TECHNICAL MEMORANDUM

To: Ted Palma, US EPA

From: Arlene Rosenbaum, Michael Huang, and Jonathan Cohen

Date: September 30, 2002

Re: Comparison of HAPEM5 and APEX3

INTRODUCTION

The Hazardous Air Pollutant Exposure Model, version 4 (HAPEM4) was used to estimate exposure concentration to 10 demographic groups as part of the National Air Toxics Assessment (NATA) national-scale assessment (NSA). After reviewing the NATA-NSA, EPA's Science Advisory Board suggested that a "case study" exposure assessment be conducted for benzene using HAPEM. In the companion memorandum "Benzene Case Study for PO3-NTA006-ICF" we described the results of such a case study for a small geographical area: the central portion of the Houston MSA for which EPA has already conducted air dispersion modeling at a fine spatial scale.

HAPEM4 incorporates several sources of variability in its exposure concentration predictions: activity patterns, commuting patterns, and air quality among census tracts. However, some of the other input parameters and variables are treated as point values, even though they are likely to vary: microenvironment factors and ambient concentrations within census tracts. In addition, the extrapolation of available short-term activity pattern data (i.e., 1 to 3 day duration) to annual sequences is an uncertain process that could be implemented in alternative ways.

In the companion memorandum we presented results for a series of HAPEM simulations, some requiring a modified version of HAPEM, HAPEM5, to explore the implications of introducing variability in microenvironment factors and ambient concentrations within census tracts, as well an alternative approach for extrapolating short-term activity pattern data to annual sequences. Note that the version of HAPEM5 used for the analyses in this memo includes the alternative approach for extrapolating short-term activity pattern data to annual sequences. In the companion memo this model is referred to as variability simulation 7.

The goal of the analysis in this memo is to compare the results of HAPEM5 benzene exposure simulations with results using the APEX3 exposure model. There are several differences between the two models. One major structural difference is that APEX3 takes hourly air quality data while HAPEM5 uses annual average air quality data. Another structural difference is that HAPEM5 generates a fixed number of replicates of annual exposure concentrations for each demographic group in a census tract, whereas APEX3 produces annual exposure concentrations for a set of randomly selected persons in randomly selected census tracts. Thus the number of annual exposure estimates per demographic group and census tract is fixed for HAPEM5 but varies for APEX3.

For the comparison between the two models, the HAPEM5 model was set up as described under variability simulation 7 in the companion memorandum. The APEX3 model was set up using the inputs described in Table 1 and a sample of 19,499 persons in 60 of the 61 Houston census tracts. For a fair comparison, the HAPEM5 data for the tract not simulated with APEX3 was not used in these analyses. Thus the HAPEM5 results were based on 18,000 simulated persons, i.e., 30 replications of 1 person in each of 10 demographic groups and 60 census tracts.

STATISTICAL METHODOLOGY

For the HAPEM5 model, the methods for estimating the mean and variance of total exposure are given in the companion memo and are repeated for convenience here. The mean exposure for a person selected at random from the entire population is estimated by averaging across all combinations of tract and demographic group:

$$\text{Estimated Mean Exposure (Run i, All)} = \frac{\sum_{\text{tracts, groups}} \text{Exposure (Run i, Group j)} \times \text{Population (Group j)}}{\sum_{\text{tracts, groups}} \text{Population (Group j)}}$$

To estimate the mean exposure from 30 runs, the estimated mean exposure per run is averaged across runs:

$$\text{Estimated Mean Exposure} = \frac{\sum_{\text{runs}} \text{Estimated Mean Exposure (Run i)}}{\text{Number of runs}}$$

Since the estimates from each run are statistically independent, the uncertainty (i.e., variance) of the estimated mean exposure is estimated by:

Variance of Mean Exposure = Uncertainty of Mean Exposure =

$$\frac{\sum_{\text{runs}} \text{Estimated Mean Exposure (Run i)}^2 - \text{Estimated Mean Exposure}^2}{\text{Number of runs} - 1}$$

An unbiased estimate for the variance of the exposure across the entire population is given by:

$$\text{Estimated Variance of Exposure (Runs 1 - k, All)} =$$

$$\frac{\sum_{\text{tracts, groups}} \text{Sample Mean Exposure (Group j)}^2 \times \text{Population (Group j)}}{\sum_{\text{tracts, groups}} \text{Population (Group j)}}$$

$$- \left\{ \frac{\sum_{\text{tracts, groups}} \text{Sample Mean Exposure (Group j)} \times \text{Population (Group j)}}{\sum_{\text{tracts, groups}} \text{Population (Group j)}} \right\}^2$$

$$+ \frac{\sum_{\text{tracts, groups}} \text{Sample Variance Exposure (Group j)} \times \text{Population (Group j)} \times (\text{k} - 1)/\text{k}}{\sum_{\text{tracts, groups}} \text{Population (Group j)}}$$

$$+ \frac{\sum_{\text{tracts, groups}} \text{Sample Variance Exposure (Group j)} \times \text{Population (Group j)}^2 /\text{k}}{\left\{ \sum_{\text{tracts, groups}} \text{Population (Group j)} \right\}^2}$$

The sample means and sample variances are the within tract arithmetic mean and variance across the k runs.

Therefore, using 30 runs, an estimate of the variance of exposure is given by:

$$\text{Estimated Variance of Exposure} = \frac{\sum_{i=1}^{10} \text{Estimated Variance of Exposure (Runs } 3i\text{-2, } 3i\text{-1, } 3i\text{)}}{\text{Number of groupings (} = 10\text{)}}$$

and the uncertainty (i.e., variance) of the estimated variance of exposure is given by the formula:

Variance of Variance of Exposure = Uncertainty of Variance of Exposure =

$$\frac{\sum_{\text{runs}} \text{Estimated Variance of Exposure (Run i)}^2 - \text{Estimated Variance of Exposure}^2}{\text{Number of groupings - 1}}$$

Similar calculations apply for specific demographic groups.

For the APEX3 model, the calculations are easier since the sample can be assumed to behave as a simple random sample from the entire Houston population, rather than a stratified random sample with replacement. The mean and variance of the exposure are estimated by the sample mean $\overline{X}$ and sample variance S^2 of all the exposure estimates, across the entire population or across a specified demographic group. Assuming the exposure distribution is approximately

normally distributed, then the uncertainty (i.e., variance) of the mean is given by S^2 / N, where N is the population size, and the uncertainty of the variance is given by $2S^4 /(N-1)$.

These estimates are tabulated in Table 2. For each demographic group (including "All") and for the two statistics "Mean" and "Variance," Table 2 gives the estimated mean and variance of exposure for the two models. Table 2 also gives a Z statistic for testing whether the statistics are equal for the two models and its p-value. P-values below 0.05 are statistically significant at the 5 percent significance level. The Z statistic, which is approximately standard normally distributed is defined by:

$$Z = \frac{\text{Mean (Statistic, HAPEM5) - Mean (Statistic, APEX3)}}{\sqrt{\{\text{Variance (Statistic, HAPEM5)} + \text{Variance (Statistic, APEX3)}\}}}$$

where "Statistic" is either the mean or the variance of the exposure.

RESULTS

For the estimates of mean exposure, the two models give statistically significantly different results only for demographic groups 5, 9, and 10, and for the entire population (at the five percent significance level). These are males and females over 65 and females 18-64. In nearly every case the HAPEM5 estimated mean exposure is higher than the APEX3.1 estimated mean exposure.

For the estimates of the variance of exposure, the two models did not give statistically significantly different results for any of the demographic groups, nor for the entire population (at the five percent significance level). In nearly every case the HAPEM5 estimated variance of exposure is lower than the APEX3.1 estimated variance of exposure.

These results suggest that APEX3.1 and HAPEM5 give similar estimated distributions for the population exposure.

Table 1. Input Files for APEX3 and HAPEM5 Comparison Study

Input Files				
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	Comparability/Differences
Unit 11- District location file This file provides the latitudes and longitudes of air data monitoring locations. The file is used along with the user-defined air-radius to define the geographical area covered by the air quality data. The air quality data from a monitoring location are used for the census tracts within its covered area.	The latitudes and longitudes of locations slightly off the centroid of 61 Houston census tracts are provided in this file. The air-radius of 1 mile is used to ensure that the average air quality data for a census tract will be used only for this census tract.	No corresponding file in HAPEM5. However, HAPEM5 requires the user to provide the census-tract specific air quality data. There is no need to assign the air quality data to a census tract.	N/A	File configurations are different, but implementation should be comparable. (See discussion of Unit 17.)
Unit 13- Temperature zone Location file This file provides the latitude and longitude of the meteorological station. The file is used along with the user-defined Zone-radius to determine the area covered by the temperature data.	The latitude and longitude of Houston airport national weather station is provided in this file	No corresponding file	N/A	N/A
Unit-14 Age-group and employment file This file provides a list of probabilities of employment for each of age groups in	The following probabilities provided by Ted Palma were used in the case study: <u>Age G</u> <u>Employment Prob</u> 0-4 0 5-9 0	In HAPEM5, any age/gender/daytype group will be have the potential for commuting if more than 10% of all CHAD activity records in that	Set up internally in Hapem5 0-4 0 5-11 0 12-17 1 18-65 1	Not comparable

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
population input files. The probabilities are used to determine if a simulated individual will commute to another census tract to work.	10-14 0.1 15-17 0.5 18-19 0.6 20-20 0.8 21-21 0.9 22-24 1.0 25-29 1.0 30-34 1.0 35-39 1.0 40-44 1.0 45-49 1.0 50-54 1.0 55-59 1.0 60-61 1.0 62-64 0.8 65-66 0.5 67-69 0.5 70-74 0.2 75-79 0.1 80-84 0.0 85-99 0.0	age/gender/daytype group indicate commuting. Whether the simulated individual representing the age/gender/day type group actually commutes depends on the activity patterns selected.	>65 1 0 - non commuting 1 - commuting	
Unit-15 Commuting flow file This file provides probabilities of a worker commuting to various destination census tracts from any given home tract.	N/A	COMM2000.txt This file provides probabilities of a worker commuting to various destination census tracts from any given home tract.	N/A	Comparable. The HAPEM5's COMM2000.txt file was developed based on APEX3.1's commuting flow file.
Unit 16 - Temperature data file.	The 1995 daily maximum temperature data for the	No corresponding files in HAPEM5. No temperature	N/A	Not comparable

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
This file provides the daily maximum temperature for the period of simulation. The daily maximum temperature is used to determine window positions and group activity pattern pools in APEX3.1	Houston Airport national weather station was provided in this file. The data was extracted from the Hourly United States Weather Observations 1990-1995 (HUSWO) CD. However, the year of the dates was changed to 1996 in order to be consistent with the dates of air quality data. Since 1996 is a leap year, the last day of 1995 February was repeated and recorded as the last day of 1996 February (i.e., 02/29/02)	data are needed to run HAPEM5		
Unit 17 - Air quality data file This file provides the hourly air quality data for each of air monitoring locations listed in the District location file.	Four sources of Houston ISC modeling data were processed according to the following procedures: - -Average hourly data across all the receptor locations within a census tract; - The average hourly data were treated as monitoring data for the centroid of the census tract. Because of difference in ME factors between on-road source and other sources, two hourly air data files were	Air quality data file HAPEM5 allows the user to provide multiple sources of 8 time blocks (3 hours each block) of annual average air quality data and variable background concentrations in a single file. In addition, HAPEM5 also allows multiple records of 8 time blocks of annual average air quality data for each census tract. Then HAPEM5 randomly selects one record from each census tract for each replicate of simulation.	Four sources (onroad, nonroad, major, and area) of annual average, 3-hour time blocks of Houston ISC air quality data were provided in this file. The source-specific ISC data was processed according to the following procedures: - Average hourly data across all the days in the simulation period; - Consolidate the 24 hours of hourly annual average data into (8)	File configurations are different, but implementation should be comparable.

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
	created for the case study. The first file contains the hourly on-road source data. The second file contains the hourly data for sum of non-road source, major source, area source, and ambient background concentrations. The current APEX3.1 has limit for # of air districts. Thus, the above two files were broken into six smaller files with 20 or 21 air districts in each file. These files were used with 2 ME factor files in six separate APEX3.1 runs.		annual average 3-hour time blocks of air data; - Combine all four sources of processed air data and ambient background concentration into a single hapem5 air data file.	
Unit 18 - Activity specific MET file The file provides distribution types and parameters for calculating the MET value for each CHAD activity code. A MET value is a dimension- less ratio of the activity-dependent energy expenditure rate to the resting energy expenditure rate. This file is not used for calculating exposure concentrations.	This file is not used in the case study because the DOSE calculation was not implemented for the case study runs	N/A	N/A	N/A
Unit 19 - Physiology data				

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
file This file provides tables of certain physiological parameters by age and gender. The file is not used for calculating exposure concentrations. However, it is needed for running APEX3.1 even if the DOSE calculation is turned off.	The default file provided by ManTech will be used in the case study runs	N/A	N/A	N/A
Unit 20 - Profile functions file This file provides the definitions of the following user-definable functions: **TempCat** - Binning temperatures into categories **DiaryPools** - Assigning diary pools using day of week and TempCat **IDGRP** - Group number for output labeling not used in internal calculation **Has_GasStove** - Defines probability of having gas stove in a residence **HasPilot** - Probability of having a pilot light, based on HasGasStove **AC_Home** - Probability of having air conditioning at home	The default file provided by ManTech was used in the case study, except for **Has_GasStove** function. In this case study, **Has_GasStove** function was used to define the probability of a residence having an attached garage, since the ME factor data used for the cased study showed no differences between residences with and without gas stoves, but did show differences between homes with and without attached garages. The national average probability of 0.26 was used in the APEX3.1	No corresponding file.	N/A	N/A

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
AC_Car - Probability of having air conditioning in a car **WindowPos** - Probability of windows open or closed, based on AC_Home and TempCat **SpeedCat** - Probability of average car speed categories	run to determine if a simulated individual has an attached garage. If he/she does, an indoor benzene release value was randomly selected based on the distribution data for a residence with an attached garage. The national average probability of 0.26 for a residence having a garage was calculated based on the CHAD activity database.			
Unit 21 - CHAD- Micro mapping file This file provides the mapping from CHAD location codes to APEX3.1 microenvironments. The file lists all the CHAD location codes and their corresponding codes of user-defined microenvironments	The CHAD location codes were be mapped to 34 of the 37 microenvironments defined in HAPEM5. The three HAPEM5 microenvironments not defined in this file are residence-gas stove (14), residence with garage (15), residence with garage and gas stove (16). These microenvironments have the same CHAD location code	No corresponding file. HAPEM5 maps the CHAD location codes to 37 microenvironments. The CHAD records are pre-processed into a file containing time spent in 37 microenvironments at each of 24 hours in a day.	N/A	File configurations are different but implementation should be comparable. (See discussion of Unit 24.)

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
	as residence with no gas stove (13). Thus, they cannot be defined in APEX3.1. However, APEX3.1 allows the user to define the probability of having gas stove or garage in the Profile Function file. (See discussion of Unit 20)			
Unit 22 - CHAD personal info file This file provides information relating to each 24 hour CHAD activity diary	The default file was used	Durhw.fix.txt This file was pre-processed from the CHAD data base. It contains information from each 24 hour CHAD activity diary record about time spent in each of 37 microenvironments and home and work tract for each of the 24 hours. The records are categorized according the age, gender, and day type	The default file was used	File configuration are different, but Unit 22 and Unit 23 together should be comparable to HAPEM5's Durhw.fix.txt
Unit 23 - CHAD Diary events file This file provides the 24 hour event descriptions for all the diary days in the CHAD database.	The default file was used	Durhw.fix.txt This file was pre-processed from the CHAD data base. It contains information from each 24 hour CHAD activity diary record about time spent in each of 37 microenvironments and	The default file was used	Configuration is different, but Unit 22 and Unit 23 together should be comparable to HAPEM5's Durhw.fix.txt

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
		home and work tract for each of the 24 hours. The records are categorized according the age, gender, and day type		
Unit 24 - Micro descriptions file This file contains the definitions of the microenvironments and the microenvironment factors used to determine the exposure concentrations in user-defined microenvironments	The ME factors for 34 of the HAPEM5 microenvironments are defined using the distribution data provided by EC/R. Because of differences in proximity factor between onroad sources and the other three sources, two microenvironment files were created for separate on-road source and non-onroad source runs. The penetration factors are the same in both files. However, the proximity factors (*PROX*) and indoor source factors (*CS* in APEX3.1 and *ADD* in HAPEM5) are different. In the **onroad** file, *PROX* and *CS* were set to distribution data provided by EC/R. In the **non-onroad** file, *PROX* was set to 1.0 and CS was set to zero (to prevent double counting the indoor source in the output file).	Factors.txt This file provides the distribution data for the three mocroenvironment factors (i.e., penetration, proximity, and indoor source) for the 37 microenvironments specified in HAPEM5.	The microenvironment factors (i.e., penetration, proximity, and add) for 37 HAPEM5 microenvironments were specified with the distribution data provided by EC/R	File configurations are different but implementation should be comparable.

Input Files				Comparability/Differences
APEX3.1		HAPEM5		
Uses	Setup in Case Study	Uses	Setup in Case Study	
Unit 24 - Micro descriptions file (Continued)	Both files were set up in a way that the penetration and *CS* factors were selected randomly once every 3 hours (i.e., 8 times a day). The *PROX* factor was selected once and used for all 24 hours. The same set of selected 24 hours values were used for each day in the simulation period. However, a new set of PROX and CS was generated for each work tract. The sampling frequency was set up this way to correspond tot the sampling frequency in HAPEM5. In the on-road microenvironment file, two sets of *CS* distribution data for the residence ME (#13) were specifed - one for residence with garage and the other for residence with no garage.			

Table 2. Comparison of Mean and Variance of Exposure between APEX3 and HAPEM5

		APEX3.1			HAPEM5				
Demographic Group	Statistic	Population	Estimate	Uncertainty (Variance)	No. of Estimates	Estimate	Uncertainty (Variance)	Z	P-value for Testing No Difference Between Models.
01	Mean	1912	5.592	0.0013	30	5.531	0.0280	-0.36	0.72
01	Variance	1912	2.471	0.0064	10	1.626	0.5070	-1.18	0.24
02	Mean	2471	5.171	0.0008	30	5.291	0.0231	0.78	0.44
02	Variance	2471	1.893	0.0029	10	1.296	0.2843	-1.11	0.27
03	Mean	1617	5.239	0.0011	30	5.280	0.0209	0.28	0.78
03	Variance	1617	1.855	0.0043	10	1.268	0.2813	-1.10	0.27
04	Mean	2514	5.424	0.0006	30	5.496	0.0282	0.42	0.67
04	Variance	2514	1.440	0.0016	10	1.641	0.6637	0.25	0.81
05	Mean	996	5.671	0.0023	30	6.249	0.0221	3.70	0.00
05	Variance	996	2.259	0.0103	10	1.722	0.1173	-1.50	0.13
06	Mean	1877	5.586	0.0013	30	5.589	0.0183	0.03	0.98
06	Variance	1877	2.415	0.0062	10	1.602	0.7140	-0.96	0.34
07	Mean	2519	5.203	0.0008	30	5.300	0.0169	0.73	0.47
07	Variance	2519	1.928	0.0030	10	1.426	0.3659	-0.83	0.41
08	Mean	1608	5.311	0.0012	30	5.445	0.0162	1.02	0.31
08	Variance	1608	1.992	0.0049	10	1.641	0.6521	-0.43	0.66
09	Mean	2482	5.452	0.0007	30	5.864	0.0259	2.53	0.01
09	Variance	2482	1.685	0.0023	10	1.266	0.0604	-1.67	0.09
10	Mean	1503	5.752	0.0017	30	6.248	0.0384	2.48	0.01
10	Variance	1503	2.558	0.0087	10	1.691	0.2878	-1.59	0.11
All	Mean	19499	5.412	0.0001	30	5.627	0.0055	2.87	0.00
All	Variance	19499	2.028	0.0004	10	1.550	0.1057	-1.47	0.14