Appendix I

Model-to-Monitor Comparison Methods

Appendix I

Protocol for Model-to-Monitor Comparisons for National Air Toxics Assessment

Office of Air Quality Planning and Standards

Revised Draft
February 29, 2000

In the interest of simplicity and readability, this document is written using Aplain language@. The use of Awe@ refers to the US EPA.

I. Introduction.

The purpose of this document is to describe a project we=re undertaking at the EPA which involves comparing ambient air quality modeling system estimates to ambient monitoring concentrations for hazardous air pollutants (HAPs), or air toxics. Namely, we would like to have a better understanding of the performance of the ASPEN model in predicting ambient toxics concentrations for our initial national-scale assessment.

As one of the tools that will be used in the National Air Toxic Assessments (NATA) initial national-scale assessment, the ASPEN model is a dispersion model which can estimate the annual average ambient concentration of a HAP at the census tract level. So, for example, we can ask ASPEN, AWhat is the annual average concentration for ambient benzene in 1996 at census tract X ?@ ASPEN will then give us a quantitative estimate of the annual average concentration at that given location.

Dispersion models in general have many uses. First, an accurate model will reduce the need for a dense monitoring network, thus saving monitoring costs. Second, a model can answer forward-looking questions such as, AIf we reduce emissions of HAP *X* by *y*%, how will that affect ambient concentrations in the future?@ Monitors cannot answer this question directly. Third, the model can estimate the ambient concentrations of pollutants at places where there are no monitors or a monitor is not feasible. In lieu of the model, we=d have to use Aspatial interpolation@ methods such as kriging to estimate ambient concentrations where there are no monitors.

In fact, EPA considers the predictions of dispersion models such an important component of their risk-based air toxic program that model-to-monitor comparison considerations are one of the key considerations being included in the design of a National Air Toxics Monitoring Network (US EPA, 2000a). We=d like to design our network so that we can evaluate dispersion models in a wide variety of situations - for urban areas and nonurban areas, areas dominated by certain emissions sources, areas with different climates, etc. Table 1 is a preliminary attempt at delineating some of these situations.

One way to evaluate the usefulness and limitations of any dispersion modeling system is to compare its ambient concentration predictions to available concurrent monitoring data. As part of the NATA initial national-scale assessment, the ASPEN model will predict annual average ambient concentrations for 33 urban HAPs at approximately 60,000 census tract locations nationwide for the year 1996. In 1996 there were only several hundred air toxics monitoring sites all across the US. Many of these sites, which were primarily designed and maintained under existing criteria air pollutant monitoring programs (e.g., Photochemical Assessment Monitoring Stations, or PAMS; State and Local Air Monitoring Stations, or

SLAMS; Interagency Monitoring of Protected Visual Environments, or IMPROVE), and only monitored a handful of HAPs for a limited period of time. Where data allow us, we can look at the 1996 annual averages from these sites for whichever HAPs they monitor, and compare these annual averages to the 1996 annual averages generated by ASPEN for the appropriate geographic location. By comparing the ASPEN predictions with the available monitoring data, we hope to gain a better understanding of the overall performance and limitations of the quantitative ASPEN model predictions. It is these ASPEN model predictions that will be used in the initial national-scale assessment to predict exposure and risk values nationwide. These predicted exposure and risk levels will subsequently help the Agency in setting priorities for future control efforts of our air toxic programs.

II. Uncertainties and Limitations.

To better understand what the results of such a comparison may tell us and make the most use of its results, we must first realize the limitations that both the NATA initial national-scale model predictions and monitoring measurements have. These are discussed below.

A. ASPEN Modeling Uncertainties.

i) Emissions Inventory. The model takes emissions data and meteorological data as inputs. For the initial national-scale assessment the emissions data come from the 1996 National Toxics Inventory (NTI), a composite of emissions estimates generated by state and local regulatory agencies, industry, and EPA. Because the estimates in the NTI originated from a variety of sources and estimation methods, as well as being developed for a variety of different purposes, they will vary in quality (i.e., pollutants, level of site detail, and geographic coverage). In some cases, the inventory may serve as an excellent site specific representation for a dispersion model effort, in other cases key model input parameters may not have been available. Further, for area and mobile emissions in particular, representative surrogates (e.g., industrial land, roadway miles, population density and inverse population density) were used to geographically distribute county wide inventories to a small geographic scale (i.e, census tract). Thus, the usefulness of these data to represent smaller spatial scales may be in question.

ii) The Modeling Simulation. A dispersion model in general makes many simplified assumptions as to the fate and transport of the emission plume. One of the key simplifications of the ASPEN model is that it does not include a terrain component in its prediction algorithms. Further, the model relies on steady-state long-term sector-averaged climate summary data to represent the conditions at the plume site. The model also simplifies some complex atmospheric chemical processes and captures only pollution transport within 50 km of any individual source.

B. Monitoring Uncertainties.

Though often overlooked, there can be significant uncertainties as to what the monitored value actually represents. We must consider uncertainties in both monitoring methods as well give considerations as to what the monitor was initially sited for before we can consider a comparison with model predictions. Unlike the criteria air pollutant world, there currently is not a formal national air toxics monitoring network which follows standardized EPA guidelines or established national monitoring procedures. While some States and local agencies have collected some high quality HAP monitoring data, some of the data has not undergone any formal quality assurance tests, and the data come from several different monitoring networks, which may differ in precision and accuracy. The number of monitoring sites varies by pollutant. The monitor siting (proximity of sites to emission sources) also varies by pollutant, with some monitors reflecting general urban concentrations while others are more reflective of specific source impacts. In general, most the of available HAP data to be considered in this evaluation were not collected every day. Instead, they were produced every 12th day or every 6th day throughout the calendar year. This introduces another layer of uncertainty as to the representativeness of the data for comparison with estimated annual average concentrations. Thus, we are proposing a set of completeness criteria be applied to the monitoring data to help assure its representativeness of 1996 annual averages (see Appendix A).

C. Comparison Uncertainties.

In addition to the Aindependent@ model and monitoring uncertainties mentioned above, there are a few reasons why the comparison itself is uncertain. The monitors have a level below which measurements are uncertain (called the Amethod detection limit@, or AMDL@), because routine monitoring methods are not sensitive enough to detect low levels precisely. The ASPEN model does not have a MDL, and thus careful consideration as to the treatment of the MDL must be made when aggregating short-term monitoring concentrations into annual average values. Another consideration in the comparison is the question of resolution, or scale. ASPEN was designed to give estimates at the census tract level. In contrast, the monitors are sensitive to very local fluctuations in air pollution, especially close to dominant emissions sources. If we had a series of monitors lined up right next to each other, for certain pollutants and in certain situations, the readings could vary quite a bit. But we would expect the model to generate very similar values for nearby locations. The model by necessity has to Asmooth out@ the local fluctuations in pollutant levels because it=s predicting impacts on a much larger scale. Still, since we=re dealing with annual averages here and not short-term values, we would expect that these fluctuations in spacial monitoring data would in all likelihood be smoothed out spatially over a 1-year period.

It is important that even if the comparison does not result in a Aperfect@ comparison for all pollutants and monitoring locations, trends in the comparison can be used to better understand the strengths and weaknesses and uncertainty bounds on the predicted model results. A comparison can help us both improve the model and give us a better idea what its predictions are useful for. By discovering where the model is performing well and where it isn=t, we can refine our understanding of the processes which

underlie air pollution and its transport. The model might perform well for certain pollutants but not others. It might perform better in certain geographic areas than in others. It might do a better job in urban areas than it does in rural areas; it might always overestimate or underestimate; it might do better when concentrations are higher; it might do better for areas near large sources of emissions. Any information of this type will help modelers improve their future assessments. If only an order-of-magnitude accuracy is required to make a certain decision from the model results, then one may be happy with the model=s expected performance based on the comparison. On the other hand, if detailed accuracy is required for a given pollutant and geographical area, the model may be deemed inappropriate for this use or further comparisons may be warranted.

III. The Basic Components of the Comparison.

The model evaluation study will present graphs, tables, maps, and charts which show the results of the model-to-monitor comparison for the year 1996. We will initially do this comparison for nine HAPs: benzene, 1,3-butadiene, formaldehyde, acetaldehyde, acrolein, tetrachloroethylene, cadmium, chromium, and lead. These were chosen because 1) they are a subset of the Aurban HAPs@ considered in the initial national-scale assessment; 2) they represent a range of physical HAP parameters (i.e., organic, volatile, particulate) and 3) there is a significant amount of Ahigher-quality@ monitoring data available for them.

A. Processing of Monitoring Data and Elimination of Questionable Sites.

Appendix A describes the method by which annual means were calculated from hourly (or daily) toxics monitoring data. We will discard a (pollutant, site) pair from the comparison if it falls into any of the following categories, discussed in more detail below:

- It fails the completeness test in Appendix A.
- A large portion or all of the measured data is below the MDL.
- We have reason to doubt the accuracy of the monitoring average.
- The site is very close to an international border.
- There is some uncertainty as to the appropriate model emissions parameters in the vicinity of the monitoring site which affect the model estimate for the given pollutant.

i) Completeness test. None of the monitors in this study measure concentrations every day, so we want to be confident that the days measured are representative of all the days of the year. The completeness test is designed to discard (pollutant, site) pairs which may not be representative of atmospheric fate and transport processes as well as expected emissions throughout the year. However, we were careful to avoid making the completeness test too stringent. A more stringent test would eliminate so many sites that we would have too few sites left to conduct a meaningful comparison.

ii) Data below the MDL. In general, daily values below the MDL are replaced with

MDL/2. However, if most or all of the daily observations are below the MDL, we can=t be confident that the computed annual average accurately reflects concentrations in the air. All (pollutant, site) pairs for which all daily values were below the MDL were discarded from the comparison. Other sites with only one or two values above the MDL for the whole year were also discarded.

iii) Questionable monitoring averages. We eliminated sites for which we had reason to doubt the accuracy of the monitoring annual average. In areas with few major point sources, we would expect both toluene and benzene (both HAPs emitted primarily by mobile sources) to be highly correlated in the ambient air. Thus, if a site had an odd ratio of toluene to benzene, we discarded the site from the site list for benzene. The same is true of formaldehyde and acetaldehyde; sites with odd ratios of formaldehyde to acetaldehyde were eliminated from the site list for both pollutants. Sites which had extremely high monitoring values compared to other values for that pollutant were also discarded.

iv) Border sites. Some of the monitoring sites are very close to the US - Mexico border or the US - Canada border. Because we have no emissions data for Mexico or Canada, we do not have confidence in the model estimates for these sites. This is especially true of sites in Calexico, CA; El Paso, TX; Brownsville, TX; and Bellingham, WA. These sites are close to large cities on the other side of the US border.

v) Inventory check. As noted above, a fringe benefit of the model-to-monitor comparison is that it can serve as a check system for the emissions inventory. We=ll see some model estimates which are very far away from the monitoring average. Some of these can be explained by a mislocated emissions source or missing emissions source information. In these situations appropriate modifications can be made to the emissions inventory which result in revised model predictions or we can eliminate these comparisons from our analysis.

B. Figures for Each Pollutant.

For each pollutant, we will have a table showing the raw data, along with four figures:

- a scatter plot of model estimates and monitoring values;
- side-by-side box plots comparing the overall distributions of model estimates and monitoring values;
- a Aprobability plot@ estimating the probability that the model agrees with the monitoring data for a site pair, with regard to the direction of the inequality; and
- a map showing the locations of monitoring sites.

These are discussed below.

i) Scatter Plots. An example is Figure 1. The scatter plot is a straightforward way to show the agreement between the model and the monitors. Each ordered pair on the graph is (monitor, model), for each monitoring site for that pollutant. For example, let=s say we have a benzene monitor at

latitude *X*, longitude *Y*. It gives an annual average for 1996 of 10 ìg/m^3. We give the lat/long coordinates *(X,Y)* to the ASPEN model, and it gives us an annual average of 8 ìg/m^3 at these coordinates. So we plot the point *(10,8)* on the graph.

As noted above, if most of the points are around the *Y=X* line on the graph, which is also on the graph in Figure 1, the model is performing well - its estimates are close to the monitored concentrations.

ii) Side-by-side Box Plots. An example is Figure 2. This is another relatively simple way to show the agreement between the model and the monitors. Let=s say there are 120 monitors across the US for benzene. To construct the box plot for the monitored annual averages, we compute certain percentiles and the mean of these 120 numbers. Then, we input the 120 (lat,long) coordinates of the monitors into the ASPEN model, and get 120 model estimates. We compute the same percentiles and the mean of these numbers, and make another box plot.

If there is good agreement between the model and the monitors, the two box plots will look like they came from the same distribution - they will be side by side, instead of one being higher or longer than the other. The corresponding percentiles and the means should match up fairly closely.

iii) Probability Plots. An example of this type of plot is shown in Figure 3. The idea behind this graph is that it might be unreasonable to expect the model to match up its estimates with the monitors on an absolute scale; but maybe the model is performing well in a relative way. We would hope that if the cadmium monitors say that Site A has a higher concentration than Site B, then the model agrees. So this plot assesses whether the model is getting the Adirection of the inequality@ correct. If the monitors say that Site A=s concentration is only slightly higher than Site B=s, then maybe we can Aexcuse@ the model for getting the direction of the inequality wrong; but if Site A=s is much higher than Site B=s according to the monitors, we would hope that the model does better. The plot looks at the Arelative@ performance of the model for different ratios of Site A=s monitored concentration to Site B=s monitored concentration (assuming A=s is higher than B=s). If the model is performing well, then the probability it agrees with the monitors for site pairs will increase as the ratio gets higher, and will be well over 50% for ratios near 1.3 and 1.4. If the model is performing poorly, its probability of agreement will hover around 50%, which is what you=d get by pure chance, even as the ratio gets well above 1. The plot uses a statistical technique called logistic regression (Agresti, 1990) to estimate the probability of agreement.

iv) Maps. An example is Figure 4. The map will do nothing more than show the locations of the sites on a national scale. It will give some idea of the geographic coverage of the monitors and of the number of sites.

C. Figure for All Pollutants at Once.

In addition to these 9H4=36 figures, we will also present a graph showing Aratio box plots@ for all nine pollutants, on the same set of axes. An example is Figure 5 (which only shows three pollutants for

now). The box plots will show the distribution of monitor/model ratios. So if we have 150 monitors for acrolein, we will have 150 monitor/model ratios to compute. We then compute percentiles and the mean of these 150 numbers, and create a box plot. If the model is performing well, the box plots will be short, and centered at 1. By putting the box plots side by side for each pollutant, we can easily compare the model=s performances for the different pollutants. We=ll be able to surmise which HAPs are being overestimated and underestimated, and which are being estimated consistently and inconsistently.

D. Stratification Tables.

The last group of presentations will show the results of Astratifying@ the sites based on several variables. By Astratification@, we mean that we will place the monitoring sites into certain categories, and evaluate the model=s performance in each of these categories separately. The idea is that we=d like to get a better idea of the model=s performance in specific situations, as described in the introduction above. So far, every figure we=ve looked at aggregates all the sites for each pollutant across the country. We will use three variables to stratify the comparison, for each pollutant:

- urban vs. nonurban;
- geographic/climatological region; and
- pollutant level.

i) Urban vs. Nonurban. We will look at the urban and nonurban sites separately. For the sake of this analysis we will defineAurban@ as a monitoring site being located in a Metropolitan Statistical Area.

ii) Geographic/Climatological Region. We will divide the US into climatological regions using a subset of the Köppen Climate Classification (Godfrey, 1999). Where appropriate, or data limits, we=ll try to merge the classifications into fewer super-classifications.

iii) Pollutant Level. For each pollutant, we will divide the sites into four quartiles, based on their monitored annual averages. We=d like to see if the model performs better for high levels of pollutants. One reason the model might perform better at high levels is that there are fewer MDL issues at high levels.

Instead of presenting the results for these stratifying variables in graphs and box plots, we=ll use some summary statistics. We will present three tables, one for each of the above headings. Table 2 is an example for the APollutant Level@ heading. In each cell of the table, we will present six statistics:

- the mean of all the monitor/model ratios;
- the standard deviation of all the monitor/model ratios;
- the number of sites;
- the proportion of monitored averages which are covered by the range of model estimates for the

county containing the monitoring site;
- Spearman=s correlation coefficient, è; and
- the *p*-value from the hypothesis test that è>0.

i) Mean of Ratios. This number is the same as the asterisk in the box plot in Figure 5. It is just the mean of all the monitor/model ratios in that Astratum@ for that pollutant. Numbers close to 1 here suggest that the model is giving unbiased estimates in this stratum.

ii) Standard Deviation of Ratios. This number gives some idea of the length of the box plot in Figure 5. It measures the variability of the ratios. The smaller this number is, the better. If the model is giving good estimates, the mean should be near 1 and the standard deviation should be small.

iii) Number of Sites. Hopefully these numbers will not be too small. If this number is small for a particular stratum, we can=t have much confidence that the statistics are representative of the model=s performance for that stratum.

iv) Proportion of Sites Covered By County. This statistic is one which helps addresses spatial resolution issues with the model predictions. This statistic compares the monitored average to the range of all estimates in the county containing the monitor. Independent of this model evaluation project, the ASPEN model is being run to generate estimates for 1996 for every census tract centroid in the US (although the results of the air quality predictions will only be presented at a county wide level with a statistical range of census tract concentrations presented). Given a monitoring site, we can determine its county. Then, we can see if the monitored average is covered by the interval *(l,u)*, where *l* is the 10^{th} percentile of all the model estimates for the county, and *u* is the 90^{th} percentile of all the model estimates of the county. The statistic will be the proportion of all sites in the stratum for which this is true. The higher the proportion, the better the performance of the model.

v) Spearman=s correlation coefficient. Like the probability plots discussed above, this statistic tests the relative performance of the model instead of its absolute performance. It is calculated the same way as Pearson=s correlation coefficient, commonly called *r* in statistics textbooks, except that instead of using the actual data, it uses the ranks of the values. The statistic is always between -1 and 1. Values near 1 suggest a strong positive correlation between the model estimates and monitoring averages.

For example, let=s say we have 100 monitors for tetrachloroethylene. We get estimates from ASPEN for each of the sites. Then, we assign ranks for the monitoring averages from 1 to 100, and for the model estimates from 1 to 100. We then form ordered pairs (monitor rank, model rank) for each of the 100 sites, and calculate the correlation between the monitor ranks and the model ranks. If the rankings are exactly the same, we=d get a correlation coefficient of 1; if the model ranks the sites in the reverse order of the monitors, we=d get a correlation coefficient of -1. Positive values suggest a positive association and negative values suggest a negative association. We hope that all the correlation

coefficients in the table are positive.

vi) *P*-value from hypothesis test that è>0. We statistically test the hypothesis that Spearman=s correlation coefficient is greater than 0 using SAS software. The *p*-value from the test will always be between 0 and 1. Small *p*-values suggest that there actually is a positive association between the model estimates and the monitoring averages. It is common to think of *p*-values less than .05 as indicating Astatistically significant@ results. The *p*-values will get smaller as the sample size (statistic (iii) above) and the estimate of Spearman=s correlation coefficient (statistic (v) above) get larger.

IV. Next Steps.

We plan to apply the procedures described above to the 1996 model estimates and the 1996 monitoring averages. Then, we will evaluate the results and make our findings publically available. The procedures may be revised based on the analytical results and on the recommendations from the peer review process.

Climate/ Geographical Regions		EMISSION SOURCES/POPULATION CHARACTERISTICS										
		Population size for Area/Mobile Source Dominated					Population size for Major Source Dominated					Non-Urban
		>3M	1M-3M	1M-500k	500k-250k	<250k	>3M	1M-3M	1M-500k	500k-250k	<250k	<50k
Maritime - North												
Maritime - South												
Continental - North												
Continental - South												
Desert												

NOTES: Population Distribution of MSAs >3M (13), 1M-3M (35), 1M-500k (29), 500k-250k(64), <250k (134)

Table 1. Categorization of Ambient Air Toxics Monitoring Sites and Annual Average Concentration to Prioritize Monitoring Network Design for Model Evaluation.

	Pollutant Level (Monitoring Percentiles)			
Pollutant	0-25	26-50	51-75	76-100
Pollutant A	5.72 6.78 45 .921 .643 .045	4.03 4.47 45 .887 .501 .121	5.44 9.62 45 .793 .438 .187	7.45 6.89 45 .825 .555 .103
Pollutant B	...	...	...	...
!	!	!	!	!

Table 2. Stratification Table by Pollutant Level. The sites are placed into one of four strata based on their monitored annual averages: the lowest 25% are in the first stratum, etc. The six statistics are described in order in Section II(C) of the paper. For example, the point estimate of Spearman=s correlation coefficient for the 45 most polluted sites (according to the monitored annual averages) for Pollutant A is .555.

Appendix A. Completeness Criteria and Calculation of Annual Averages for Monitoring Data.

Since the ASPEN model predicts 1996 annual averages, in order to have a valid comparison, we must also compute 1996 annual averages for the monitoring sites. Some of the monitoring sites have very limited data for 1996. So the first step is to eliminate the (pollutant, site) pairs which are not complete enough to compute an annual average. For example, if we have a site which conducted a short-term study of benzene in the month of June, we would not want to compare this average to model predictions, because June might not be representative of the whole year.

Here is a step-by-step explanation of how we would determine whether a (pollutant, site) pair has a Acomplete year@ for 1996.

1) If we have measurements for at least eighteen hours of the day, then the day is complete.
2) Determine the sampling frequency for 1996, for each quarter (January - March is quarter 1, . . . , October - December is quarter 4). Some (pollutant, site) pairs are regular for 1996 - for example, there is a site in Concord, California which measured styrene every twelve days, from January 4 to December 29. So the sampling frequency for this site is once every twelve days for all four quarters. For irregular sites, just take the mode of the sampling frequencies for each quarter - in other words, take the most Acommon@ sampling interval in the quarter.
3) Use the sampling frequencies in Step 2 to determine the percent of complete days in the quarter. For example, in our Concord, CA site, the sampling frequency is once every twelve days. So there are about 365/4/12, or about 8 sampling days in each quarter. So if quarter 1 has 6 complete days in it, then quarter 1 has 6/8=75% completeness. If the percent of complete days in the quarter is 75% or more, the quarter is complete.
4) If the (pollutant, site) pair has at least one complete Acool@ quarter (quarters 1 and 4) and at least one complete Awarm@ quarter (quarters 2 and 3), this site has a complete year for this pollutant.

If the (pollutant, site) pair does not have a complete year for 1996, it is discarded, and it is not used in any of the model-to-monitor comparisons.

To calculate the annual average, we begin with the daily values. Daily values below the MDL are replaced with MDL/2. We then calculate quarterly averages by averaging all the daily values in each quarter. Finally, we calculate an annual average by averaging two numbers: the average of the one or two Acool@ quarters, and the average of the one or two Awarm@ quarters. Here are three examples:

Site A: Quarter 1 5.0, Quarter 2 incomplete, Quarter 3 6.0, Quarter 4 9.0
Site B: Quarter 1 incomplete, Quarter 2 incomplete, Quarter 3 6.0, Quarter 4 9.0

Site C: Quarter 1 incomplete, Quarter 2 7.0, Quarter 3 6.0, Quarter 4 incomplete
The annual average for Site A is 6.5. The average of Quarters 1 and 4 is 7.0, and the average of Quarter 3 alone is 6.0. The average of 7.0 and 6.0 is 6.5.

The annual average for Site B is 7.5. The average of Quarter 3 alone is 6.0, the average of Quarter 4 alone is 9.0, and the average of 6.0 and 9.0 is 7.5.

Because Site C does not have a complete Acool@ quarter, it does not have a complete year. So it is discarded.

References

Agresti, A. *Categorical Data Analysis*, pp. 84-91. (John Wiley & Sons, New York, 1990.)

Godfrey, B. *Köppen Climate Classification for the Conterminous United States*. On web at **http://snow.ag.uidaho.edu//Clim_Map/koppen_usa_map.htm**.

US EPA. AAir Toxics Monitoring Strategy Concept Paper,@ draft version (2000a).

US EPA. AAssessment System For Population Exposure Nationwide (ASPEN) Model,@ Vol. II, draft version (2000b).