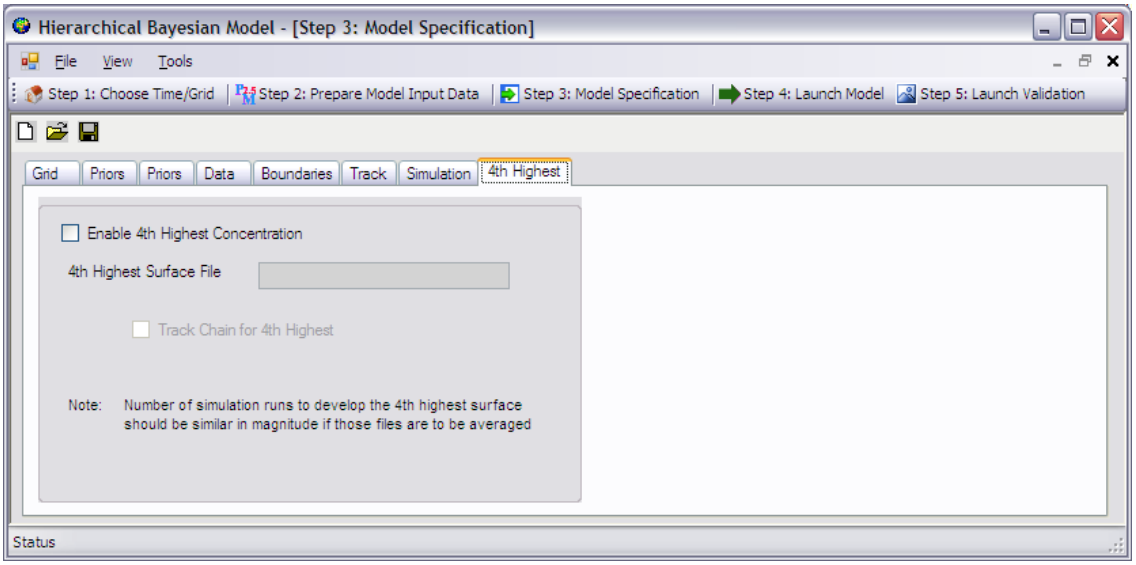


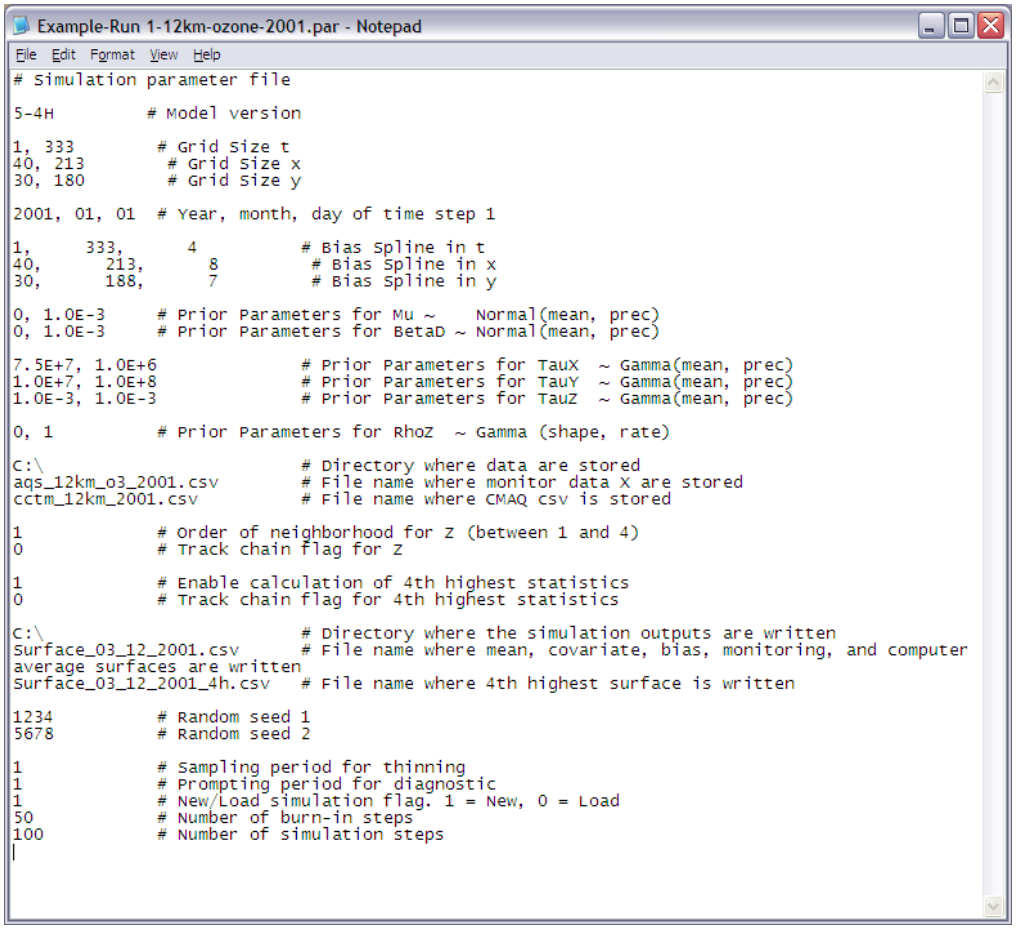
Figure 32. 4th highest concentration surface.



4.5.10 *.PAR file

Prior to moving to Step 4, the user will want to save their choices to a *.PAR (parameter) file. This file is excellent for archiving the model runs. The *.PAR file is a text file with a specific format that is needed to perform the simulation. As shown in Figure 33, the third line of the file indicates the version of the file. Currently, we are using version "5-4H". If you have an earlier *.PAR file, you may enter your parameter choices manually in Step 3 and create a new *.PAR file, or you may compare the previous *.PAR file to the current version and add the necessary lines.

Figure 33. Simulation.PAR file.



Example-Run 1-12km-ozone-2001.par - Notepad

File Edit Format View Help

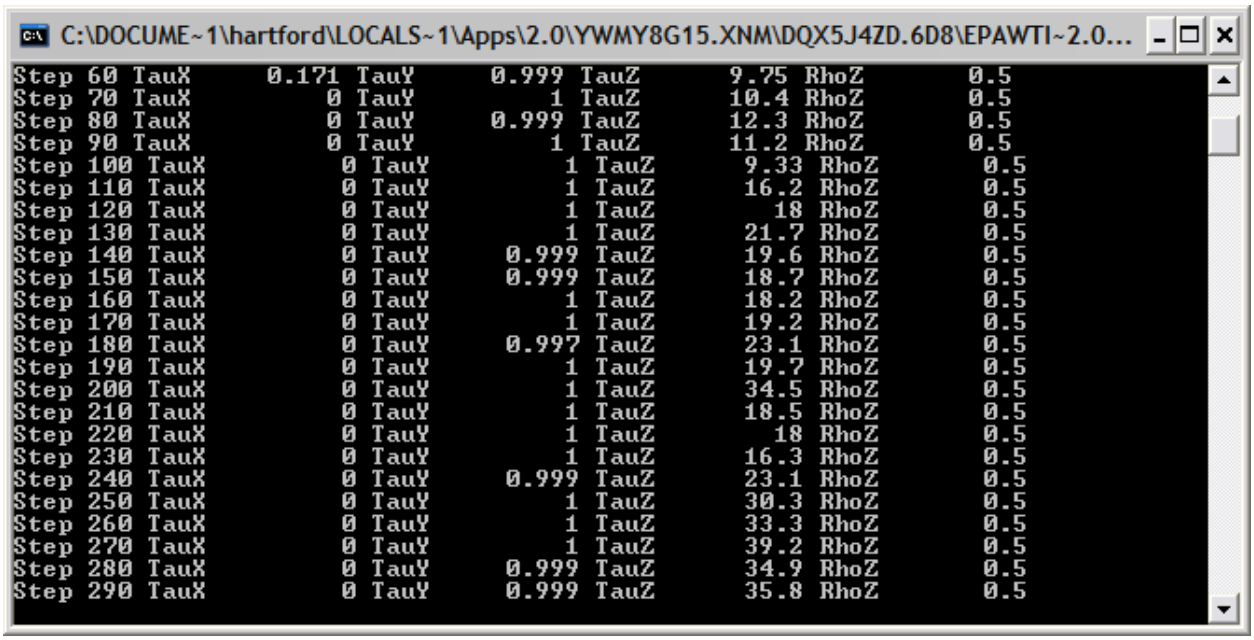
```
# Simulation parameter file

5-4H          # Model version

1, 333         # Grid Size t
40, 213         # Grid Size x
30, 180         # Grid Size y

2001, 01, 01  # Year, month, day of time step 1

1,      333,      4           # Bias Spline in t
40,       213,      8          # Bias Spline in x
30,       188,      7          # Bias Spline in y

0, 1.0E-3     # Prior Parameters for Mu ~    Normal(mean, prec)
0, 1.0E-3     # Prior Parameters for BetaD ~ Normal(mean, prec)

7.5E+7, 1.0E+6              # Prior Parameters for TauX  ~ Gamma(mean, prec)
1.0E+7, 1.0E+8              # Prior Parameters for TauY  ~ Gamma(mean, prec)
1.0E-3, 1.0E-3              # Prior Parameters for TauZ  ~ Gamma(mean, prec)

0, 1          # Prior Parameters for RhoZ  ~ Gamma (shape, rate)

C:\                         # Directory where data are stored
aqs_12km_o3_2001.csv        # File name where monitor data X are stored
cctm_12km_2001.csv          # File name where CMAQ csv is stored

1             # Order of neighborhood for Z (between 1 and 4)
0             # Track chain flag for Z

1             # Enable calculation of 4th highest statistics
0             # Track chain flag for 4th highest statistics

C:\                         # Directory where the simulation outputs are written
Surface_03_12_2001.csv      # File name where mean, covariate, bias, monitoring, and computer
average surfaces are written
Surface_03_12_2001_4h.csv   # File name where 4th highest surface is written

1234          # Random seed 1
5678          # Random seed 2

1             # Sampling period for thinning
1             # Prompting period for diagnostic
1             # New/Load simulation flag. 1 = New, 0 = Load
50            # Number of burn-in steps
100           # Number of simulation steps
```

4.6 Step 4: Launch Model

Prior to launching Step 4, the user must specify the simulated air pollution surface file in Step 2 (NetCDF format) and must open a *.PAR file or create a *.PAR file in Step 3. Once the user has specified the model information and has created a simulation.PAR file, the user should choose the "Step 4: Launch Model" tab. Hitting this tab once will initiate the T-SpACE model run.

It may take a few minutes for a DOS screen to appear. This screen will show the progression of the T-SpACE model run, including reading the data into the model, executing burn-in loop steps, and executing the simulation loop. Figure 34 illustrates an example of the output that is written to the DOS screen.

Figure 34. Step 4: DOS Screen output – progression of the simulation.



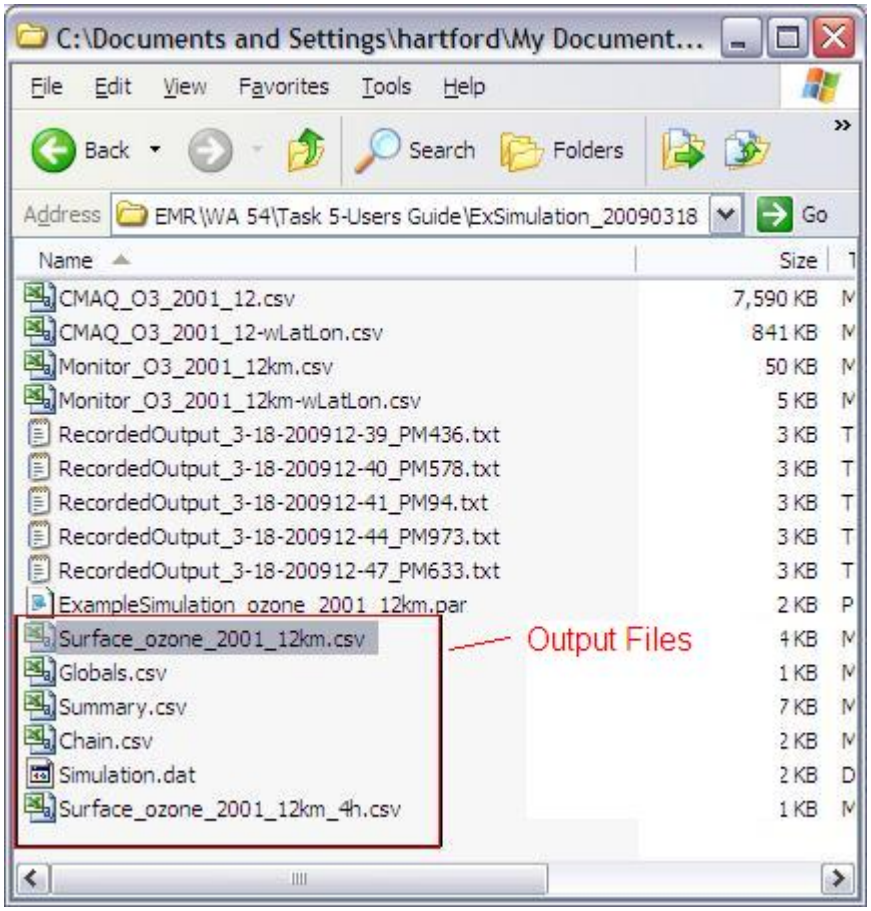
```
C:\DOCUME~1\hartford\LOCALS~1\Apps\2.0\YWMY8G15.XNM\DQX5J4ZD.6D8\EPAWTI~2.0...
Step 60 TauX      0.171 TauY      0.999 TauZ      9.75 RhoZ       0.5
Step 70 TauX          0 TauY          1 TauZ      10.4 RhoZ       0.5
Step 80 TauX          0 TauY      0.999 TauZ      12.3 RhoZ       0.5
Step 90 TauX          0 TauY          1 TauZ      11.2 RhoZ       0.5
Step 100 TauX          0 TauY          1 TauZ      9.33 RhoZ       0.5
Step 110 TauX          0 TauY          1 TauZ      16.2 RhoZ       0.5
Step 120 TauX          0 TauY          1 TauZ        18 RhoZ       0.5
Step 130 TauX          0 TauY          1 TauZ      21.7 RhoZ       0.5
Step 140 TauX          0 TauY      0.999 TauZ      19.6 RhoZ       0.5
Step 150 TauX          0 TauY      0.999 TauZ      18.7 RhoZ       0.5
Step 160 TauX          0 TauY          1 TauZ      18.2 RhoZ       0.5
Step 170 TauX          0 TauY          1 TauZ      19.2 RhoZ       0.5
Step 180 TauX          0 TauY      0.997 TauZ      23.1 RhoZ       0.5
Step 190 TauX          0 TauY          1 TauZ      19.7 RhoZ       0.5
Step 200 TauX          0 TauY          1 TauZ      34.5 RhoZ       0.5
Step 210 TauX          0 TauY          1 TauZ      18.5 RhoZ       0.5
Step 220 TauX          0 TauY          1 TauZ        18 RhoZ       0.5
Step 230 TauX          0 TauY          1 TauZ      16.3 RhoZ       0.5
Step 240 TauX          0 TauY      0.999 TauZ      23.1 RhoZ       0.5
Step 250 TauX          0 TauY          1 TauZ      30.3 RhoZ       0.5
Step 260 TauX          0 TauY          1 TauZ      33.3 RhoZ       0.5
Step 270 TauX          0 TauY          1 TauZ      39.2 RhoZ       0.5
Step 280 TauX          0 TauY      0.999 TauZ      34.9 RhoZ       0.5
Step 290 TauX          0 TauY      0.999 TauZ      35.8 RhoZ       0.5
```

At the completion of the T-SpACE model run, the DOS window will close. Figure 35 illustrates the files that are created (and stored) in the directory selected by the user for the output files. This example shows the files when the 4th highest air pollution concentration surface is also generated.

The following are the simulation results files available to the user:

Chain.CSV	Contains output that was printed to the DOS screen for each parameter in the model.
Summary.CSV	Contains summary statistics for <u>all</u> calculated parameters in the model including the bias surface. Section E provides a detailed summary of the information in this file.

Figure 35. Files saved from the simulation run.



Globals.csv Contains a brief summary of the number of burn-in loop and simulation loop steps and thinning choices made for the simulation.

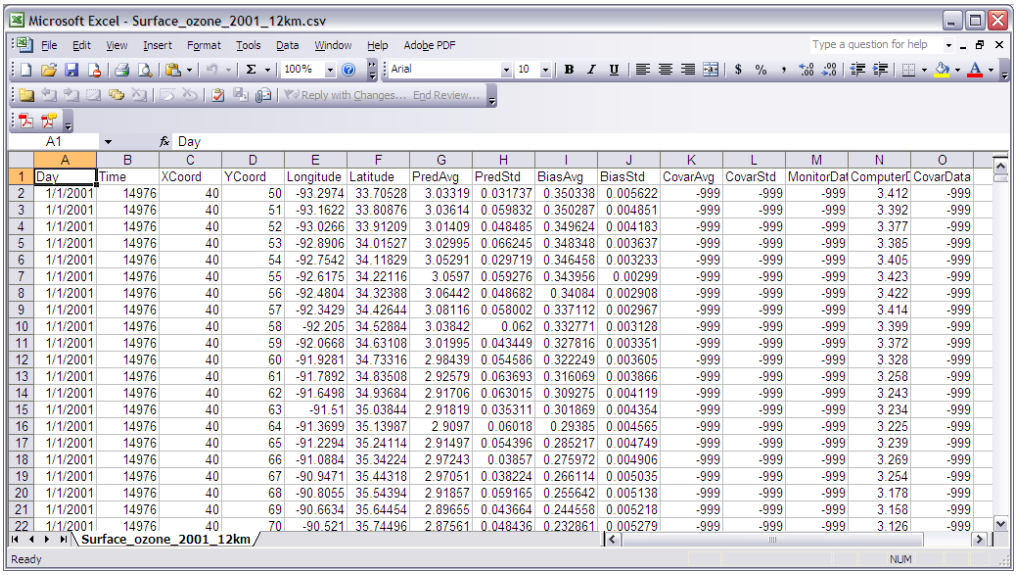
Simulation.dat This file is used for another application. It is in a machine readable format.

Surface*.csv Contains the estimated air pollution concentration surface from the T-SpACE simulation. Figure 36 provides a snapshot of the variables seen in the file. A brief description of each variable in the **Surface.csv** file is listed below.

Day	Date for the estimated surface in a MM/DD/YYYY format.
Time	This is the date in Julian format (i.e., number of Julian Day)
XCoord	The grid x-coordinate
YCoord	The grid y-coordinate
Longitude	The transformed grid x-coordinate to a Longitude (degrees)
Latitude	The transformed grid y-coordinate to a Latitude (degrees)

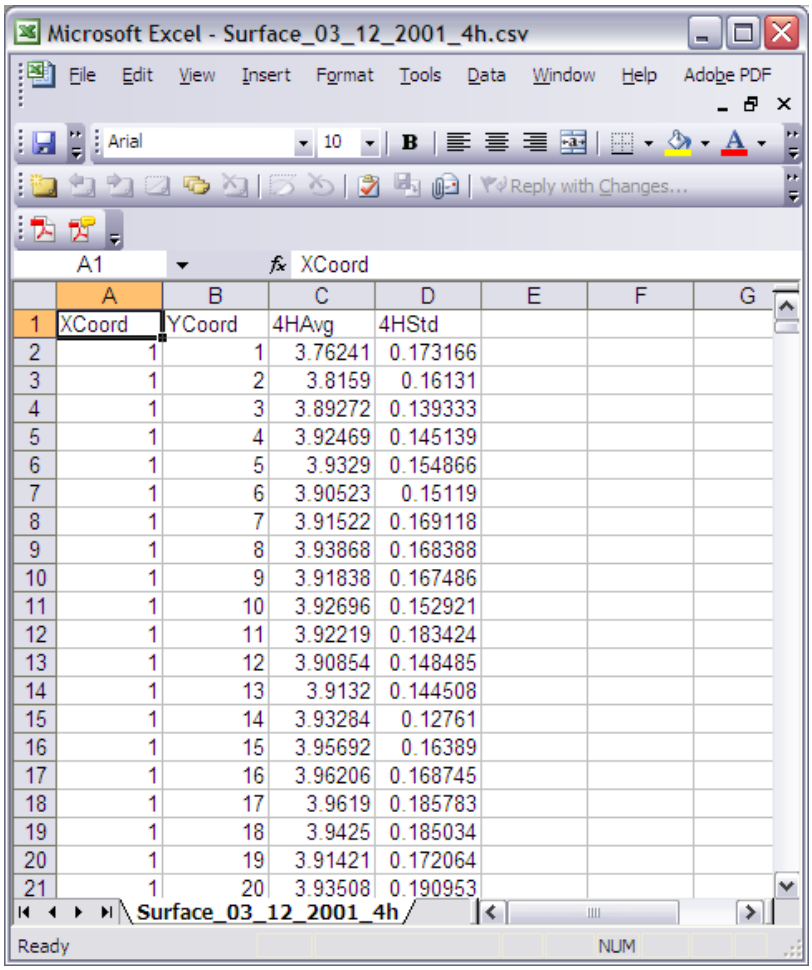
PredAvg	The natural log of the mean predicted air pollution concentration for each day and grid cell. (Values of -999 denote a missing value.) The predicted daily concentration represents a daily average for $PM_{2.5}$, and an eight-hour maximum concentration for ozone. Section A contains a detailed description of the calculation of this value.
PredStd	The natural log of the standard error for the mean predicted concentration for each day and grid cell. (Values of -999 denote a missing value.) Section A contains a detailed description of the calculation of this value.
BiasAvg	Mean of the bias surface. (Values of -999 denote a missing value)
BiasStd	Standard error of the mean of the bias surface. (Values of -999 denote a missing value.)
CovarAvg	This parameter is obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua and Terra satellites. The number -999 denotes a missing value. [Implemented in Model 6 version of T-SpACE.]
CovarStd	This parameter is obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua and Terra satellites. The number -999 denotes a missing value. [Implemented in Model 6 version of T-SpACE.]
MonitorData	The natural log-transformed air pollution measurement from any air pollution monitor present at the given location.
ComputerData	The natural log-transformed simulated concentration (CMAQ) for the day and grid cell.
CovarData	This parameter is obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua and Terra satellites. The number -999 denotes a missing value. [Implemented in Model 6 version of T-SpACE.]

Figure 36. Screen shot of the Surface.csv file.



Day	Time	XCoord	YCoord	Longitude	Latitude	PredAvg	PredStd	BiasAvg	BiasStd	CovarAvg	CovarStd	MonitorDat	ComputerD	CovarData
1/1/2001	14976	40	50	-93.2974	33.70528	3.03319	0.031737	0.350338	0.005622	-999	-999	-999	3.412	-999
1/1/2001	14976	40	51	-93.1622	33.80876	3.03614	0.059832	0.350287	0.004851	-999	-999	-999	3.392	-999
1/1/2001	14976	40	52	-93.0266	33.91209	3.01409	0.048485	0.349624	0.004183	-999	-999	-999	3.377	-999
1/1/2001	14976	40	53	-92.8906	34.01527	3.02995	0.066245	0.348348	0.003637	-999	-999	-999	3.385	-999
1/1/2001	14976	40	54	-92.7542	34.11829	3.05291	0.029719	0.346458	0.003233	-999	-999	-999	3.405	-999
1/1/2001	14976	40	55	-92.6175	34.22116	3.0597	0.059276	0.343956	0.00299	-999	-999	-999	3.423	-999
1/1/2001	14976	40	56	-92.4804	34.32388	3.06442	0.048682	0.34084	0.002908	-999	-999	-999	3.422	-999
1/1/2001	14976	40	57	-92.3429	34.42644	3.08116	0.058002	0.337112	0.002967	-999	-999	-999	3.414	-999
1/1/2001	14976	40	58	-92.205	34.52884	3.03842	0.062	0.332771	0.003128	-999	-999	-999	3.399	-999
1/1/2001	14976	40	59	-92.0668	34.63108	3.01995	0.043449	0.327816	0.003351	-999	-999	-999	3.372	-999
1/1/2001	14976	40	60	-91.9281	34.73316	2.98439	0.054586	0.322249	0.003605	-999	-999	-999	3.328	-999
1/1/2001	14976	40	61	-91.7892	34.83508	2.92579	0.063693	0.316069	0.003866	-999	-999	-999	3.258	-999
1/1/2001	14976	40	62	-91.6498	34.93684	2.91706	0.063015	0.309275	0.004119	-999	-999	-999	3.243	-999
1/1/2001	14976	40	63	-91.51	35.03844	2.91819	0.035311	0.301869	0.004354	-999	-999	-999	3.234	-999
1/1/2001	14976	40	64	-91.3699	35.13987	2.9097	0.06018	0.29385	0.004565	-999	-999	-999	3.225	-999
1/1/2001	14976	40	65	-91.2294	35.24114	2.91497	0.054396	0.285217	0.004749	-999	-999	-999	3.239	-999
1/1/2001	14976	40	66	-91.0884	35.34224	2.97243	0.03857	0.275972	0.004906	-999	-999	-999	3.269	-999
1/1/2001	14976	40	67	-90.9471	35.44318	2.97051	0.038224	0.266114	0.005035	-999	-999	-999	3.254	-999
1/1/2001	14976	40	68	-90.8055	35.54394	2.91857	0.059165	0.255642	0.005138	-999	-999	-999	3.178	-999
1/1/2001	14976	40	69	-90.6634	35.64454	2.89655	0.043664	0.244558	0.005218	-999	-999	-999	3.158	-999
1/1/2001	14976	40	70	-90.521	35.74496	2.87561	0.048436	0.232861	0.005279	-999	-999	-999	3.126	-999

Figure 37. Screen shot of the surface_4h.csv file.



	A	B	C	D
1	XCoord	YCoord	4HAvg	4HStd
2	1	1	3.76241	0.173166
3	1	2	3.8159	0.16131
4	1	3	3.89272	0.139333
5	1	4	3.92469	0.145139
6	1	5	3.9329	0.154866
7	1	6	3.90523	0.15119
8	1	7	3.91522	0.169118
9	1	8	3.93868	0.168388
10	1	9	3.91838	0.167486
11	1	10	3.92696	0.152921
12	1	11	3.92219	0.183424
13	1	12	3.90854	0.148485
14	1	13	3.9132	0.144508
15	1	14	3.93284	0.12761
16	1	15	3.95692	0.16389
17	1	16	3.96206	0.168745
18	1	17	3.9619	0.185783
19	1	18	3.9425	0.185034
20	1	19	3.91421	0.172064
21	1	20	3.93508	0.190953

Surface_4h.csv Contains the results of the calculation of the 4th highest air pollution concentration surface for the specified grid and time frame. The result is a single air pollution concentration surface. Below is a description of the variables included in this file. Section F provides a summary of the calculations for results shown.

XCoord The cell's x-coordinate (east-west) within the grid (where the coordinate value increases by 1 with each grid cell as you move from west to east).

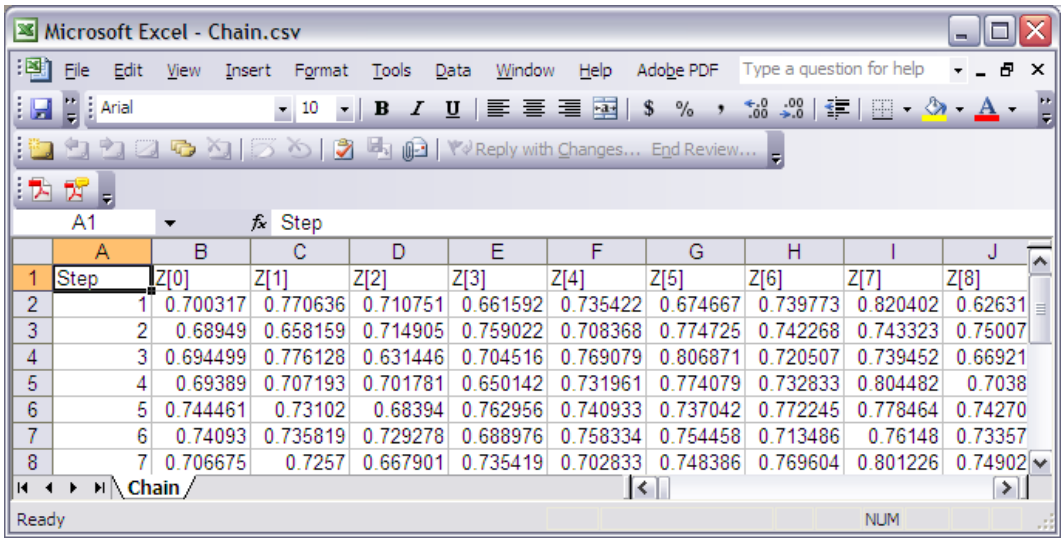
YCoord The cell's y-coordinate (north-south) within the grid (where the coordinate value increases by 1 with each grid cell as you move from south to north).

PredAvg The (natural log-transformed) T-SpACE-predicted value of the fourth highest concentration for the given grid cell.

PredStd The (natural log-transformed) T-SpACE-predicted value of the standard error for the fourth highest concentration for the given grid cell.

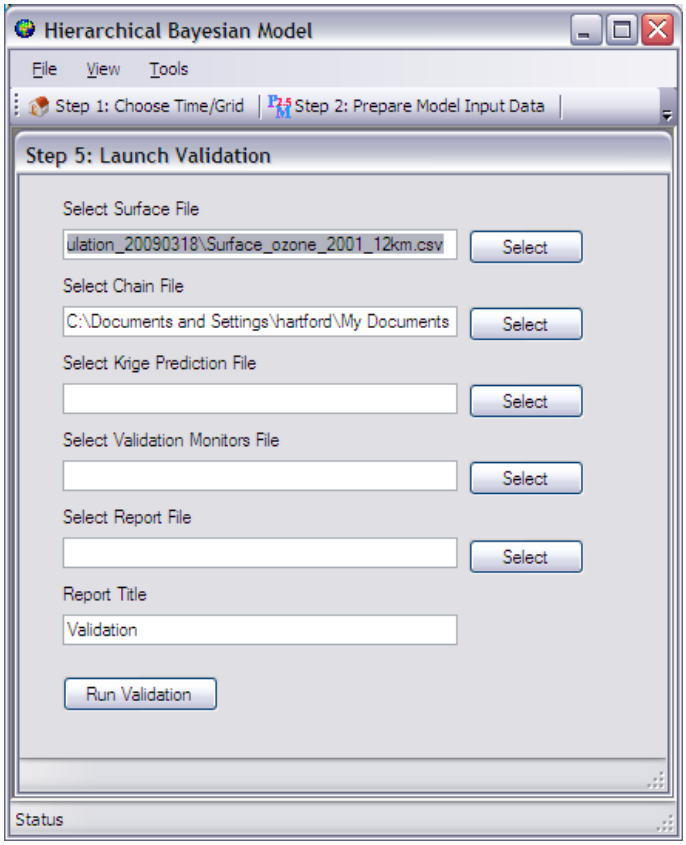
Chain.csv As seen in Figure 38, this file contains the values of Z for each T-SpACE simulation loop and each cell in the grid. This file is needed for Step 5.

Figure 38. Chain.csv.



Step	Z[0]	Z[1]	Z[2]	Z[3]	Z[4]	Z[5]	Z[6]	Z[7]	Z[8]
1	0.700317	0.770636	0.710751	0.661592	0.735422	0.674667	0.739773	0.820402	0.62631
2	0.68949	0.658159	0.714905	0.759022	0.708368	0.774725	0.742268	0.743323	0.75007
3	0.694499	0.776128	0.631446	0.704516	0.769079	0.806871	0.720507	0.739452	0.66921
4	0.69389	0.707193	0.701781	0.650142	0.731961	0.774079	0.732833	0.804482	0.7038
5	0.744461	0.73102	0.68394	0.762956	0.740933	0.737042	0.772245	0.778464	0.74270
6	0.74093	0.735819	0.729278	0.688976	0.758334	0.754458	0.713486	0.76148	0.73357
7	0.706675	0.7257	0.667901	0.735419	0.702833	0.748386	0.769604	0.801226	0.74902

Figure 39. Step 5: Launch Validation.



4.7 Step 5: Launch Validation

This step allows the user to run several validation procedures for the results from the T-SpACE simulation. The validation procedures used to analyze the output from T-SpACE are written in SAS and requires SAS to be installed on the computer used to validate the model. The file that is used to run the validation is called **ValidationTemplate.sas**. The instance of the file is loaded along with the T-SpACE software at installation time. It can be found in the main directory of the T-SpACE software. The user should not make any changes to this file, because Step 5 needs to have it in this format to appropriately run the validation process. A listing of the file is provided in Section B.

If a user has gone from Step 1 to Step 5 sequentially, then all information needed for the validation will have been generated. If a user is choosing to run the validation separately from a model run, then the following MUST be specified in Step 1 in order for the SAS program to be configured correctly:

- The appropriate dates associated with the T-SpACE model output file need to be specified here.

- The grid that was used in the T-SpACE model run needs to be specified here.

Figure 39 illustrates the initial screen when Step 5 is chosen.

4.7.1 Select Surface File

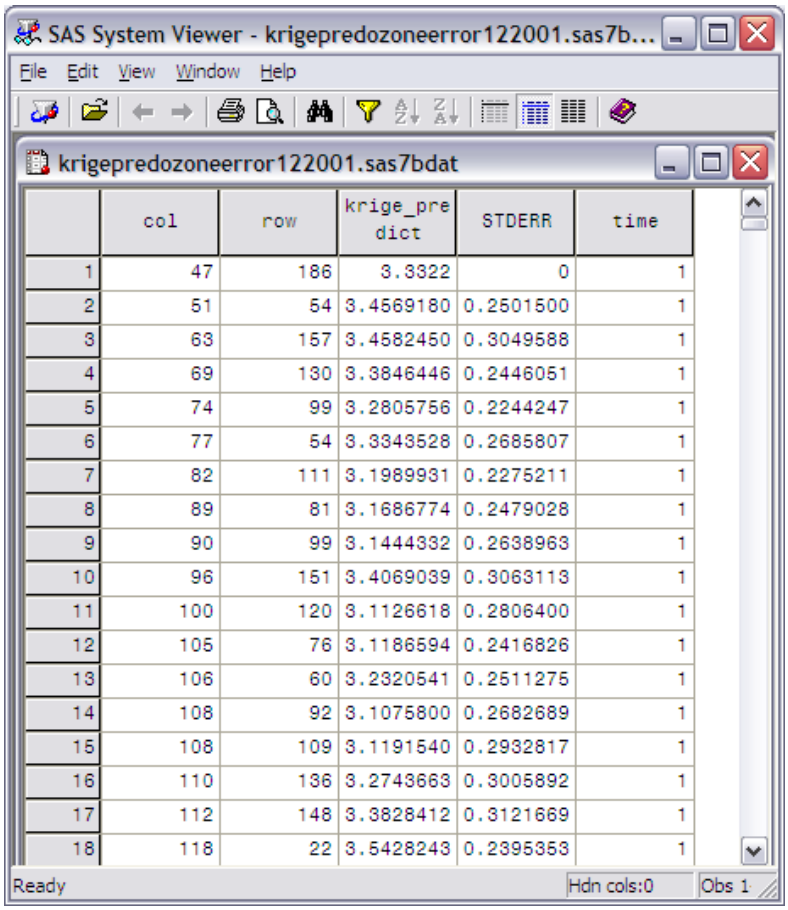
This is the file where the predicted air pollution surface concentration(s) is written. By default, if Step 3 was completed during this session, then the name of output .CSV file specified on the **Simulation** tab will be listed. Otherwise, this entry will list the last file used for a validation and the user will have to supply the name of an existing surface file that contains the predicted air pollution surface concentrations. Note that this file must have the same format as the file illustrated in Figure 36.

Please note that if real-time scanning software is installed on the computer and is running, then choosing a surface file can take several minutes if the file is large, such as a surface file for a full year of data (e.g., a full 12 km x 12 km grid).

4.7.2 Select Chain File

The validation process requires that a Chain file be available for the analysis. If Step 3 (Simulation) was completed and "Track Z" was checked on the **Track** tab, then a file named "**Chain.CSV**" will be in the same folder as the surface file specified in Step 3 on the **Simulation** tab. Otherwise, this entry will list the last file used for a validation, and the user will need to supply an existing chain file. This is a required file to run the T-SpACE model validation.

Figure 40. Krige Prediction File.



	col	row	krige_predict	STDERR	time
1	47	186	3.3322	0	1
2	51	54	3.4569180	0.2501500	1
3	63	157	3.4582450	0.3049588	1
4	69	130	3.3846446	0.2446051	1
5	74	99	3.2805756	0.2244247	1
6	77	54	3.3343528	0.2685807	1
7	82	111	3.1989931	0.2275211	1
8	89	81	3.1686774	0.2479028	1
9	90	99	3.1444332	0.2638963	1
10	96	151	3.4069039	0.3063113	1
11	100	120	3.1126618	0.2806400	1
12	105	76	3.1186594	0.2416826	1
13	106	60	3.2320541	0.2511275	1
14	108	92	3.1075800	0.2682689	1
15	108	109	3.1191540	0.2932817	1
16	110	136	3.2743663	0.3005892	1
17	112	148	3.3828412	0.3121669	1
18	118	22	3.5428243	0.2395353	1

4.7.3 Select Krige Prediction File

The krige prediction file is a SAS data set that has been developed using a series of programs detailed in Section G. Figure 40 illustrates the format of one of these files for ozone, 2001, 12 km x 12 km.

Col	Is the x-coordinate on the grid where a monitor was available for kriging.
Row	Is the y-coordinate on the grid where a monitor was available for kriging.
Krige_predict	Is the predicted natural log transformed concentration using a typical kriging method that is detailed in [3].
StdErr	Is the log of the standard error of the predicted concentration.
Time	number of days (Julian Days) since January 1.

Four files that have been prepared provided to EPA are as follows:

krigepredozoneerror122001.sas7bdat	2001, 12 km predicted error file for kriging the CASTNET ozone data.

krigepredozoneerror122002.sas7bdat 2002, 12 km predicted error file for kriging the CASTNET ozone data.

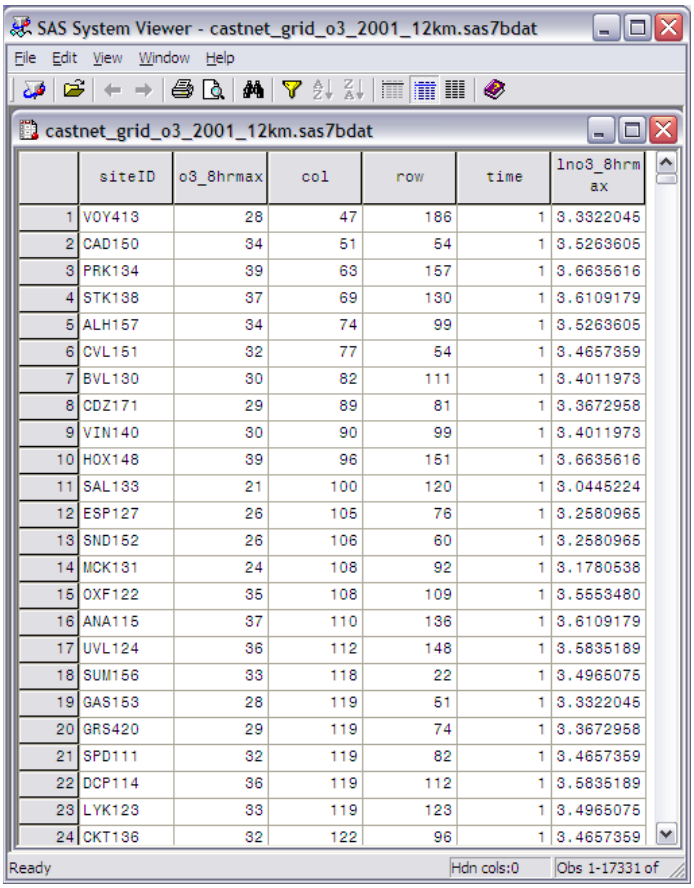
krigepredozoneerror362001.sas7bdat 2001, 36 km predicted error file for kriging the CASTNET ozone data.

krigepredozoneerror362002.sas7bdat 2002, 36 km predicted error file for kriging the CASTNET ozone data.

4.7.4 Select Validation Monitors File

The validation monitor file is a SAS data set that has been developed for EPA using the program and data detailed in Section G. Figure 41 illustrates the format of one of these files for ozone, 2001, 12 km x 12 km.

Figure 41. Validation file for ozone – uses CASTNET data.



	siteID	o3_8hrmax	col	row	time	lno3_8hrmax
1	VOY413	28	47	186	1	3.3322045
2	CAD150	34	51	54	1	3.5263605
3	PRK134	39	63	157	1	3.6635616
4	STK138	37	69	130	1	3.6109179
5	ALH157	34	74	99	1	3.5263605
6	CVL151	32	77	54	1	3.4657359
7	BVL130	30	82	111	1	3.4011973
8	CDZ171	29	89	81	1	3.3672958
9	VIN140	30	90	99	1	3.4011973
10	HOX148	39	96	151	1	3.6635616
11	SAL133	21	100	120	1	3.0445224
12	ESP127	26	105	76	1	3.2580965
13	SND152	26	106	60	1	3.2580965
14	MCK131	24	108	92	1	3.1780538
15	OXF122	35	108	109	1	3.5553480
16	ANA115	37	110	136	1	3.6109179
17	UVL124	36	112	148	1	3.5835189
18	SUM156	33	118	22	1	3.4965075
19	GAS153	28	119	51	1	3.3322045
20	GRS420	29	119	74	1	3.3672958
21	SPD111	32	119	82	1	3.4657359
22	DCP114	36	119	112	1	3.5835189
23	LYK123	33	119	123	1	3.4965075
24	CKT136	32	122	96	1	3.4657359

As seen in Section G, the data needed to create the SAS data set are the CASTNET and IMPROVE/STN monitors for ozone and $PM_{2.5}$, respectively. EPA provided the appropriate monitor data, and using the SAS programs highlighted on the diagram in Section G, this data was

then placed on the appropriate CMAQ/CAMx grid for comparison to the AQS monitors. The four ozone data files created are listed below.

castnet_grid_o3_2001_12km.sas7bdat 2001, 12 km CASTNET ozone data.

castnet_grid_o3_2002_12km.sas7bdat 2002, 12 km CASTNET ozone data.

castnet_grid_o3_2001_36km.sas7bdat 2001, 36 km CASTNET ozone data.

castnet_grid_o3_2002_36km.sas7bdat 2002, 36 km CASTNET ozone data.

A similar set of $PM_{2.5}$ files using the IMPROVE/STN are also available. All of these files are available to the user.

4.7.5 Select Report File

The user is asked to choose a file name for the file where the SAS program will write its output. This will be a plain text file. The SAS log will be written to the same folder as the report file. This is a required entry.

4.7.6 Report Title

The user is asked to supply a title for the SAS reports that are generated during this session. This is a required entry.

4.7.7 Run Validation

Clicking this button will launch the SAS program. The cursor will change to an hourglass while the validation program is executing and will return to the default shape when it completes. Note that if you have a large simulation, this can take up to 20 minutes to run.

Immediately, upon initiation, the SAS program is written to the directory as

{Report File name}-SAS Code.sas

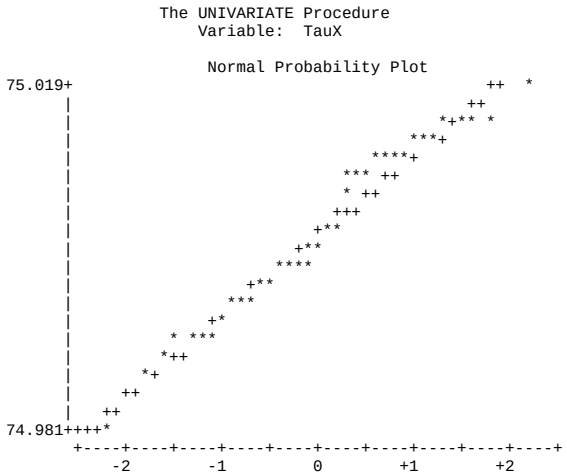
This provides the user an archive of the code that was run for the validation and allows the user to run the code through SAS separately if they desire. Once complete, the specified report file and the SAS log file will be available in the directory that the user specified.

Section C provides a listing of an example report for a validation performed on a 12 km, 2001, ozone model run.

The output provided in the validation report is as follows:

1. Distributional summary statistics for each of the model parameters τ^X, τ^Y, τ^Z is listed below:

```
The UNIVARIATE Procedure
                        Variable:  TauX

                             Moments

N                          50    Sum Weights                 50
Mean                75.002386    Sum Observations     3750.1193
Std Deviation      0.00890517    Variance             0.0000793
Skewness           -0.1340276    Kurtosis            -0.8349036
Uncorrected SS     281267.899    Corrected SS         0.0038858
Coeff Variation    0.01187318    Std Error Mean      0.00125938


                    Basic Statistical Measures

           Location                    Variability

      Mean     75.00239     Std Deviation            0.00891
      Median   75.00090     Variance               0.0000793
      Mode     74.99030     Range                    0.03710
                            Interquartile Range      0.01490

NOTE: The mode displayed is the smallest of 3 modes with a count of 2.


                    Tests for Location: Mu0=0

          Test           -Statistic-    -----p Value------

          Student's t    t  59554.95    Pr > |t|    <.0001
          Sign           M        25    Pr >= |M|   <.0001
          Signed Rank    S     637.5    Pr >= |S|   <.0001


                     Quantiles (Definition 5)

                      Quantile      Estimate

                      100% Max       75.0184
                      99%            75.0184
                      95%            75.0152
                      90%            75.0138
                      75% Q3         75.0105
                      50% Median     75.0009
                      25% Q1         74.9956
                      10%            74.9905
                      5%             74.9881
                      1%             74.9813
                      0% Min         74.9813


                         Variable:  TauX

                       Extreme Observations

             ------Lowest-----        -----Highest-----

                Value      Obs           Value      Obs

              74.9813        1         75.0141       36
              74.9872       23         75.0146       48
              74.9881       44         75.0152       28
              74.9903       34         75.0154       37
              74.9903       33         75.0184        4


           Stem Leaf                     #             Boxplot
          75018 4                        1                |
          75016                                           |
          75014 1624                     4                |
          75012 604                      3                |
          75010 0455688                  7             +-----+
          75008 3558                     4             |     |
          75006 2                        1             |     |
          75004                                        |     |
          75002 0101                     4             |  +  |
          75000 153                      3             *-----*
          74998 0347908                  7             |     |
          74996 689                      3             |     |
          74994 22456                    5             +-----+
          74992 8                        1                |
          74990 3376                     4                |
          74988 1                        1                |
          74986 2                        1                |
          74984                                           |
          74982                                           |
          74980 3                        1                |
                ----+----+----+----+
            Multiply Stem.Leaf by 10**-3
```



2. A summary of all three parameters, τ^X, τ^Y, τ^Z, number of iterations, the mean of the simulated values, and the standard error observed from the simulated values.

Obs	n TauX	n TauY	n TauZ	mnTauX	mnTauY	mnTauZ	seTauX	seTauY	seTauZ
1	50	50	50	75.0024	0.16231	3.70759	.001259381	.000035409	0.11092

3. A summary of the performance of the Mean Square Error (MSE) and bias for the T-SpACE, kriging, and the CMAQ model surface.

Obs	Overall MSE HBM	Overall MSE Krige	Overall MSE CMAQ	Overall Bias HBM	Overall Bias Krige	Overall Bias CMAQ
1	0.10	0.04	0.06	0.03	0.06	-0.09

4. A summary of the estimated percentage of time (across all monitoring days and locations) that the calculated prediction intervals (i.e., 95% credible intervals for the T-SpACE model) actually included the observed monitor value, using data for those locations in the validation dataset that contain monitors.

Obs	ndays	Pct_Time_ Krige_ worse	krigeworse	Pct_Time_ CMAQ_worse	CMAQworse
1	333	3.60	12	11.71	39

Obs	nsites	Pct_site_ Krige_ worse	krigeworse	Pct_site_ CMAQ_worse	CMAQworse
1	41	9.76	4	31.71	13

5. A summary of the prediction intervals that represent bounds on the range of measurements that are expected to contain the (unknown) true value a certain percent of the time (typically, 95%) if the model assumptions are correct.

```
Obs    modelCI    krigeCI         predStd

 1     0.97616    0.84095    0.4232558534
```

A more detailed discussion of these summary statistics can be found in [2].

5.0 References

[1] McMillan, N.J., Holland, D.M., Morara, M., and Feng, J., Combining different sources of particulate data using Bayesian space-time modeling, Environmetrics, 2010, Volume 21, pp 48 – 65, DOI: 10.1002/env.984

[2] USEPA, Final Report Overview of EPA's Hierarchical Bayesian Model For Predicting Air Quality Patterns In The United States Over Space And Time, For Use With Public Health Tracking Data, March 13, 2009, Contract No. EP-D-04-068, Work Assignment 54.

[3] USEPA, Draft Report Hierarchical Bayesian Model Evaluation For Ozone Data, March 13, 2009, Contract No. EP-D-04-068, Work Assignment 54.

SECTION A:
Detailed Description Of The
Markov Chain Monte Carlo (MCMC) Model

COVARIATE SPACE-TIME AUTO-REGRESSIVE (STAR) {MCMC} MODEL

Model 5.1

Data:

N^T Number of time points
N^P Number of space points
$N = N^T \times N^P$ Number of events (space-time points)

$x_i \in \mathbf{R}^{N_i^X} \quad i = 1,\ldots,N$	Monitoring data at event i
$X_i \in \mathbf{R} \quad i = 1,\ldots,N$	Sum of the monitoring data at event i (the number of monitoring data for each event *i* can be 0, 1 or more than 1).
$y_i \in \mathbf{R} \quad i = 1,\ldots,N$	CMAQ data at event i (the number of CMAQ data for each event *i* is always and only 1).
$D_{ij} \in \mathbf{R} \quad i = 1,\ldots,N \quad j = 1,\ldots,N^D$	Bias *jth* basis function evaluated at event *i*

The bias is evaluated as a linear combination of 2nd order uniform B-spline functions defined over a 3-dimensional lattice of uniform knots. The coefficients of the linear combination represent unidimensional control points, that is,

$$Bias = \sum_{j=1}^{N^D} D_{ij}\beta_j^D \ .$$

If we indicate with N_1, N_2, N_3 the dimensions of the CMAQ grid (that is, $N_1 = N^T$, $N_2 \cdot N_3 = N^P$ and $N_1 \cdot N_2 \cdot N_3 = N$), and with M_1, M_2, M_3 the dimensions of the control-points grid (this defines the degrees of freedom of the bias, that is, $M_1 \cdot M_2 \cdot M_3 = N^D$), and we decompose the indexes as: $i = i_1 + N_1(i_2 + N_2 i_3)$, $j = j_i + M_1(j_2 + M_2 j_3)$, then the bias matrix is then defined as

$$D_{ij} = b_{j_1}(i_1) b_{j_2}(i_2) b_{j_3}(i_3)$$

where $b_k(u)$ is the 2nd order *k*th B-spline basis function evaluated at the parameter point *u*. Setting $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ the domain over which the bias is defined (such domain must be bigger than or equal to the CMAQ grid domain), the uniform knots vectors over which the B-spline basis functions are defined are respectively

$U_r = \{a_r, a_r, a_r, a_r + s_r, a_r + 2s_r \ldots, a_r + (M_r - 3)s_r, b_r, b_r, b_r\} \quad r = 1,2,3$

where $s_r = \dfrac{b_r - a_r}{M_r - 2} \quad r = 1,2,3$.

Using B-splines as basis functions for the bias allows control of the degrees of freedom of the bias through the number of control points. Furthermore, the piece-wise nature of the B-spline functions respects the principle of locality, that is, local information does not affect regions far from the region where the local information is defined. On the numerical side, B-splines allow a tensor factorization of the bias matrix into three matrixes $B^r_{i_r j_r} = b_{j_r}(i_r), r = 1,2,3$ for a total dimension of $N_1M_1 + N_2M_2 + N_3M_3$, very much less than the total dimension of the full D-matrix, which is $N_1M_1 \cdot N_2M_2 \cdot N_3M_3$.

Parameters

Z_i	$i = 1,\ldots,N$	CAR process at event i.
μ_t	$t = 1,\ldots,N^T$	Mean level of the CAR process.
$\beta^D \in \mathbf{R}^{N^D}$		Vector of coefficient for the bias.
τ^X		Precision of the measurement error in the monitor observations.
τ^Y		Precision of the measurement error in the computer observations.
τ^Z		Precision of the mean process.
ρ^Z		Temporal autocorrelation parameter of the mean process.

Likelihood:

$$\left[x_{ik} \mid \mu_{t(i)}, Z_i, \tau^X\right] = N\left(w_i, \tau^X\right),$$

$$\left[y_i \mid \mu_{t(i)}, \beta^D, Z_i, \tau^Y\right] = N\left(w_i + \beta^D D_i, \tau^Y\right),$$

where

$$w_i = \mu + Z_i .$$

Priors:

$$\left[Z \mid \tau^Z, \rho^Z\right] = N\left(0, \tau^Z\left(\Lambda^T\left(\rho^Z\right) \otimes \Lambda^P\right)\right)$$

where $\Lambda^T(\rho)$ is the precision matrix corresponding to a time autoregressive model with parameter ρ and Λ^P is the precision matrix corresponding to a space autoregressive model of order r with a zero boundary condition.

Writing

$$n_t^T = \begin{cases} 1 & t = 1, N^T \\ 1 + \left(\rho^Z\right)^2 & 1 < t < N^T \end{cases}$$

$$n^P = \text{ number of spatial neighbors}$$

and

$$\mu_i^z = \frac{\rho^Z}{n_{t(i)}^T} \sum_{j \in \partial_t i} Z_j + \frac{1}{n^P} \sum_{j \in \partial_p^r i} Z_j - \frac{\rho^Z}{n_{t(i)}^T n^P} \sum_{j \in \partial_t i \times \partial_p^r i} Z_j$$

$$\tau_i^z = \tau^Z \frac{n_{t(i)}^T n^P}{1 - \left(\rho^Z\right)^2}$$

where $\partial_t i$ denotes the time first nearest neighbors' events of the event i, $\partial_p^r i$ is the set of space r nearest neighbors' events of the event i, the prior conditional distribution for a single element of Z can be written as

$$\left[Z_i \mid Z_{i-}, \tau^Z, \rho^Z\right] = N\left(\mu_i^z, \tau_i^z\right)$$

where the minus after a subscript denotes the set of all subscripts not including the one shown.

The priors corresponding to the other parameters are

$$\left[\mu_t\right] = N\left(\theta^\mu, \tau^\mu\right)$$

$$\left[\beta_i^D\right] = N\left(\theta^D, \tau^D\right)$$

$$\left[\tau^X\right] = G\left(\gamma^X, \delta^X\right)$$

$$\left[\tau^Y\right] = G\left(\gamma^Y, \delta^Y\right)$$

$$\left[\tau^Z\right] = G\left(\gamma^Z, \delta^Z\right)$$

$$\left[\rho^Z\right] \sim U\left(a^Z, b^Z\right)$$

Full Model

$$\left[Z, \mu, \beta^D, \tau^X, \tau^Y, \tau^Z, \rho^Z \mid X, Y, S\right] \propto$$

$$\prod_{i=1}^{N} \prod_{k=1}^{N_i^X} \left[x_{ik} \mid Z_i, \mu_{t(i)}, \tau^X\right] \times$$

$$\prod_{i=1}^{N} \left[y_i \mid Z_i, \mu_{t(i)}, \beta^D, \tau^Y\right] \times$$

$$\prod_{i=1}^{N}\left[Z_i \mid Z_{i-}, \tau^Z, \rho^Z\right]\times$$

$$[\mu][\beta^D][\tau^X][\tau^Y][\tau^Z][\rho^Z]$$

Explicitly:

$$\left[Z, \mu, \beta^D, \tau^X, \tau^Y, \tau^Z, \rho^Z \mid X, Y\right]\propto$$

$$\prod_{i=1}^{N}\left(\frac{\tau^X}{2\pi}\right)^{\frac{N_i^X}{2}} \exp\left\{-\frac{\tau^X}{2}\sum_{i=1}^{N}\sum_{k=1}^{N_i^X}\left[x_{ik}-\left(\mu_{t(i)}+Z_i\right)\right]^2\right\}\times$$

$$\prod_{i=1}^{N}\left(\frac{\tau^Y}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\tau^Y}{2}\sum_{i=1}^{N}\left[y_i-\left(\mu_{t(i)}+Z_i+\beta^D D_i\right)\right]^2\right\}\times$$

$$\left(\tau^Z\right)^{\frac{N}{2}}\left|\Lambda^T\left(\rho^Z\right)\otimes\Lambda_r^P\right|^{\frac{1}{2}} \exp\left\{-\frac{\tau^Z}{2}(Z)^T\left[\Lambda^T\left(\rho^Z\right)\otimes\Lambda_r^P\right]Z\right\}\times$$

$$[\mu][\beta^D][\tau^X][\tau^Y][\tau^Z][\rho^Z]$$

Full Conditional Distributions

Variable: $Z_i \in \mathbf{R}$

$$\left[Z_i \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A = \tau^X N_i^X + \tau^Y + \tau_i^z$$

$$B = \tau^X\left[X_i - N_i^X \mu_{t(i)}\right] + \tau^Y\left[y_i - \left(\mu_{t(i)} + \beta^D D_i\right)\right] + \tau_i^z \mu_i^z$$

Variable: $\mu_t \in \mathbf{R}$

$$\left[\mu_t \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A = \tau^X \sum_{i\in I_t} N_i^X + \tau^Y N^T + \tau^\mu$$

$$B = \tau^X \sum_{i\in I_t}\left[X_i - N_i^X Z_i\right] +$$

$$\tau^Y \sum_{i\in I_t}\left[y_i - \left(Z_i + \beta^D D_i\right)\right] + \tau^\mu \theta^\mu$$

Variable: $\beta^D \in \mathbf{R}^{N^D}$

$$\left[\beta^D \mid -\right] = N\left(A^{-1}B, A^{-1}\right)$$

where

$$A_{kl} = \tau^Y \sum_{i=1}^{N} N_i^Y D_{ik} D_{il} + \delta_{kl}\tau^D$$

$$B_k = \tau^Y \sum_{i=1}^{N} D_{ik}\left[y_i - \left(\mu_{t(i)} + Z_i\right)\right] + \tau^D \theta^D$$

Variable: τ^X

$$\left[\tau^X \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}\sum_{i=1}^{N} N_i^X + \gamma^X$$

$$B = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N_i^X}\left[x_{ik} - \left(\mu_{i(t)} + Z_i\right)\right]^2 + \delta^X$$

Variable: τ^Y

$$\left[\tau^Y \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}N + \gamma^Y$$

$$B = \frac{1}{2}\sum_{i=1}^{N}\left[y_i - \left(\mu_{t(i)} + Z_i + \beta^D D_i\right)\right]^2 + \delta^Y$$

Variable: τ^Z

$$\left[\tau^Z \mid -\right] = G\left(A, B\right)$$

where

$$A = \frac{1}{2}N + \gamma^z$$

$$B = \frac{1}{2\tau^Z}\sum_{i=1}^{N}\tau_i^z Z_i\left(Z_i - \mu_i^z\right) + \delta^Z$$

Variable: ρ^Z

$$\left[\rho^Z \mid -\right] \propto \chi_{\left(a^Z, b^Z\right)}\left(\rho^Z\right)\left(1-\left(\rho^Z\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^Z\right)Z_i\left(Z_i - \mu_i^z\left(\rho^Z\right)\right)\right\}$$

This full conditional is not a recognized form, so it has to be sampled using a Metropolis-Hastings step.

Jump:

$$\rho' \sim J\left(\left|\rho' - \rho^Z\right|\right)$$

Acceptance:

$$\min\left\{\frac{\chi_{\left(a^Z, b^Z\right)}\left(\rho'\right)\left(1-\left(\rho'\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho'\right)Z_i\left(Z_i - \mu_i^z\left(\rho'\right)\right)\right\}}{\left(1-\left(\rho^Z\right)^2\right)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\tau_i^z\left(\rho^Z\right)Z_i\left(Z_i - \mu_i^z\left(\rho^Z\right)\right)\right\}}, 1\right\}$$

SECTION B:
VALIDATION SAS PROGRAM EXAMPLE

```
*-----------------------------------------------------*
Program Name:     ValidationTemplate.sas
Purpose:          Summarizes the model parameters.
*-----------------------------------------------------*;
*=====================================================================
* Inputs
* This program requires the following
* Hierarchical Bayes Model (HBM) output files:
*   Chain.csv
*   SurfaceModel.csv
*
* Kriged surface output SAS dataset:
*
* Monitor data for validation SAS dataset:
*
* This script provides statistics on the differences between
* the HBM, Kriged, and CMAQ predicted surfaces and the actual monitor data, on a log scale;
* (technically, CMAQ is modeled, not predicted).
* Statistics computed for each monitor location are 'mse' and 'bias'
* where mse is squared error for each monitor location, averaged either daily or by site;
* and bias provides the raw error for each location.
* (The error is the difference between the prediction and the monitor measured value).
* Statistics are computed by day ('time'), by site, and for all the data overall.
*
* Performance is indicated by the average differences in mse,
* and by how often the HBM mse is smaller than the Kriged or CMAQ surfaces.
* Specifically:
* Average differences between the Kriged mse and HBM mse,
* and between CMAQ mse and HBM mse,
* indicate whether the HBM squared-error is smaller or larger on average
* relative to Kriging or CMAQ.
* HBM performance is also rated by how often the daily HBM mse
* is smaller than those of the Kriged or CMAQ predictions;
* and by how often site-by-site mse is smaller.
* Finally, the prediction intervals of both the HBM model and the Kriging method
* are checked to determine how often the intervals
* include the true value measured by the monitor;

%let LocSurfaceData = $$SurfaceFileFolder$$;
%let LocChainData = $$ChainFileFolder$$;
%let LocMonitorData = $$MonitorFileFolder$$;
%let LocKrigeData = $$KrigeFileFolder$$;
```

```
/* HBM data (Hierarchical Bayesian Model) */
%let inChainFile = $$ChainFileName$$;
%let modelSurfaceFile = $$SurfaceFileName$$;

/* Kriged data*/
%let krigedSurface = $$KrigeFileName$$;

/* Monitor data: Combined Improve and STN for PM2.5; CASTNet for O3*/
%let monitorData = $$MonitorFileName$$;

* 'afterDate' is the day before the start date;
%let afterDate = '$$AfterDate$$'d;

/* Report Title */
%let rptTtl = $$ReportTitle$$;

/* Define the range */
%let minX = $$xMin$$;
%let maxX = $$xMax$$;
%let minY = $$yMin$$;
%let maxY = $$yMax$$;

/* These are the reading columns for data */
%let monRdg = $$monRdg$$;
%let logRdg = $$logRdg$$;
%let Valcols = &monRdg. &logRdg.;

libname sd "&LocMonitorData";
libname sd2 "&LocKrigeData";

filename rptFile '$$ReportFile$$';
filename logFile '$$LogFile$$';

/* macro to determine if a column exists in a dataset */
%macro ColumnExist(DS=,Var=);
%global found;
%let dsid = %sysfunc(open(&DS,i));
%if &dsid le 0 %then %do;
  %put WARNING: %sysfunc(sysrc()) -- %sysfunc(SysMsg());
  %let found = -1;
%end;
%else %do;
  %let dsname = %sysfunc(DsName(&dsid));
  %put NOTE: Opened dataset: &dsname;
  %let NVar = %sysfunc(attrn(&dsid,NVARS));
```

```
  %put NOTE: There are &Nvar Variables in &dsname;
  %let found = 0;
    %do i = 1 %to &Nvar;
      %let vn&i = %sysfunc(varname(&dsid,&i));
      %if %upcase("&&vn&i") = %upcase("&VAR") %then %do;
     %let found = 1;
  %goto ColdStop;
      %end;
 %ColdStop:
 %end;
  %if &found = 1 %then %put NOTE: Variable &Var was found.;
  %else %put NOTE: Variable &Var was not found.;
%end;

%let rc = %sysfunc(close(&dsid));
%mend ColumnExist;

/* end macro */

proc printto log = logFile print = rptFile New;
run;

*-------------------------------------------------------------
* Calculate the summary statistics on the HBM model parameters
*-------------------------------------------------------------;
proc import out      = chain
           datafile = "&LocChainData.\&inChainFile"
           dbms     = csv replace;

/* tau is the precision parameter in HBM
   x denotes FRM monitor precision
   y denotes CMAQ 'precision'
   z denotes precision for Conditional autoregressive (CAR) process
These are estimated from the characteristics of the HBM Markov chain
seTauX is se_tauX; etc.
*/

proc univariate data = chain plot;
    var tauX tauY tauZ;
    output out = sumParms
          mean = mnTauX mnTauY mnTauZ
          stdErr = seTauX seTauY seTauZ
          n = nTauX nTauY nTauZ;

proc print data = sumParms;
       title "Summary Statistics for &rptTtl";
```

```
run; quit;

/* create a couple of macro variables */
/* posterior variance to be computed from x and z precision ******/
data _null_;
   set sumParms;
   call symput("sigmaX", 1 / mnTauX);
   call symput("sigmaZ", 1 / mnTauZ);
run;

/* dump values to log */
%put &sigmaX;
%put &sigmaZ;

*--- Comparison summaries ---*;
* Read in HBM predicted surface, referred to as modelsurface;
proc import out      = hbm
           datafile = "&LocSurfaceData.\&modelSurfaceFile"
           dbms     = csv replace;

data hbm_surfacePreds(keep = time xCoord yCoord predAvg predStd computerData test
                    rename = (xCoord = col yCoord = row predAvg = predSurface computerData =
cmaqSurface));
   set hbm;

   *--- Remove the data where a prediction was not made ---*;
   time = time - &afterDate.;
   if predAvg ^= -999;
   test = predStd * predStd;

   *--- This step exponentiates the results back to normal units ---*;
   array vars [4] predAvg biasAvg monitorData computerData;
   do i = 1 to 4;
      if vars[i] ne -999 then vars[i] = exp(vars[i]);
   end;

   *--- The bias is transformed ---*;
   if vars[2] ne -999 then vars[2] = (vars[2] - 1) * vars[1];
   drop i;

   *--- For the validation results, any monitor result greater than 600 change to -999 ---*;
   if vars[3] > 600 then vars[3] = -999;

   /* covarData is the 15th variable (and last) in the record. Check to see if it contains
      a carriage return (ascii code 13. or '0D'x). If not, take all of the 15th variable as
      the value of covarData. If it does contain a carriage return then only use the data up
```

```
     to, but not including, the carriage return as the value of covarData */
   if index(var15,'0D'x) = 0 then covarData = input(var15, 8.);
   else covarData = input(substr(var15, 1, index(var15, '0D'x) - 1), 8.);

proc sort data = hbm_surfacePreds;
   by time col row;

data check;
   set hbm_surfacePreds;
   diff = predSurface - cmaqSurface;

proc sort data = check;
   by time;

proc means data = check noprint;
   by time;
   var diff;
   output out = sumt(drop = _type_ _freq_)
         mean = mnDiff
         stdErr = seDiff
         n = nDiff;
run; quit;

proc sort data = check;
   by col row;

proc means data = check noprint;
   by col row;
   var diff;
   output out = sumcr(drop = _type_ _freq_)
          mean = mnDiff
          stdErr = seDiff
          n = nDiff;
run; quit;

*--- Kriging data for comparison ---*;
data krige_predict;
   set sd2.&krigedSurface.;
   krigePred = exp(krige_predict);
run; quit;

proc sort data=krige_predict;
   by time col row;
run;

*--- IMPROVE, STN, or CASTNet, data for comparison ---*;
```

```
%ColumnExist(DS=sd.&monitorData.,Var=site);
%let foundsite = &found;

%ColumnExist(DS=sd.&monitorData.,Var=site_code);
%let foundsitecode = &found;

%ColumnExist(DS=sd.&monitorData.,Var=logpm);
%let foundlogpm = &found;

data tempValidation;
   set sd.&monitorData.;
   if &foundsite = 1 then;
      rename site = siteID;
run; quit;

data tempValidation;
   set tempValidation;
   if &foundsitecode = 1 then;
      rename site_code = siteID;
run; quit;

data tempValidation;
   set tempValidation;
   if &foundlogpm = 1 then;
      pm = exp(logpm);
run; quit;

data surf_validation (keep = time col row site &Valcols.);
   set tempValidation (rename = (siteID = site));
   where (col ge &minX.) and (col le &maxX.) and (row ge &minY.) and (row le &maxY.);
run; quit;

proc sort data = surf_validation;
   by time col row;
run;

*-----------------------------------------------------------------------------
*     Combine the model prediction, Kriging results, and the IMPROVE/STN data
*   NOTE:  The bias is calculated as bias = validation value - predicted value
*-----------------------------------------------------------------------------;
data combined;
   merge krige_predict
         surf_validation (in = inValidation)
         hbm_surfacePreds (in = inPred);
```

```
   by time col row;

   *--- Data to keep in the file ---*;
   if inValidation;
   if predSurface ^= .;

   if cmaqSurface = -999 then cmaqSurface = .;
   if predSurface = -999 then predSurface = .;

   *--- Calculate the summary stats ---*;
   *--- regular units ---*;
   mse_model = (predSurface - &monRdg.) ** 2;
   mse_krige = (krigePred - &monRdg.) ** 2;
   mse_CMAQ = (cmaqSurface - &monRdg.) ** 2;

   mse_diff = mse_krige - mse_model;
   mse_diff_2 = mse_CMAQ - mse_model;

   bias_model = &monRdg. - predSurface;
   bias_krige = &monRdg. - krigePred;
   bias_CMAQ = &monRdg. - cmaqSurface;

   *--- log units ---*;
   mse_lmodel = (log(predSurface) - &logRdg.) ** 2;
   mse_lkrige = (krige_predict - &logRdg.) ** 2;
   mse_lCMAQ = (log(cmaqSurface) - &logRdg.) ** 2;

   * compare MSE of Krige to HBM and of CMAQ to HBM;
   mse_ldiffKrige = mse_lkrige - mse_lmodel;
   mse_ldiffCMAQ = mse_lcmaq - mse_lmodel;

   bias_lmodel = &logRdg. - log(predSurface);
   bias_lkrige = &logRdg. - krige_predict;
   bias_lCMAQ = &logRdg. - log(cmaqSurface);
run;

*--- Summarize performance by time ---*;
proc sort data = combined out = combined_pltnt;
   by time;
run;

*--- Average daily MSE for HBM, Kriging, and CMAQ ---*;
* avg mse_ldiff... > 0 implies HBM error is smaller on average;
* se_mse... provides measure of variation in the magnitude of square errors;

proc means data = combined_pltnt noprint;
```

```
   by time;
   var mse_lmodel mse_lkrige mse_lcmaq mse_ldiffkrige mse_ldiffcmaq
      bias_lmodel bias_lkrige bias_lCMAQ;
   output out = mse_pltnt (drop = _type_ _freq_)
         mean = mse_lmodel mse_lkrige mse_lcmaq
               mse_ldiffkrige mse_ldiffcmaq
               bias_lmodel bias_lkrige bias_lCMAQ
         stdErr = se_mse_lmodel se_mse_lkrige se_mse_lcmaq
                 se_mse_ldiffkrige se_mse_ldiffcmaq
                 se_bias_lmodel se_bias_lkrige se_bias_lCMAQ
         n = n_mse_lmodel n_mse_lkrige n_mse_lcmaq
            n_mse_ldiffkrige n_mse_ldiffcmaq
            n_bias_lmodel n_bias_lkrige n_bias_lCMAQ;

*--- Annual MSE for HBM, Kriging, and CMAQ.  Bias summary for HBM, kriging, and CMAQ
---*;
proc means data=combined_pltnt noprint;
   var mse_lmodel mse_lkrige mse_lcmaq mse_ldiffkrige mse_ldiffcmaq
      bias_lmodel bias_lkrige bias_lCMAQ;
   output out = overall_mse_pltnt (drop = _type_ _freq_)
         mean = mse_lmodel mse_lkrige mse_lcmaq
               mse_ldiffkrige mse_ldiffcmaq
               bias_lmodel bias_lkrige bias_lCMAQ;

proc print data = overall_mse_pltnt label;
   var mse_lmodel mse_lkrige mse_lcmaq
      bias_lmodel bias_lkrige bias_lCMAQ;
   label mse_lmodel = 'Overall MSE HBM'
         mse_lkrige = 'Overall MSE Krige'
         mse_lcmaq = 'Overall MSE CMAQ'
         bias_lmodel = 'Overall Bias HBM'
         bias_lkrige = 'Overall Bias Krige'
         bias_lcmaq = 'Overall Bias CMAQ';
   format mse_lmodel mse_lkrige mse_lcmaq
          bias_lmodel bias_lkrige bias_lCMAQ 5.2;
run; quit;

*--- Summarize performance by site ---*;
proc sort data = combined_pltnt;
   by site;

proc means data = combined_pltnt noprint;
   by site;
   var mse_lmodel mse_lkrige mse_lcmaq
      mse_ldiffkrige mse_ldiffcmaq
      bias_lmodel bias_lkrige bias_lCMAQ;
```

```
   output out = mse_pltnt2
         mean = mse_lmodel mse_lkrige mse_lcmaq
               mse_ldiffkrige mse_ldiffcmaq
               bias_lmodel bias_lkrige bias_lCMAQ;

*--- Comparing HBM to CMAQ and Krige based on MSE ---*;
*    mse_pltnt is daily statistics (by 'time');
data mse_pltnt1;
   set mse_pltnt;
   retain sum_1 sum_2;

   if mse_ldiffkrige > 0 then do;
      krige_larger = 1;
      sum_1 + 1;
   end;

   if mse_ldiffcmaq > 0 then do;
      cmaq_larger = 1;
      sum_2 + 1;
   end;

proc means data = mse_pltnt1 noprint;
   var sum_1 sum_2 time;
   output out = Improve_Time(drop = _type_ _freq_ t1 t2)
         max = krigeworse CMAQworse
         n = t1 t2 ndays;

data Improve_Time;
   set Improve_Time;
   Pct_Time_Krige_worse = (krigeworse / ndays) * 100;
   Pct_Time_CMAQ_worse = (cmaqworse / ndays) * 100;

proc print data = Improve_Time;
   title1 "&rptTtl: % Improvement over Kriging and CMAQ";
   var ndays pct_time_krige_worse krigeworse pct_time_CMAQ_worse cmaqworse;
   format pct_time_krige_worse pct_time_CMAQ_worse 7.2;
run; quit;

* mse_pltnt2 is statistics by site;
data mse_pltnt3;
   set mse_pltnt2;
   retain sum_1 sum_2;

   if mse_ldiffkrige > 0 then do;
      krige_larger = 1;
      sum_1 + 1;
```

```
   end;

   if mse_ldiffcmaq > 0 then do;
      cmaq_larger = 1;
      sum_2 + 1;
   end;
run; quit;

proc means data = mse_pltnt3 noprint;
   var sum_1 sum_2 _freq_;
   output out = Improve_site(drop = _type_ _freq_ t1 t2)
         max = krigeworse CMAQworse
         n = t1 t2 nsites;

data Improve_site;
   set Improve_site;
   Pct_site_Krige_worse = (krigeworse / nsites) * 100;
   Pct_site_CMAQ_worse = (cmaqworse / nsites) * 100;

proc print data = improve_site;
   title1 "&rptTtl: % Improvement over Kriging and CMAQ";
   var nsites pct_site_krige_worse krigeworse pct_site_CMAQ_worse cmaqworse;
   format pct_site_krige_worse pct_site_CMAQ_worse 7.2;
run; quit;

*-----------------------------------------------------------------------------------------
* 95% Prediction Interval : this interval is calculated using modelstd=sqrt(sigmaZ+sigmaX)
* MCMC Prediction Interval: this interval is calculated using
modelstd=sqrt(predStd*predStd+sigmaX)
* here begin to calculate the prediction interval
*-----------------------------------------------------------------------------------------;
* pm_surfacePreds is HBM predicted surface;
* sigmaX is posterior variance of FRM data;
* sigmaZ is posterior variance of CAR process;

data compare;
   merge krige_predict(keep = time col row krige_predict stdErr)
         hbm_surfacePreds(keep = time col row predSurface cmaqSurface predStd)
         surf_validation(in = Validation);
   by time col row;

   if Validation;
   if predSurface ^= .;
   sigmaX = &sigmaX;
   sigmaZ = &sigmaZ;
   modelstd = sqrt(sigmaZ + sigmaX);
```

```
* keep track of confidence interval coverage of prediction vs. monitor measurement;
   if ((&logRdg. > log(predSurface) - 2 * modelstd) and (&logRdg. < log(predSurface) + 2 *
modelstd)) then modelCI = 1;
   else modelCI = 0;

   if ((&logRdg. > krige_predict - 2 * stdErr) and (&logRdg. < krige_predict + 2 * stdErr)) then
krigeCI = 1;
   else krigeCI = 0;

proc means data = compare noprint;
   var modelCI krigeCI predStd;
   output out = PredictionInterval(drop = _TYPE_ _FREQ_)
         mean = modelCI krigeCI predStd;

proc print data = predictioninterval;
   var modelCI krigeCI predStd;
   title1 "&rptTtl: Average Prediction Interval";
run; quit;

proc printto;
run;
```

Section C:
Validation Report Example

```
                    The SAS System         14:36 Friday, April 24, 2009   1

                 The UNIVARIATE Procedure
                     Variable:  TauX

                          Moments

N                          50    Sum Weights                 50
Mean                 75.002386    Sum Observations     3750.1193
Std Deviation      0.00890517    Variance             0.0000793
Skewness           -0.1340276    Kurtosis            -0.8349036
Uncorrected SS     281267.899    Corrected SS         0.0038858
Coeff Variation    0.01187318    Std Error Mean      0.00125938


                 Basic Statistical Measures

          Location                    Variability

      Mean     75.00239     Std Deviation            0.00891
      Median   75.00090     Variance               0.0000793
      Mode     74.99030     Range                    0.03710
                            Interquartile Range      0.01490

NOTE: The mode displayed is the smallest of 3 modes with a count of 2.


                 Tests for Location: Mu0=0

          Test           -Statistic-    -----p Value------

          Student's t    t  59554.95    Pr > |t|    <.0001
          Sign           M        25    Pr >= |M|   <.0001
          Signed Rank    S     637.5    Pr >= |S|   <.0001


                Quantiles (Definition 5)

                 Quantile      Estimate

                 100% Max       75.0184
                 99%            75.0184
                 95%            75.0152
                 90%            75.0138
                 75% Q3         75.0105
                 50% Median     75.0009
                 25% Q1         74.9956
                 10%            74.9905
                 5%             74.9881
                 1%             74.9813
                 0% Min         74.9813
```

The UNIVARIATE Procedure
Variable: TauX

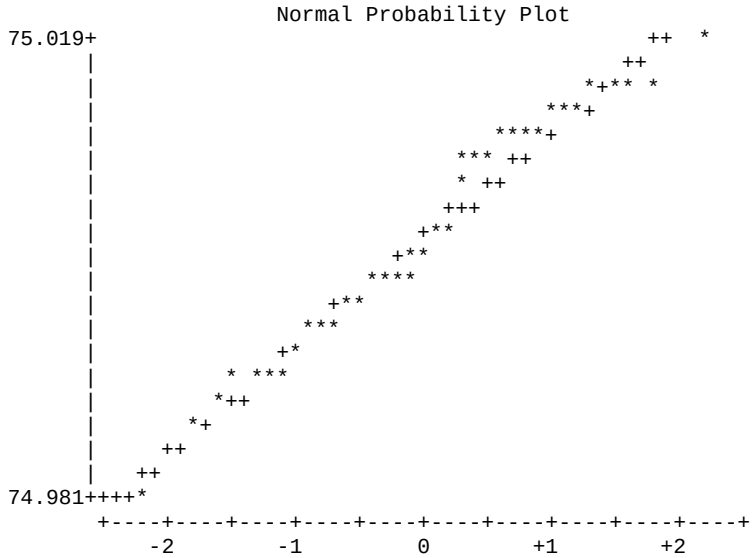
Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
74.9813	1	75.0141	36
74.9872	23	75.0146	48
74.9881	44	75.0152	28
74.9903	34	75.0154	37
74.9903	33	75.0184	4

```
 Stem Leaf                     #             Boxplot
75018 4                        1                |
75016                                           |
75014 1624                     4                |
75012 604                      3                |
75010 0455688                  7             +-----+
75008 3558                     4             |     |
75006 2                        1             |     |
75004                                        |     |
75002 0101                     4             |  +  |
75000 153                      3             *-----*
74998 0347908                  7             |     |
74996 689                      3             |     |
74994 22456                    5             +-----+
74992 8                        1                |
74990 3376                     4                |
74988 1                        1                |
74986 2                        1                |
74984                                           |
74982                                           |
74980 3                        1                |
      ----+----+----+----+
  Multiply Stem.Leaf by 10**-3
```

The UNIVARIATE Procedure
Variable: TauX

Normal Probability Plot



The UNIVARIATE Procedure
Variable: TauY

Moments

N	50	Sum Weights	50
Mean	0.16230594	Sum Observations	8.115297
Std Deviation	0.00025038	Variance	6.26906E-8
Skewness	-0.3695166	Kurtosis	-1.0537027
Uncorrected SS	1.31716398	Corrected SS	3.07184E-6
Coeff Variation	0.15426482	Std Error Mean	0.00003541

Basic Statistical Measures

Location		Variability	
Mean	0.162306	Std Deviation	0.0002504
Median	0.162334	Variance	6.26906E-8
Mode	0.162334	Range	0.0008990
		Interquartile Range	0.0004300

NOTE: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	4583.72	Pr > |t|	<.0001
Sign	M	25	Pr >= |M|	<.0001
Signed Rank	S	637.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	0.162701
99%	0.162701
95%	0.162670
90%	0.162592
75% Q3	0.162512
50% Median	0.162334
25% Q1	0.162082
10%	0.161920
5%	0.161907
1%	0.161802
0% Min	0.161802

The UNIVARIATE Procedure
Variable: TauY

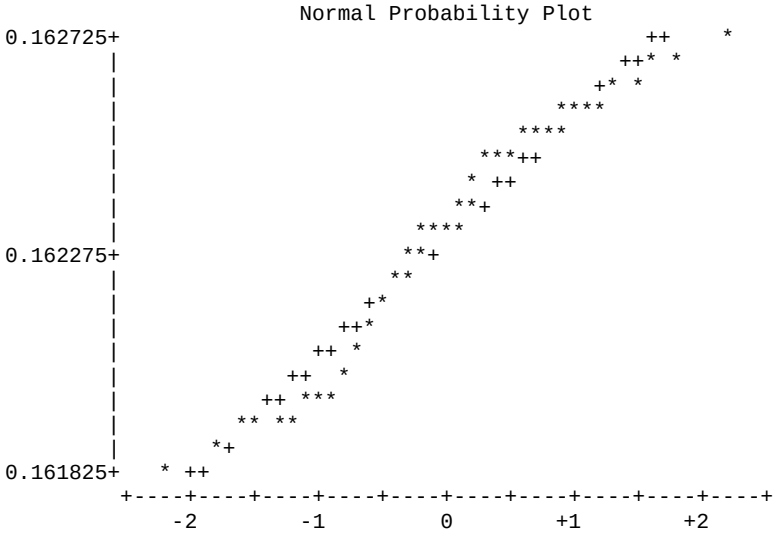
Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
0.161802	2	0.162602	46
0.161886	3	0.162615	47
0.161907	4	0.162670	48
0.161910	1	0.162682	50
0.161915	5	0.162701	49

```
Stem Leaf                     #             Boxplot
1627 0                        1                |
1626 78                       2                |
1626 02                       2                |
1625 6778                     4                |
1625 011134                   6             +-----+
1624 66777                    5             |     |
1624 3                        1             |     |
1623 589                      3             |     |
1623 013333                   6             *--+--*
1622 58                       2             |     |
1622 02                       2             |     |
1621 59                       2             |     |
1621 1                        1             |     |
1620 88                       2             +-----+
1620 02                       2                |
1619 778                      3                |
1619 1122                     4                |
1618 9                        1                |
1618 0                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-4
```

The UNIVARIATE Procedure
Variable: TauY

Normal Probability Plot



The UNIVARIATE Procedure
Variable: TauZ

Moments

N	50	Sum Weights	50
Mean	3.7075934	Sum Observations	185.37967
Std Deviation	0.78433975	Variance	0.61518885
Skewness	-0.0195176	Kurtosis	-1.1976147
Uncorrected SS	717.456695	Corrected SS	30.1442536
Coeff Variation	21.154956	Std Error Mean	0.11092239

Basic Statistical Measures

Location		Variability	
Mean	3.707593	Std Deviation	0.78434
Median	3.714860	Variance	0.61519
Mode	.	Range	2.63731
		Interquartile Range	1.34256

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	33.42511	Pr > |t|	<.0001
Sign	M	25	Pr >= |M|	<.0001
Signed Rank	S	637.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	5.00858
99%	5.00858
95%	4.90812
90%	4.77865
75% Q3	4.38390
50% Median	3.71486
25% Q1	3.04134
10%	2.62880
5%	2.48634
1%	2.37127
0% Min	2.37127

The UNIVARIATE Procedure
Variable: TauZ

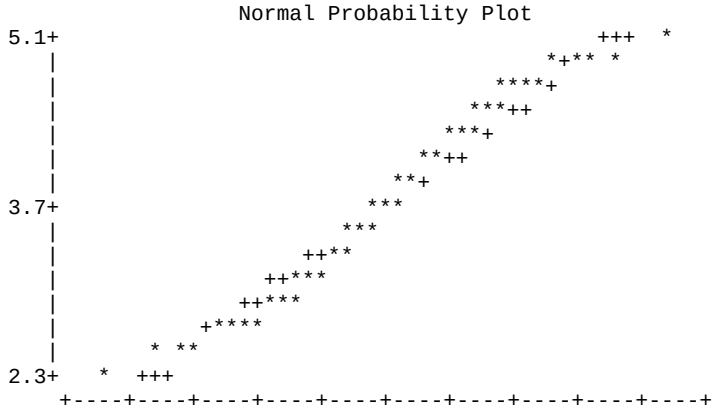
Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
2.37127	1	4.80453	46
2.43050	2	4.85586	47
2.48634	3	4.90812	48
2.54217	4	4.96257	49
2.60102	5	5.00858	50

```
Stem Leaf                     #             Boxplot
  50 1                        1                |
  48 0616                     4                |
  46 0505                     4                |
  44 495                      3                |
  42 2838                     4             +-----+
  40 0617                     4             |     |
  38 405                      3             |     |
  36 3949                     4             *--+--*
  34 2728                     4             |     |
  32 0516                     4             |     |
  30 495                      3             +-----+
  28 2839                     4                |
  26 0616                     4                |
  24 394                      3                |
  22 7                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```

The UNIVARIATE Procedure
Variable: TauZ

Normal Probability Plot



Obs	n TauX	n TauY	n TauZ	mnTauX	mnTauY	mnTauZ	seTauX	seTauY	seTauZ
1	50	50	50	75.0024	0.16231	3.70759	.001259381	.000035409	0.11092

Obs	Overall MSE HBM	Overall MSE Krige	Overall MSE CMAQ	Overall Bias HBM	Overall Bias Krige	Overall Bias CMAQ
1	0.10	0.04	0.06	0.03	0.06	-0.09

Obs	ndays	Pct_Time_ Krige_ worse	krigeworse	Pct_Time_ CMAQ_worse	CMAQworse
1	333	3.60	12	11.71	39

Obs	nsites	Pct_site_ Krige_ worse	krigeworse	Pct_site_ CMAQ_worse	CMAQworse
1	41	9.76	4	31.71	13

Obs	modelCI	krigeCI	predStd
1	0.97616	0.84095	0.4232558534

Section D
Cells.txt

cells_12km_2001.txt - Notepad

File Edit Format View Help

COL #	ROW #	Center Longitude	Center Latitude	(SW Corner) Longitude	(SW Corner) Latitude	(SE Corner) Longitude	(SE Corner) Latitude	(NE Corner) Longitude	(NE Corner) Latitude	(NW Corner) Longitude	(NW Corner) Latitude
1	1	-99.48958	28.47195	-99.54857	28.41712	-99.42727	28.42005	-99.43051	28.52675	-99.55198	28.52382
1	2	-99.49291	28.57867	-99.55198	28.52382	-99.43051	28.52675	-99.43377	28.63349	-99.55539	28.63056
1	3	-99.49625	28.68542	-99.55539	28.63056	-99.43377	28.63349	-99.43703	28.74026	-99.55882	28.73732
1	4	-99.49960	28.79221	-99.55882	28.73732	-99.43703	28.74026	-99.44030	28.84707	-99.56226	28.84413
1	5	-99.50296	28.89903	-99.56226	28.84413	-99.44030	28.84707	-99.44358	28.95391	-99.56570	28.95096
1	6	-99.50632	29.00588	-99.56570	28.95096	-99.44358	28.95391	-99.44687	29.06078	-99.56915	29.05783
1	7	-99.50970	29.11277	-99.56915	29.05783	-99.44687	29.06078	-99.45016	29.16768	-99.57261	29.16473
1	8	-99.51308	29.21969	-99.57261	29.16473	-99.45016	29.16768	-99.45348	29.27462	-99.57608	29.27166
1	9	-99.51648	29.32664	-99.57608	29.27166	-99.45348	29.27462	-99.45679	29.38159	-99.57957	29.37863
1	10	-99.51988	29.43362	-99.57957	29.37863	-99.45679	29.38159	-99.46011	29.48860	-99.58305	29.48563
1	11	-99.52329	29.54064	-99.58305	29.48563	-99.46011	29.48860	-99.46345	29.59563	-99.58656	29.59266
1	12	-99.52672	29.64769	-99.58656	29.59266	-99.46345	29.59563	-99.46679	29.70270	-99.59007	29.69972
1	13	-99.53014	29.75477	-99.59007	29.69972	-99.46679	29.70270	-99.47014	29.80979	-99.59358	29.80681
1	14	-99.53358	29.86188	-99.59358	29.80681	-99.47014	29.80979	-99.47350	29.91692	-99.59711	29.91393
1	15	-99.53703	29.96902	-99.59711	29.91393	-99.47350	29.91692	-99.47688	30.02408	-99.60065	30.02108
1	16	-99.54050	30.07619	-99.60065	30.02108	-99.47688	30.02408	-99.48026	30.13126	-99.60420	30.12827
1	17	-99.54396	30.18339	-99.60420	30.12827	-99.48026	30.13126	-99.48364	30.23848	-99.60776	30.23548
1	18	-99.54744	30.29062	-99.60776	30.23548	-99.48364	30.23848	-99.48704	30.34573	-99.61132	30.34272
1	19	-99.55093	30.39787	-99.61132	30.34272	-99.48704	30.34573	-99.49045	30.45300	-99.61490	30.44999
1	20	-99.55442	30.50516	-99.61490	30.44999	-99.49045	30.45300	-99.49386	30.56031	-99.61848	30.55729
1	21	-99.55793	30.61248	-99.61848	30.55729	-99.49386	30.56031	-99.49728	30.66764	-99.62209	30.66462
1	22	-99.56145	30.71982	-99.62209	30.66462	-99.49728	30.66764	-99.50072	30.77500	-99.62569	30.77197
1	23	-99.56497	30.82720	-99.62569	30.77197	-99.50072	30.77500	-99.50417	30.88239	-99.62931	30.87936

Note that this file is specific to a year and grid size.

Section E:
Summary.CSV file description

The program generates a file named **SUMMARY.CSV** with each fit of the T-SpACE simulation, which contains summary statistics on all parameter estimates in T-SpACE. Table E-1 presents example rows from this file. Each row contains summary statistics on the estimates which were generated for a particular model parameter (specified in the first column) within each iteration of the MCMC simulation. These summary statistics include the minimum ("Min"), maximum ("Max"), average ("Avg"), and variance ("Var") and are calculated on the set of estimates associated with the specified model parameter. Thus, the number of estimates entering into these statistics equals the number of iterations performed within the simulation. Also, estimates generated within the first and last iteration of the simulation are presented in columns labeled "First" and "Last," respectively. The average and variance are of greatest importance within this file, as they represent unbiased estimates of the true mean and variance of the marginal posterior distribution for the given model parameter. The last column (labeled "Success") has a more technical interpretation; it represents the proportion of iterations for which the model parameter successfully moved to a new sample value when the analysis proposed such a move. This proportion is less than 1 whenever the algorithm incorporates a *Metropolis*-type step in which at least one move to a new sample value was rejected.

Table E-1. Example Contents of the SUMMARY.CSV Data File

Name	First	Last	Min	Max	Avg	Var	Success
Z[0]	0.981941	-0.31232	-0.59761	0.981941	0.062989	0.081035	1
Z[1]	0.685994	-0.45824	-0.47367	0.685994	0.057133	0.070435	1
…							
Z[2669]	-0.08756	-0.2092	-0.78271	0.621141	-0.06417	0.081926	1
V[0]	0.449908	-0.29966	-1.73886	1.54585	0.00787	0.267134	1
V[1]	1.29497	-0.23137	-2.1078	1.29497	0.003764	0.302053	1
…							
V[2669]	-0.5677	-0.26227	-1.11112	1.14958	0.013141	0.23977	1
Mu	1.54973	2.22671	1.54973	2.72309	2.36964	0.039279	1
Eta	-21.4706	20.2539	-73.8361	76.3331	1.04448	1083.5	1
BetaC[0]	14.6051	7.58928	-72.5916	72.2254	-0.58191	763.101	1
BetaC[1]	43.5833	27.0894	-103.352	94.3341	-0.3937	902.584	1
…							
BetaC[11]	16.7617	-35.1617	-71.7791	83.4948	2.95884	1037.14	1
BetaD[0]	-0.63135	-0.36124	-1.40918	0.204948	-0.75164	0.106207	1
BetaD[1]	-0.65789	-0.34519	-1.07664	0.06656	-0.53966	0.054132	1
…							
BetaD[27]	0.000295	0.000613	-0.00033	0.000824	0.000334	5.73E-08	1
BetaV	2.26E-05	1.54E-05	-7.66E-05	7.05E-05	-2.13E-06	1.04E-09	1
TauX	22.5191	181.391	3.00131	1380.1	142.902	44200.4	1
TauY	1.00057	1.00137	0.998501	1.00475	1.001	1.23E-06	1
TauS	1.6368	0.505814	0.007201	8.80531	1.08206	1.47834	1
TauZ	2.24921	4.60805	2.24921	5.72256	4.622	0.336919	1
TauV	1.40102	1.06579	0.697571	2.01309	1.15445	0.074964	1
RhoZ	0.448629	0.017082	0.000195	0.448629	0.040497	0.00768	0.607
RhoV	0.47431	0.014791	0.000597	0.47431	0.048789	0.008882	0.727

W[0]	2.62485	1.76478	1.46325	2.88756	2.17617	0.084941	1
W[1]	2.32892	1.61885	1.61885	3.00553	2.17031	0.072253	1
…							
W[2669]	1.55533	1.86789	1.34492	2.78993	2.04901	0.08942	1
Bias[0]	-0.53088	-0.53437	-1.12979	-0.32377	-0.76766	0.03328	1
Bias[1]	-0.50349	-0.5204	-1.06923	-0.30484	-0.73469	0.029905	1
…							
Bias[2669]	1.12976	1.1581	0.650178	1.57056	1.01738	0.030676	1
Covar[0]	-21.0207	19.9543	-73.8786	76.3927	1.05235	1082.49	1
Covar[1]	-20.1756	20.0226	-73.663	76.4343	1.04824	1084.8	1
…							
Covar[2669]	-22.0383	19.9917	-73.9745	76.2039	1.05762	1086.94	1
WeeklyAvg[0]	2.03077	2.00202	1.81888	2.5076	2.12928	0.023503	1
WeeklyAvg[1]	1.85091	1.8708	1.69304	2.53312	2.13047	0.026893	1
…							
WeeklyAvg[533]	1.59642	1.96121	1.59642	2.45346	2.09582	0.021795	1
MonthlyAvg[0]	2.03077	2.00202	1.81888	2.5076	2.12928	0.023503	1
MonthlyAvg[1]	1.85091	1.8708	1.69304	2.53312	2.13047	0.026893	1
…							
MonthlyAvg[533]	1.59642	1.96121	1.59642	2.45346	2.09582	0.021795	1
YearlyAvg[0]	2.03077	2.00202	1.81888	2.5076	2.12928	0.023503	1
YearlyAvg[1]	1.85091	1.8708	1.69304	2.53312	2.13047	0.026893	1
…							
YearlyAvg[533]	1.59642	1.96121	1.59642	2.45346	2.09582	0.021795	1

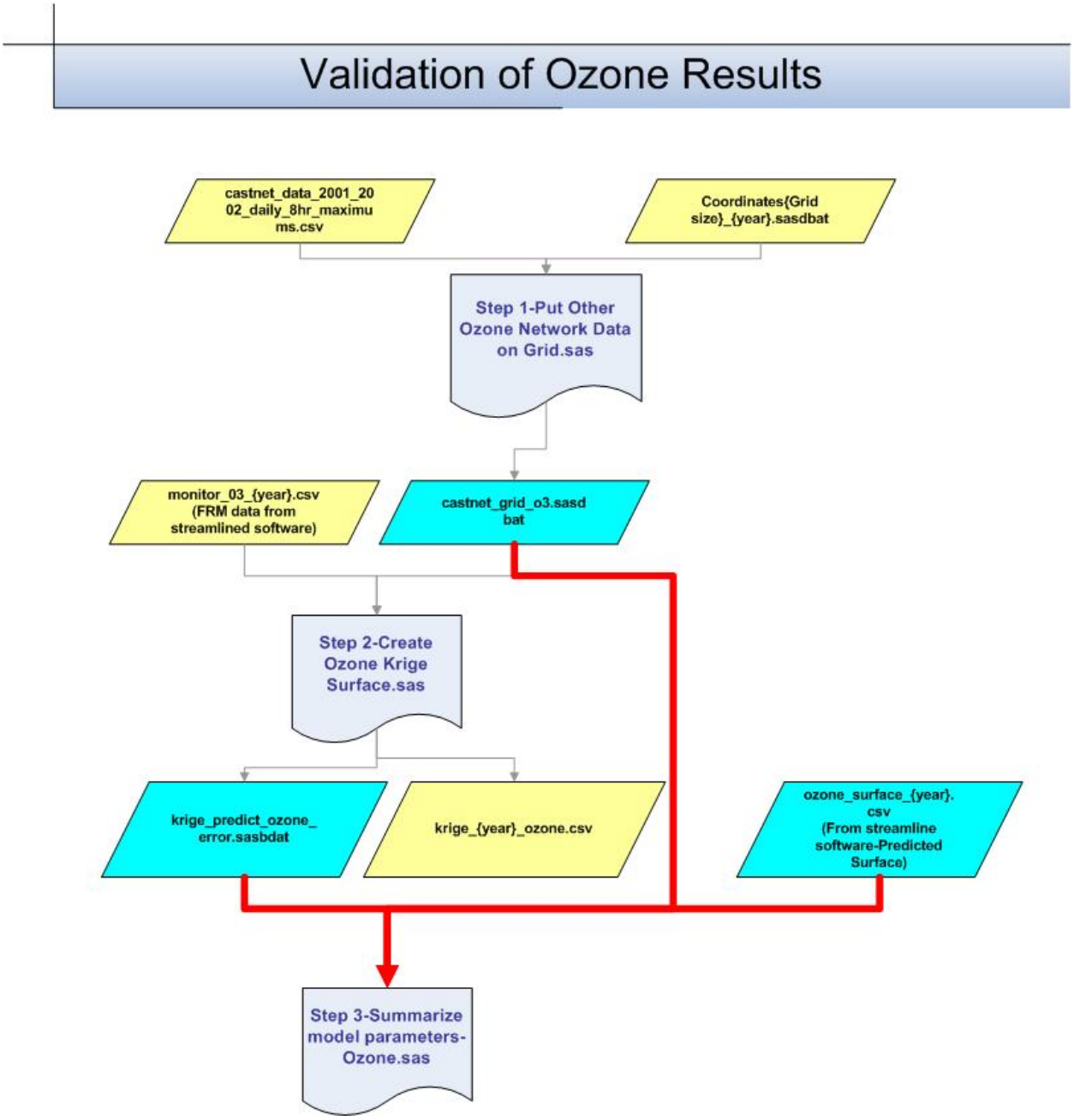
Section F:
4th Highest Concentration Surface

Within the time period of interest (such as a year), the T-SpACE software determines the fourth highest concentration within each grid cell across all simulated air pollution concentration surfaces that T-SpACE generates for each day in the time period. The results are averaged and smoothed to generate a single air pollution concentration response surface (daily maximum 8-hour average ozone O_3 concentration). T-SpACE generates a ***.CSV** file, which contains the cell-specific predictions that make up this response surface. However, only the following four columns, which document the grids and the predictions, are included in this CSV file:

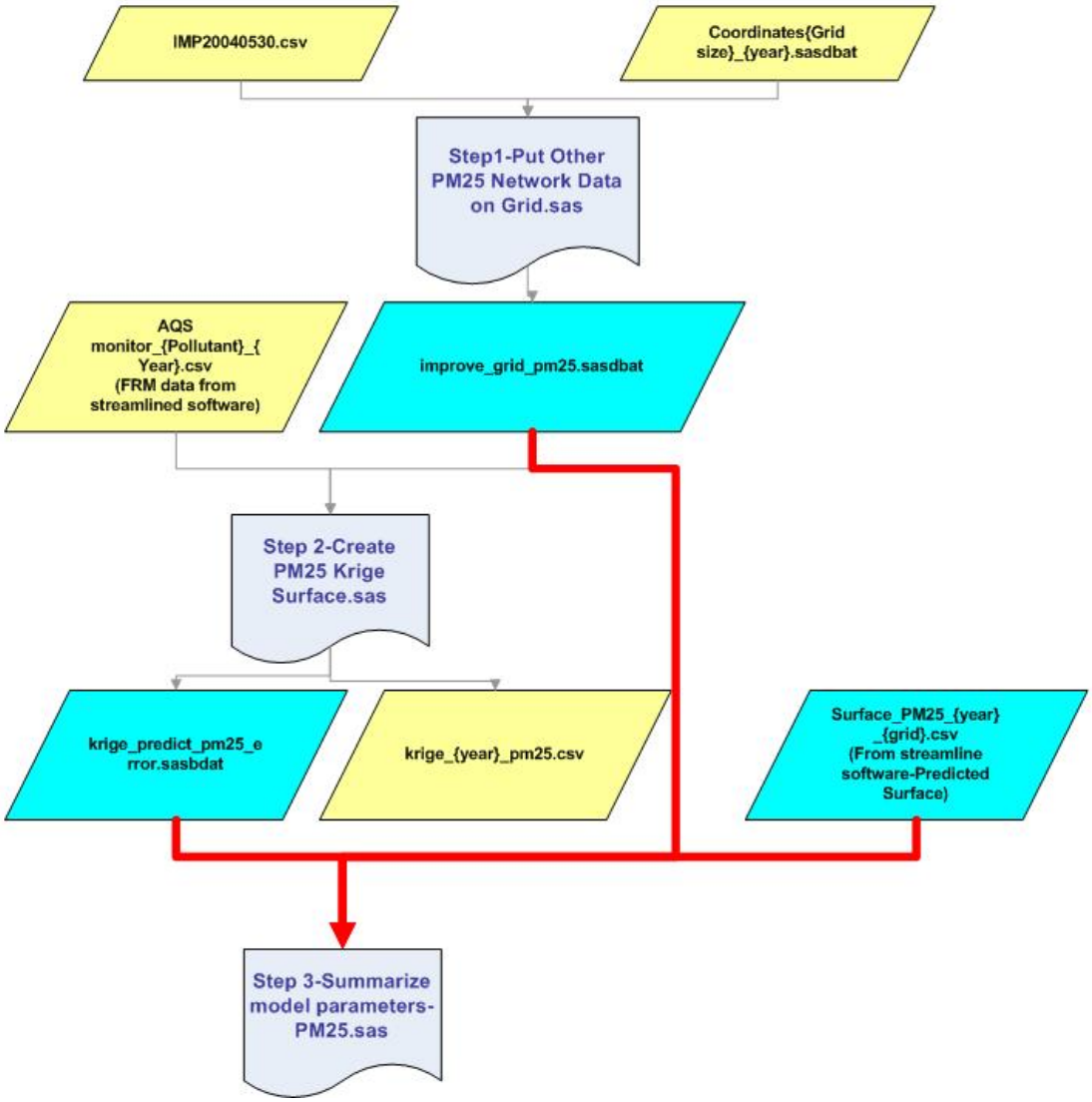
- **XCoord**: The cell's x-coordinate (east-west) within the grid (where the coordinate value increases by 1 with each grid cell as you move from west to east).
- **YCoord**: The cell's y-coordinate (north-south) within the grid (where the coordinate value increases by 1 with each grid cell as you move from south to north).
- **PredAvg**: The (natural log-transformed) T-SpACE-predicted value of the fourth highest concentration for the given grid cell.
- **PredStd**: The (natural log-transformed) T-SpACE-predicted value of the standard error for the fourth highest concentration for the given grid cell.

In particular, because the air pollution concentration response surface for the fourth highest concentration represents the entire time period of interest (e.g., one year), a date column is not necessary in this ***.CSV** file.

Section G:
Diagram of Preparation of Validation Files



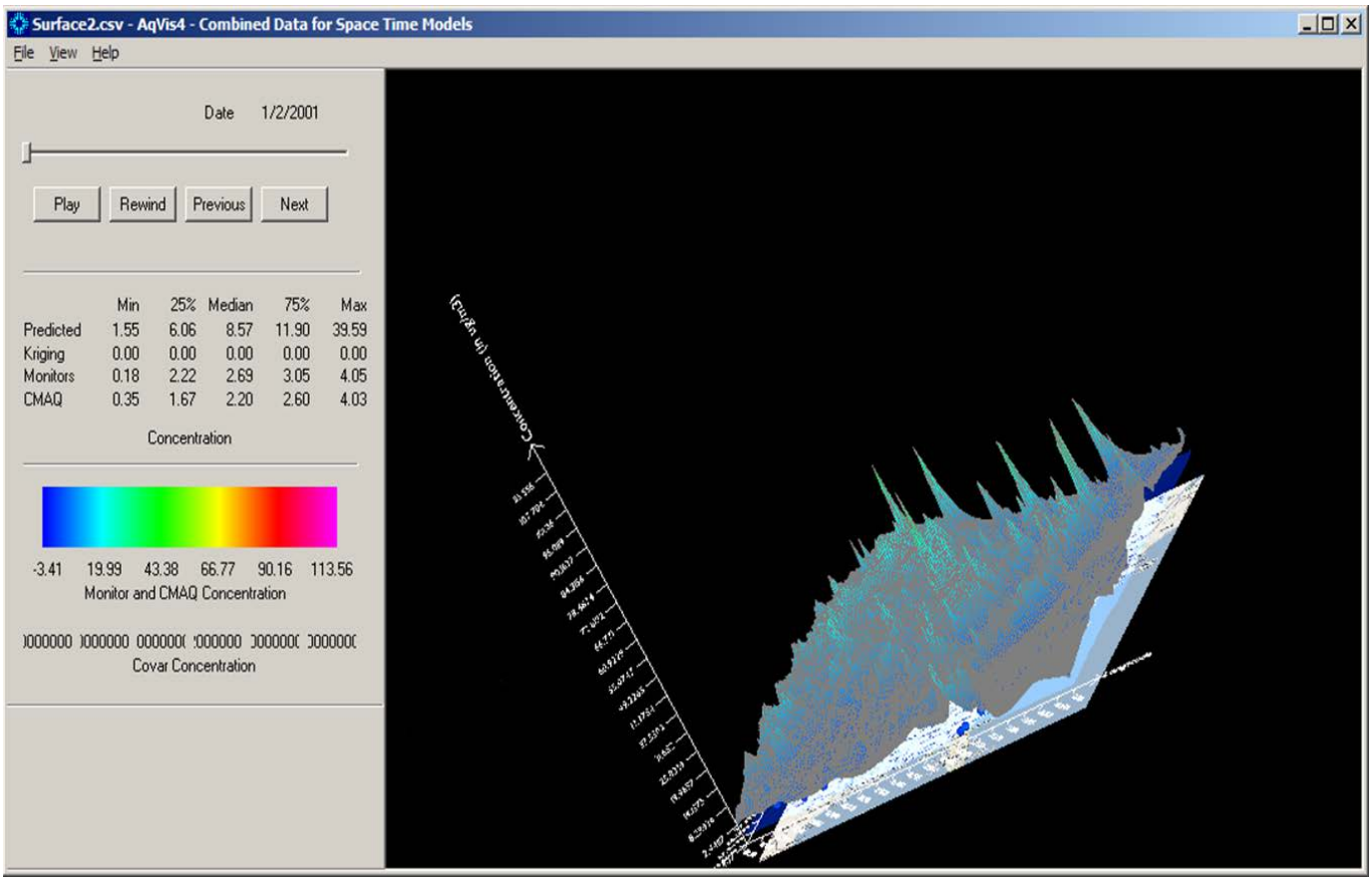
Validation of $PM_{2.5}$ Results



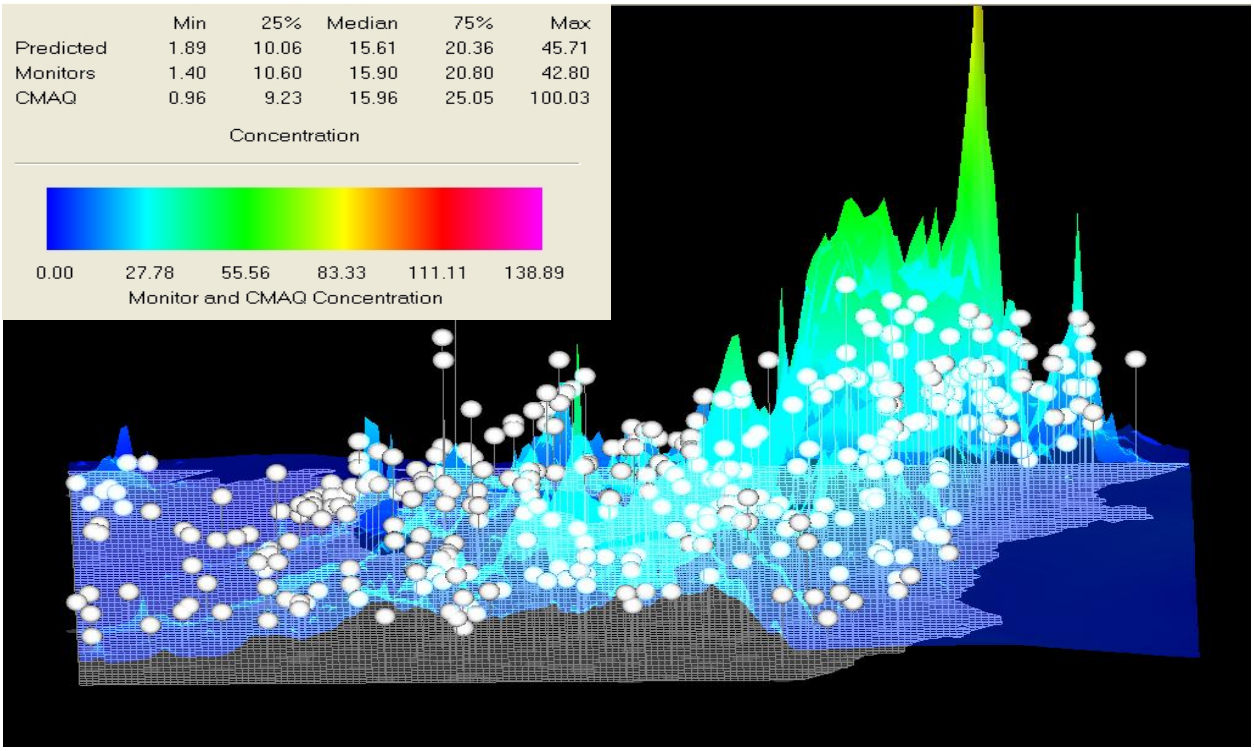
Section H: Visualization of T-SpACE Three-Dimensional Coordinate System

Step 6: Visualize Surface

This step allows the user to visualize the estimated **T-SpACE** air pollution concentration surface in a three-dimensional (3-D) coordinate system (latitude, longitude, concentration).



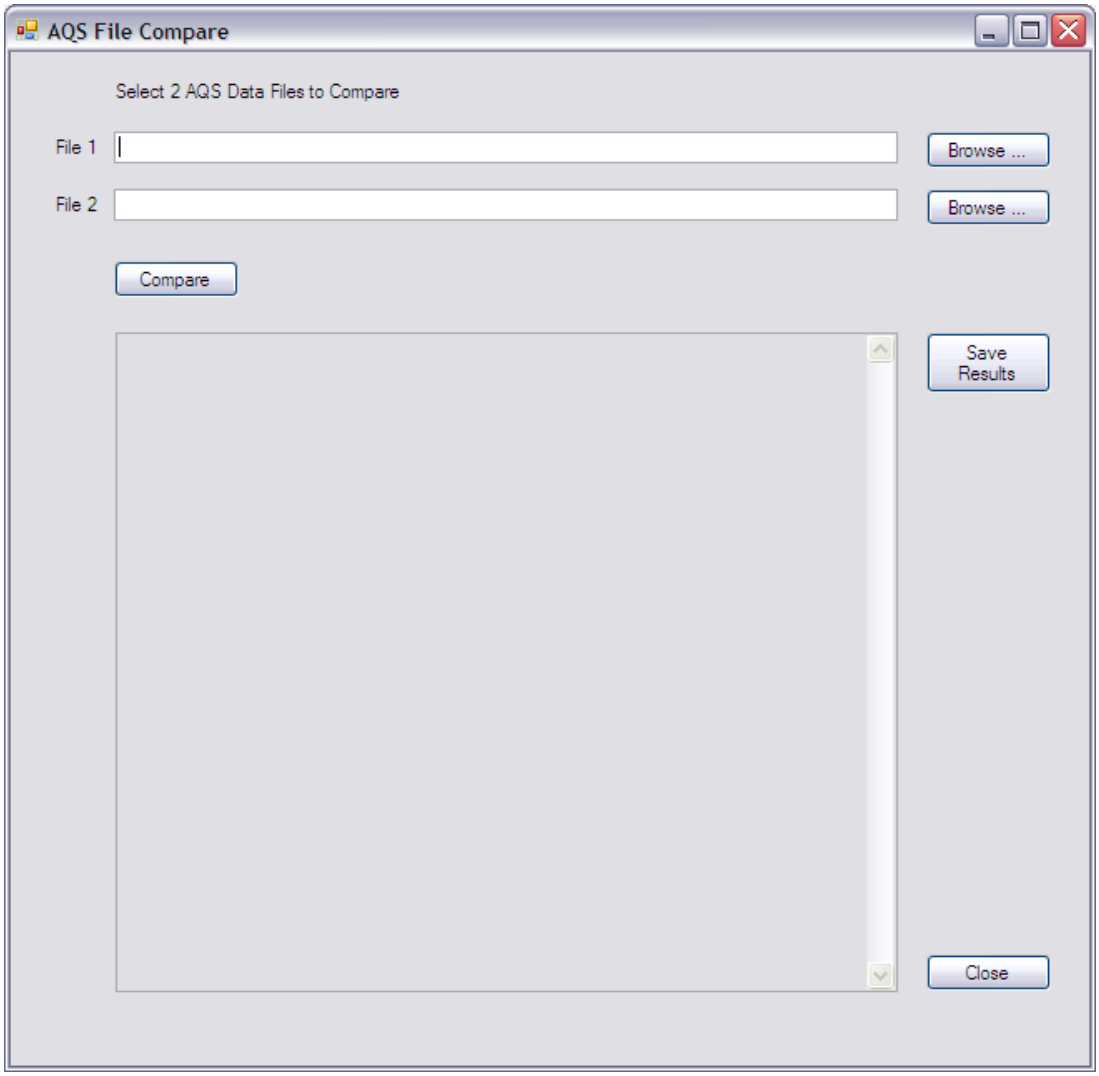
This **T-SpACE** graphic displays $PM_{2.5}$ monitor concentration values (white spheres) and CMAQ air quality model estimates (colored surfaces) for the north eastern portion of the US.



Section I: T-SpACE Utilities

T-SpACE Utilities
The seven T-SpACE utilities that are provided to assist the user in generating air pollution concentration surfaces are provided here in summary format.

Tool #1: Compare AQS Raw Data files: This tool determines which monitor locations are contained in one AQS file but not in the other AQS file.



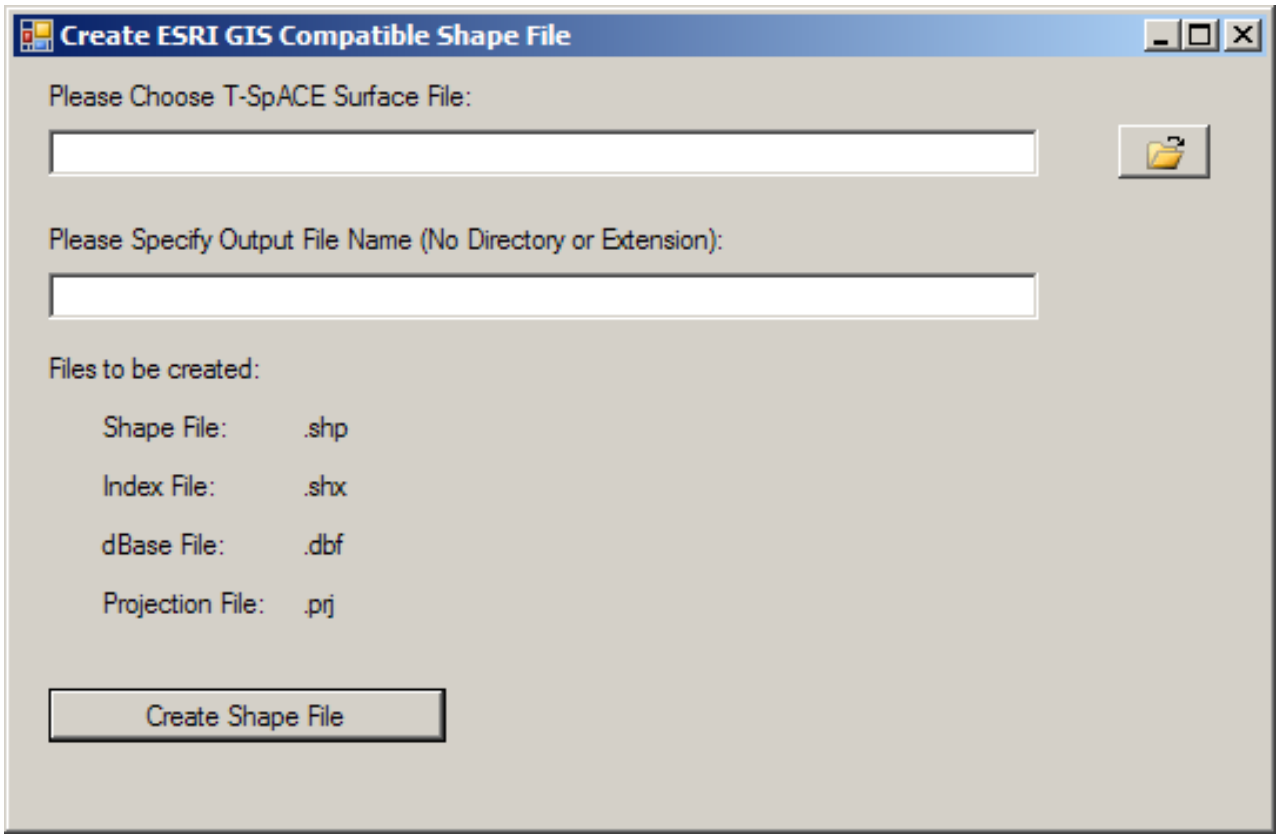
Tool #2: Compare Airsites TSV (Tab-Separated Values) files: This tool determines which monitor locations are contained in one TSV file but not in the other TSV file.

airsites.tsv - Notepad

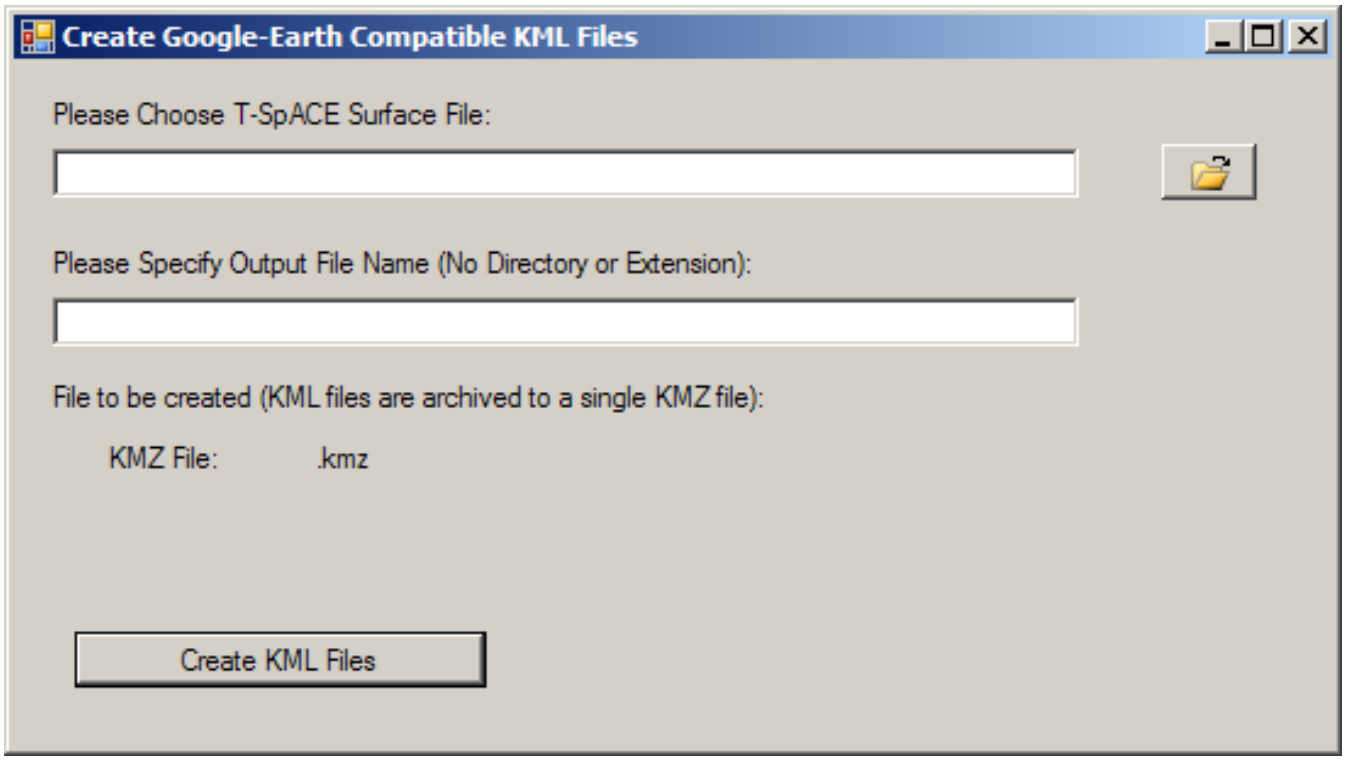
File Edit Format View Help

010010001	-86.472778	+32.437222	6
010010002	-86.443611	+32.428333	6
010010003	-86.791667	+32.332222	6
010030002	-87.706944	+30.552222	6
010030003	-87.713611	+30.555278	6
010030010	-87.881389	+30.497778	6
010050001	-85.135278	+31.891111	6
010050002	-85.606234	+31.664137	6
010070001	-87.233056	+32.903889	6
010090001	-86.464444	+33.956944	6
010090002	-86.473056	+33.949444	6
010150001	-85.824444	+33.656667	6
010150002	-85.837778	+33.685556	6
010210001	-86.637778	+32.839444	6
010210002	-86.628889	+32.839444	6
010230001	-88.051944	+32.209167	6
010230002	-88.053611	+32.205556	6
010250001	-87.781944	+31.966389	6
010250002	-87.773333	+31.694167	6
010270001	-85.802222	+33.281111	6
010310002	-87.259167	+31.056667	6
010330001	-87.886667	+34.736944	6
010330007	-87.665000	+34.760000	6
010330008	-87.643056	+34.760833	6

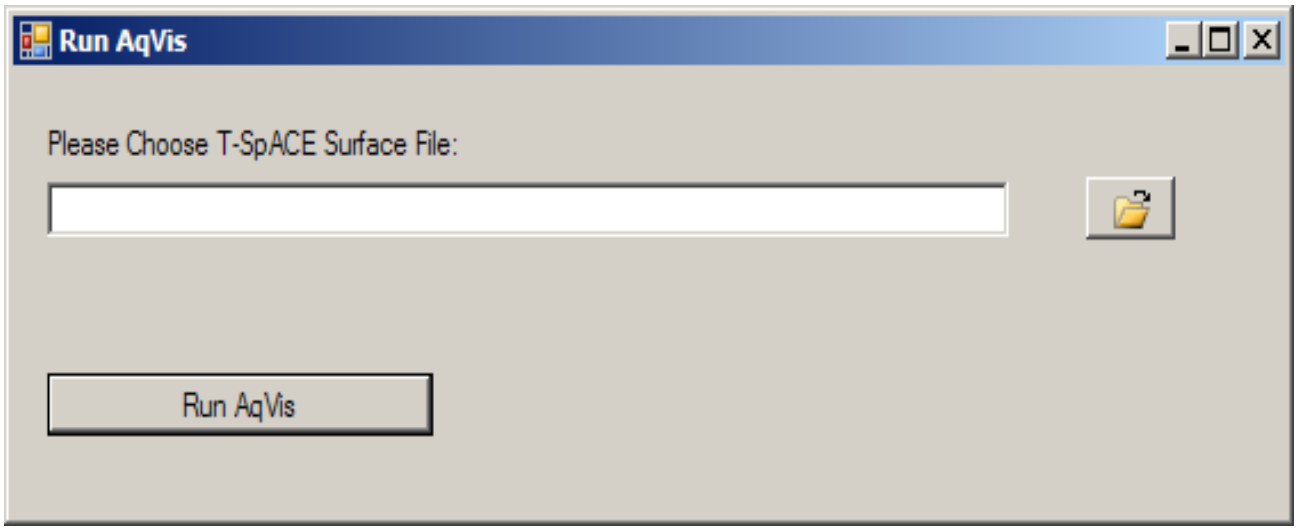
Tool #3: Create ESRI GIS Shape File: This tool converts T-SpACE model-generated surface files to the ESRI GIS format for input into other software.



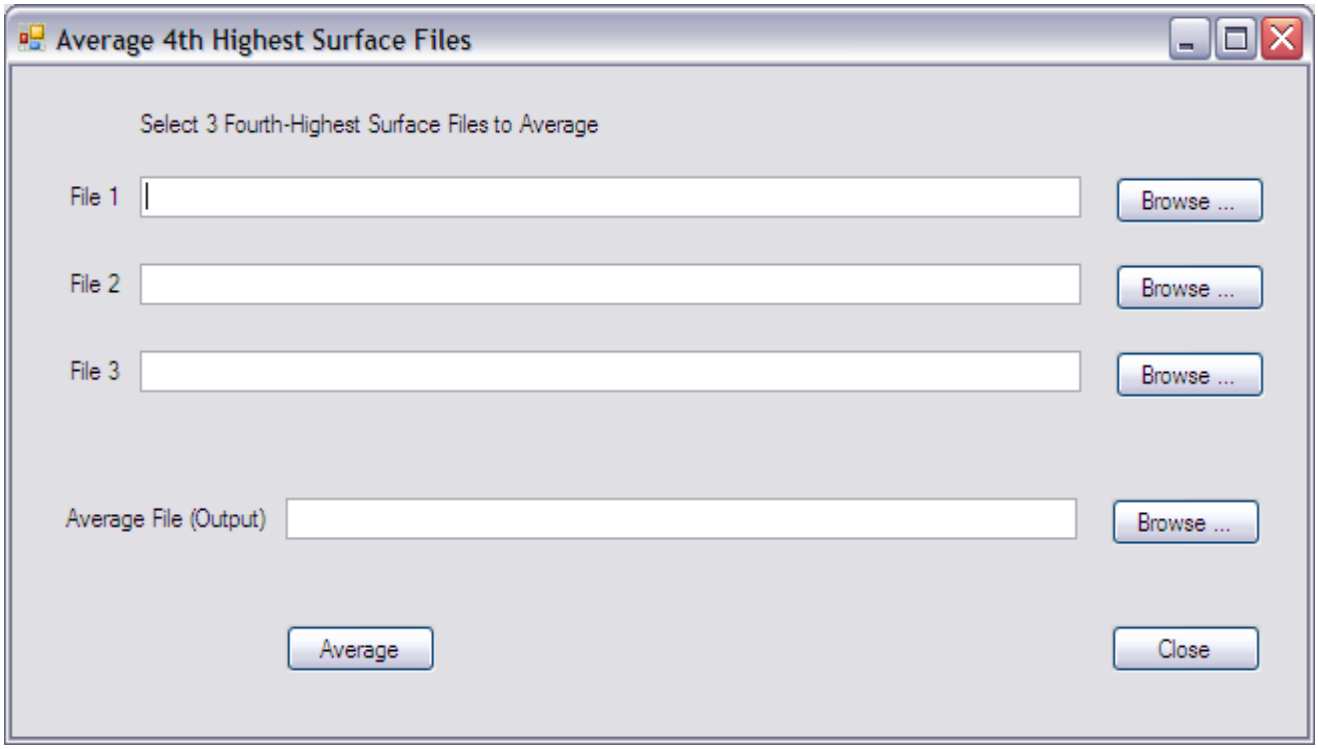
Tool #4: Create Google Earth File: This tool converts T-SpACE model-generated surface files to the proper format for visualization in Google Earth.



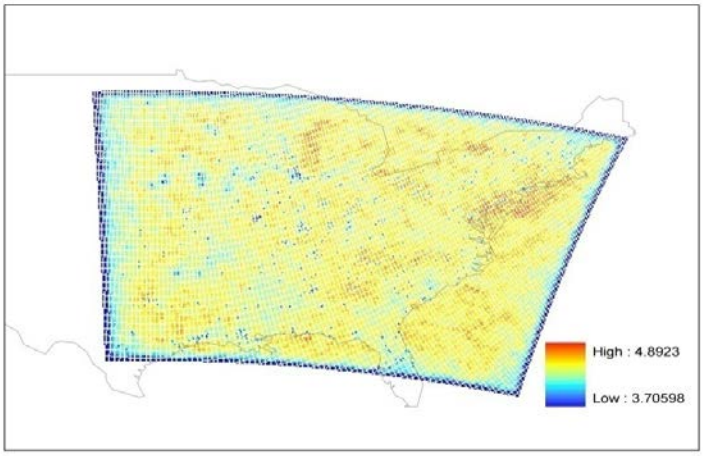
Tool #5: Visualize Surface File: This tool launches the **AQVis** application, which is the same utility that is used in T-SpACE Model Step 6.



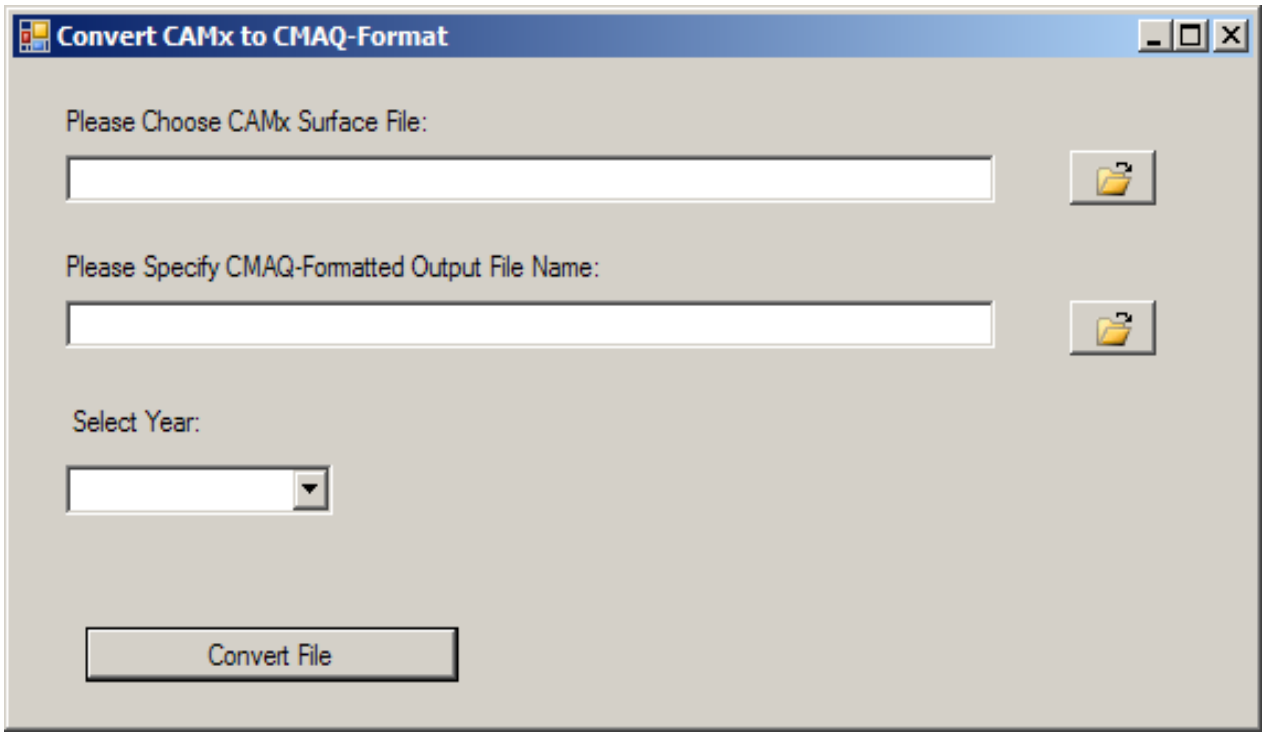
Tool #6: Average 4th Highest Surface Files: This tool averages three consecutive annual 4th highest daily maximum 8-hour average ozone (O_3) concentration surface files into a single (separate) 4th highest surface file.



CMAQ Example of 4th highest surface - **<u>Note</u>:** This graph is not provided in **T-SpACE**.



Tool #7: Convert CAMx File to CMAQ Format: This tool converts a 'raw' CAMx surface file to the CMAQ (NetCDF) format, allowing it to be used in **T-SpACE**.



SCIENCE