Comparing Air Dispersion Model Predictions with Measured Concentrations of VOCs in Urban Communities

GREGORY C. PRATT,*,[†] CHUN YI WU,[†] DON BOCK,[†] JOHN L. ADGATE,[‡] GURUMURTHY RAMACHANDRAN,[‡] THOMAS H. STOCK,[§] MARIA MORANDI,[§] AND KEN SEXTON[‡,||]

Environmental Outcomes Division, Minnesota Pollution Control Agency, 520 Lafayette Road, St. Paul, Minnesota 55155, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, and Health Science Center at Houston School of Public Health, University of Texas, Houston, Texas 77030

Air concentrations of nine volatile organic compounds were measured over 48-h periods at 23 locations in three communities in the Minneapolis−St. Paul metropolitan area. Concentrations at the same times and locations were modeled using a standard regulatory air dispersion model (ISCST3). The goal of the study was to evaluate model performance by comparing predictions with measurements using linear regression and estimates of bias. The modeling, done with mobile and area source emissions resolved to the census tract level and characterized as model area sources, represents an improvement over large-scale air toxics modeling analyses done to date. Despite the resolved spatial scale, the model did not fully capture the spatial resolution in concentrations in an area with a sharp gradient in emissions. In a census tract with a major highway at one end of the tract (i.e., uneven distribution of emissions within the tract), model predictions at the opposite end of the tract overestimated measured concentrations. This shortcoming was seen for pollutants emitted mainly by mobile sources (benzene, ethylbenzene, toluene, and xylenes). We suggest that major highways would be better characterized as line sources. The model also failed to fully capture the temporal variability in concentrations, which was expected since the emissions inventory comprised annual average values. Based on our evaluation metrics, model performance was best for pollutants emitted mainly from mobile sources and poorest for pollutants emitted mainly from area sources. Important sources of error appeared to be the source characterization (especially location) and emissions quantification. We expect that enhancements in the emissions inventory would give the greatest improvement in results. As anticipated for a Gaussian plume model, performance was dramatically better when compared to measurements that were not matched in space or time. Despite the limitations of our analysis, we found that the regulatory air dispersion model was generally able to predict space and time matched 48-h average ambient concentrations of VOC species within a factor of 2 on average, results that meet regulatory model acceptance criteria.

Introduction

Speciated volatile organic compounds (VOCs) are today routinely measured in urban air, and ambient air concentrations of some compounds, like benzene, often exceed health benchmarks (*1*, *2*). Similarly, modeling studies have predicted concentrations of some VOC species above levels of concern across large portions of the United States (*3*−*5*). These findings have increased concern about the health implications of air toxics and spurred further work on tools for evaluating air toxics concentrations and exposures.

Air dispersion modeling (ADM) is widely used to estimate ambient air pollutant concentrations and, in fact, is required for criteria pollutant (SO_2, NO_2, O_3, Pb, CO, and PM) regulatory programs in the United States. The ADM methodology includes development of an emissions inventory for pollutants of concern within the model domain, followed by calculations of downwind dispersion yielding predictions of concentrations at designated receptor locations. The most commonly used regulatory models treat downwind dispersion as a Gaussian plume formulation. The U.S. Environmental Protection Agency (EPA) has developed regulatory air dispersion models and modeling guidance, including recommended models whose performance has been tested against measurements (*6*). ADM is increasingly used for air toxics, although the performance of the ADM methodology has rarely been compared to toxics measurements, especially over averaging times of less than 1 yr. Both Pratt et al. (*2*) and Rosenbaum et al. (*5*) found that modeled annual average concentrations tended to be lower than measured values. Lorber et al. (*7*) found that EPA regulatory model predictions of 48-h average dioxin air concentrations from emissions of a solid waste incinerator were generally within a factor of 10 of measured air concentrations.

U.S. EPA regulatory air dispersion models undergo an evaluation and validation process before being accepted for use as a regulatory tool, and the EPA has developed criteria for judging model acceptability. An important part of the process involves comparing model predictions with measured data. The necessary measurement databases are typically taken from studies that have concurrently measured meteorological variables, emissions, and concentrations. These standard databases typically involve emissions of one pollutant from a single, isolated point source. It is useful to know how regulatory model predictions compare with monitored values for a number of air toxics emitted from a variety of source types within a complex metropolitan area. It is also of interest to evaluate model performance over periods of time shorter than 1 yr under simulation conditions similar to those often seen in regulatory settings.

The work described here occurred within a larger study of personal exposure to air toxics. The study design and initial monitoring results were presented in earlier publications (*8*−*10*). The concept behind the larger study was to simultaneously measure personal exposures (i.e., concentrations near the breathing zone), residential indoor air concentrations, outdoor concentrations at the home, and outdoor air

* Corresponding author phone: (651)296-7664; fax: (651)297-7709; e-mail: gregory.pratt@pca.state.mn.us.

[†] Minnesota Pollution Control Agency.

[‡] University of Minnesota.

[§] University of Texas.

[||] Present address: School of Public Health, University of Texas, Brownsville Regional Campus, RAHC Building, 80 Fort Brown, Brownsville, TX 78520.

10.1021/es030638l CCC: $27.50
Published on Web 02/28/2004

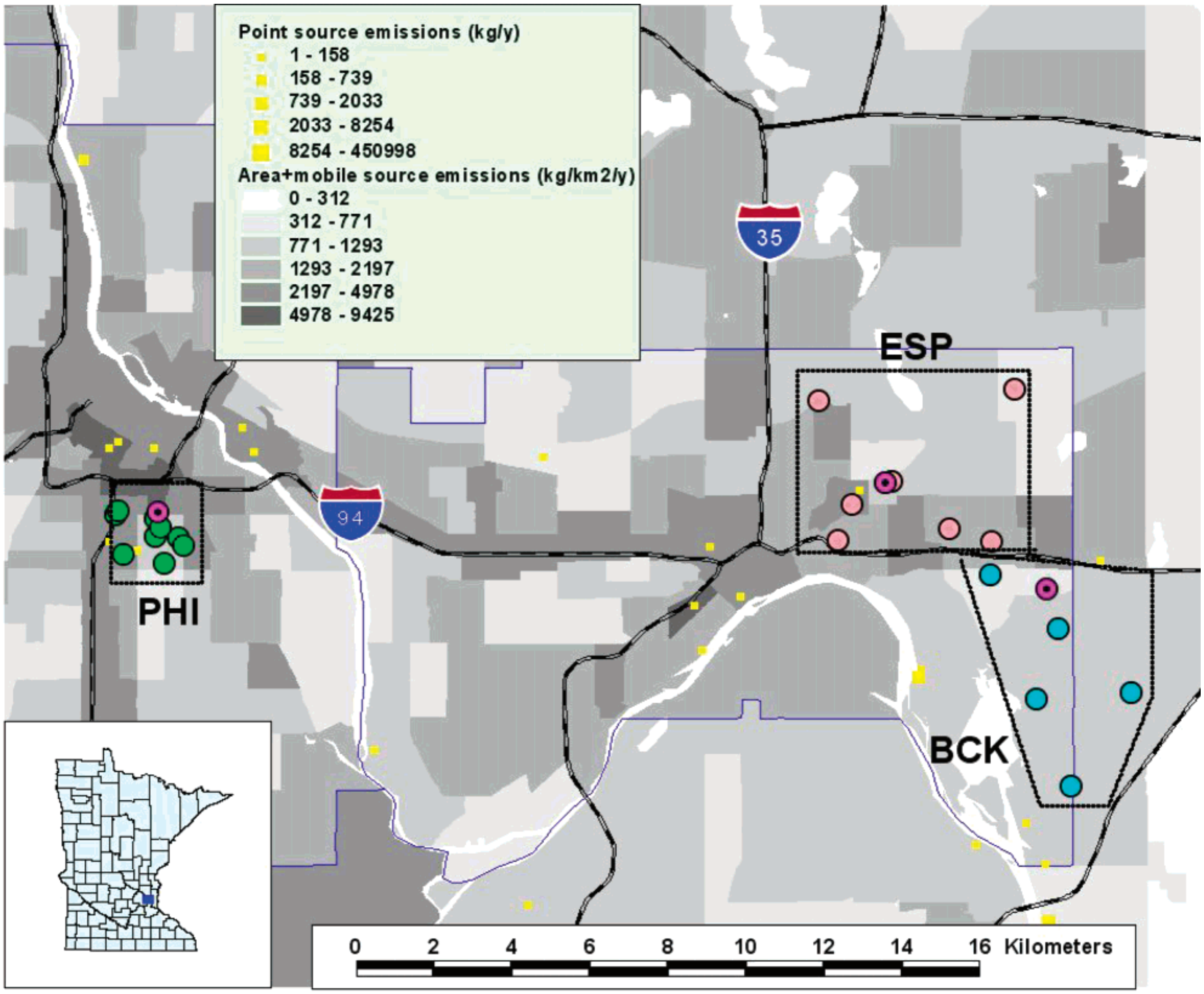

FIGURE 1. Map of the study area (i.e., the center of the Minneapolis–St. Paul metropolitan area) showing the three communities (outlined with dotted lines), the study homes (circles), the community monitoring sites (circles with center dots), and the locations of point sources emitting benzene (yellow squares). Census tracts are shaded according to the sum of area and mobile source benzene emissions in the census tract.

at community monitoring sites and finally to model ambient concentrations at the outdoor locations using standard regulatory modeling methods. The goal was to investigate the relationships among these various measures, recognizing that personal exposure is the most relevant for human health impacts. It was found in earlier work (*8*, *9*) that outdoor air concentrations were poorly correlated with indoor and personal concentrations. This paper examines one facet of the overall study (i.e., the usefulness of regulatory ADM for estimating air toxics concentrations in outdoor air). The analysis focuses on comparing concentrations measured at outdoor locations with predictions from a widely used regulatory air dispersion model.

Study Design

Three communities, Phillips (PHI), East St. Paul (ESP), and Battle Creek (BCK), were selected to cover a range of expected concentrations in the central metropolitan area determined from a pilot modeling analysis (*11*). PHI is a densely populated, predominantly minority, inner-city community in south Minneapolis where outdoor VOC concentrations were expected to be among the highest in the metropolitan area due mainly to mobile source emissions. ESP is a blue-collar, racially mixed community in St. Paul with elevated VOC concentrations expected in parts of the community due mainly to industrial point sources. BCK is a predominantly white, affluent community on the eastern edge of St. Paul with the lowest expected VOC concentrations of the three communities. Figure 1 is a map of the study area showing benzene emissions from point, area, and mobile sources and the locations of monitors.

Monitoring locations were chosen in each community from among the homes and community monitoring sites available from the larger personal exposure study ($n = 6$, 7, and 10 for BCK, ESP, and PHI, respectively) and thus were a convenience sample not selected with any spatial relationship to the milieu of sources within the community. Monitors were deployed for 48-h sampling periods. Two types of monitors were used; passive diffusion-based organic vapor monitors (OVMs) were deployed at all locations (outside study participants' homes and at the central community monitoring site) and stainless steel canisters were deployed at central community monitoring sites (one site per community).

Sampling days were chosen based upon the design of the larger personal exposure study. There were three sampling seasons: spring, summer, and fall in 1999. During each season, 48-h sampling periods were begun every third day, with one idle day between samples. Measurements were made at the centralized community sites on every sampling day. OVM samples were collected outside study participants'

TABLE 1. Analytical Detection Limits (ADLs), Number of Measurements (both Canister and OVM) below the ADL, Number ≤0 That Were Substituted with Half the ADL, and Number of Measurements below the ADL but >0 That Were Retained in the Analysis

	canister (n = 160)			OVM (n = 223)		
	ADL	total no. below ADL	no. = 0 substituted with half ADL	ADL	total no. below ADL	no. ≤ 0 substituted with half ADL
benzene	0.13	0	0 (0%)	0.11	0	0 (0%)
chloroform	0.12	56	56 (35%)	0.11	170	170 (76%)
ethylbenzene	0.12	1	1 (0%)	0.16	22	3 (1%)
dichloromethane	0.18	15	9 (6%)	0.73	193	34 (15%)
styrene	0.16	14	3 (2%)	0.19	148	127 (57%)
tetrachloroethylene	0.16	160	156 (98%)	0.10	28	6 (3%)
toluene	0.09	0	0 (0%)	0.11	41	40 (18%)
trichloroethylene	0.32	83	8 (5%)	0.10	111	60 (27%)
xylenes	0.16	1	1 (0%)	0.11	16	7 (3%)

homes on a subset of sampling days (to match days when personal and indoor sampling was conducted at a home). The number of OVM samples collected outside a given participant's home ranged from 1 to 11, while the number of canister and OVM samples collected at the central community sites ranged from 46 to 55. All samples were analyzed in the laboratory for a suite of nine pollutants.

Canister VOC Measurements

One canister VOC monitoring station was established in each of the three communities (Figure 1). The canister measurement methodology was described previously (*2*). Briefly, VOC samples were collected following the U.S. Federal Reference Method (TO-14A) in evacuated, summa-polished, stainless steel canisters, two-valve model (Scientific Instrumentation Specialists, Moscow, ID). The canisters were deployed using a Xon Tech model 910A canister sampler housed in an enclosure that allowed heating during the cold season (Xon Tech, Inc., Van Nuys, CA). Samples were collected for 48 h to be comparable with the sampling period for personal organic vapor monitors. Sample analysis was done using a Varian Saturn model 2000 gas chromatograph/mass spectrometer (Varian, Inc., Palo Alto, CA). Laboratory duplicates run daily showed high precision (Pearson's correlation coefficients >0.94 for all analytes except chloroform = 0.82). Collocated samplers run by the same laboratory (but at sites outside of this study) also showed high reproducibility (Pearson's correlation coefficients >0.78 for all analytes except chloroform = 0.64, trichloroethylene = 0.66, and styrene = 0.66). Xylenes were measured as two sets of isomers, *o*-xylene, and *m,p*-xylene. These two were added to give total xylenes for comparability with the modeling analysis because emissions data for the model were typically available for total xylenes.

Concentrations that were less than the analytical detection limit but that produced an instrument reading greater than zero were included in calculations. Concentrations that produced an instrument reading of zero were also included in calculations by assigning them a value of one-half the analytical detection limit defined as the standard deviation of seven replicate analyses of a standard prepared to five times the estimated detection limit divided by the square root of n (i.e., 7) and multiplied by the Student's t-value appropriate for a 99% confidence level with $n - 1$ (i.e., 6) degrees of freedom. Out of 160 total canisters deployed, the number of values below the ADL ranged from none for benzene and toluene to 160 for tetrachloroethylene (Table 1). Given that all canister measurements were below the ADL, tetrachloroethylene canister measurements were dropped from further analysis.

Personal Organic Vapor Monitors (OVMs)

Concentrations of VOCs were also measured using charcoal-based passive diffusion samplers referred to as organic vapor monitors (OVMs, model 3500, 3M Corporation, Maplewood, MN). The method was described previously (*8*, *12*). OVMs can be effectively used to measure VOC concentrations, and canister and OVM VOC measurements have been found to be in good agreement (*13*). In a pilot study, a sampling period of 48 h was required to obtain quantifiable results for the suite of pollutants of interest at the concentrations present in the study area.

Concentrations that were less than the analytical detection limit but that produced an instrument reading greater than zero were included in calculations. Concentrations that produced an instrument reading ≤0 were included in the calculations by assigning them a value of one-half the analytical detection limit (ADL). The ADL was determined from an analysis of seven solutions of each of a series of low-concentration standards. The ADL was defined as the standard deviation of the seven analyses of the lowest concentration standard that yielded a relative standard deviation of ≤10%, multiplied by the Student's t-value appropriate for a 99% confidence level with $n - 1$ (i.e., 6) degrees of freedom. Duplicate OVMs were run in approximately 10% of cases and showed generally good reproducibility (Pearson's correlation coefficients >0.81 for all analytes). The ADLs and the number of measurements below the ADL are given in Table 1. Duplicates were treated as separate samples in the statistical analysis, although we recognize that this approach slightly overweights their importance. The statistical analysis was done using SPSS (version 8.0.2, SPSS, Inc., Chicago, IL).

Modeling Methods

Air dispersion modeling was done to predict 48-h average ambient concentrations that were matched in space and time with the canister and OVM VOC measurements. The recommended U.S. EPA air dispersion model at the time of the study, ISCST3 version 02035, was used with regulatory default model options (*6*). Model receptors were located at all sampling sites. Terrain elevations were not included because the sampling areas do not have large variations in terrain (terrain elevations are often omitted from regulatory modeling analyses in the study area). Atmospheric chemistry was not considered, and although we recognize that this omission may introduce errors, we believe them to be small. For example, with average wind speeds in the study area of 5 m/s and source−receptor distances averaging about 10 km, the average source−receptor transport times are less than 1 h. This time was considered short enough so that atmospheric chemistry could be ignored given the reactive half-lives of the chemicals in our study. Meteorological data were taken from the U.S. National Weather Service site at the Minneapolis−St. Paul international airport for 1999 and processed for model input for each of the times when monitoring was done, thereby allowing the model runs to be done using the meteorological data for the precise times when the monitoring occurred. Airport data were used because they were readily available, because on-site meteorological data were not collected at the monitoring locations, and to be consistent with common practice in regulatory ADM. The most important meteorological variables in determining the predicted concentration are wind speed, wind direction, stability category, and mixing height. It is our experience that these parameters are typically fairly uniform across the metropolitan area except during frontal passages, so the use of airport data is not expected to introduce substantial errors. Furthermore, since our goal was to test the regulatory model as a tool for predicting air toxics concentrations, we chose

TABLE 2. Percentage of Pollutants Emitted from Each Major Source Category in the Modeling Analysis and Percentage of Modeled Concentrations Accounted for by Each Source Category for Each Community[a]

pollutant	source category	emissions (%)	modeled concns (%)		
			BCK	ESP	PHI
benzene	point	1	1	0	0
	area	26	12	13	9
	mobile	73	87	86	91
chloroform	point	26	6	6	4
	area	74	94	94	96
	mobile	0	0	0	0
ethylbenzene	point	5	4	4	6
	area	10	4	5	2
	mobile	85	92	91	92
dichloromethane	point	21	38	39	39
	area	79	62	61	61
	mobile	0	0	0	0
styrene	point	55	10	10	9
	area	1	1	1	0
	mobile	44	89	89	91
tetrachoroethylene	point	14	5	3	3
	area	86	95	97	97
	mobile	0	0	0	0
toluene	point	5	5	16	2
	area	37	39	37	41
	mobile	58	55	46	57
trichloroethylene	point	66	56	71	90
	area	34	44	29	10
	mobile	0	0	0	0
xylenes	point	7	6	5	5
	area	34	40	44	44
	mobile	59	54	51	51

[a] Total tons of emissions can be calculated by ratioing from the values in Table 3. Publicly owned treatment works emissions from Table 1 are included with area source emissions despite being modeled as point sources.

to apply the model using airport meteorological data as is usually done in regulatory analyses.

Annual average point, area, and mobile source emissions in 1999 for each pollutant were estimated as part of the Minnesota Pollution Control Agency (MPCA) air toxics emissions inventory using the Regional Air Pollutant Inventory Development System (RAPIDS is an emission inventory tool developed by the Great Lakes Commission; *14*). Table 2 shows the percentage of emissions of each pollutant from each major source category. Point sources were defined as larger stationary sources whose emissions are tabulated individually in the regulatory agency emission inventory system.

Air toxics emissions from point sources were determined by direct facility reporting, by the use of emission factors, and by incorporating data from the U.S. Emergency Planning and Community Right to Know Act Toxics Release Inventory. Point source locations were determined by facility self-reporting, global positioning, and geographic information system (GIS) addressing matching. Point source stack parameters were taken from (i) regulatory agency (MPCA) files; (ii) default values developed by the Ozone Transport and Assessment Group (OTAG) by source classification code (OTAG is a partnership between the U.S. EPA, the Environmental Council of the States (ECOS), and various industry and environmental groups aimed at creating agreements among industry and government for control of ground-level ozone and related pollutants in the eastern United States); or (iii) average OTAG values across all facilities. A total of 425 point sources in the metropolitan area were included in the modeling analysis. Within a given facility, stack-by-stack emissions were not available. Therefore, each facility was represented as a single stack whose location was taken as the centroid of the facility (when available) or as the location of the front entrance. Similarly, stack parameters were taken as averages across all emission points at the facility, weighted by the throughput for the emission point.

Area sources were defined as stationary sources whose emissions are not individually tabulated in the point source emissions inventory. Area source emissions for 1999 were developed by the MPCA using the RAPIDS system. Emissions were estimated from 22 area source categories, although eight of the source categories did not emit the pollutants considered in this study. One area source category, publicly owned treatment works, was modeled as individual point sources. The RAPIDS system generates area and mobile source emissions on a county basis. To capture greater spatial resolution, the county total emissions were allocated to census tracts by one of four methods. For most area source categories, the emissions assigned to a census tract were taken as the county total emissions multiplied by the fraction of the county population residing in the census tract. Landfill emissions were assigned to the census tract in which the landfill was located. Emissions from marking of traffic lanes were apportioned according to the fraction of the county total lane miles occurring in the census tract, and wildfire emissions were apportioned by land area. Table 3 gives the area source categories, the method of apportioning county-wide emissions estimates into census tracts, and the total emissions of each pollutant from each source category.

Mobile source emissions from RAPIDS were available for 1997 at the time of the study. Mobile sources consist of two major subcategories: on-road mobile sources include cars, trucks, and buses and non-road mobile sources include aircraft, watercraft, railways, construction equipment, farm equipment, snowmobiles, lawn and garden equipment, and other related subcategories. The details of emissions development for each category can be found at the Great Lakes Commission website (*14*). The resulting emissions are given as county totals. In this study, on-road mobile source emissions were apportioned to census tracts as the fraction of the 1999 vehicle miles traveled in the census tract relative to the county total. Vehicle miles traveled were determined from traffic count data collected by the Minnesota Department of Transportation, subdivided by roadway category. The traffic count data were combined with data on the total miles of each roadway category in each census tract using a geographic information system. This calculation allowed a determination of the vehicle miles traveled in each roadway category within a census tract.

Aircraft emissions were apportioned to the census tracts in which the airports were located, depending upon the proportion of air traffic occurring at each airport. Railway emissions were apportioned to census tracts according to the length of railway in the tract as a fraction of the county total. All other non-road mobile source emissions were apportioned to census tracts according to population. This apportioning by population was considered appropriate because in the metro area a large fraction of non-road mobile source emissions was attributable to lawn and garden equipment. For example, 60% of non-road mobile source benzene emissions in the largest metropolitan area county (Hennepin) was attributed to lawn and garden equipment.

Mobile and area source emissions that were apportioned to census tracts were represented as polygon area sources in the ISCST3 model. The polygons were taken from a GIS coverage of 1990 census tracts and processed so that each was represented by no more than 10 vertices. This simplification was required to reduce model calculation time and to meet model limitations. Within the center of the metropolitan area, most census tracts are simple polygons so the simplification process did not appreciably change the geometric representation of most census tracts in the area of interest for this study.

TABLE 3. Area Source Emissions Categories, Pollutants and Amounts (metric tons per year) Emitted by Category, and Method of Determining Census Tract Emissions from County Total Emissions. Column Totals Subject to Rounding Discrepancies

area source category	method of determining census tract emissions from county emissions	benzene	chloroform	dichloro-methane	ethyl-benzene	styrene	tetrachloro-ethylene	toluene	trichloro-ethylene	total xylenes
agricultural pesticide application	not done									
architectural surface coatings	population parsing	9		166	167			201		101
asphalt paving	not done									
auto body refinishing	population parsing	61						349		835
chromium electroplating	not done									
consumer and commercial solvent use	population parsing	0	2	4	77		60	913	1	432
dry cleaning	population parsing						90			
gasoline marketing	population parsing	146		27				1202		696
graphic arts	population parsing	55		8				322		35
hospital sterilizers	not done									
human cremation	not done									
industrial surface coating	population parsing			10				1279		1411
landfills	assign to census tract	1	1	12	8	–	5	29	5	35
marine vessel loading etc.	not done									
prescribed burning	not done									
public owned treatment works	done as point sources	236	53	1	56	12	57	1699	87	596
residential fuel combustion	population parsing	0						35		
residential wood combustion	population parsing	1203						453		126
solvent cleaning	population parsing	95						785		322
structure fires	not done									
traffic lane marking	lane miles parsing									1
wild fires	area parsing	91						46		4
totals		1896	56	228	309	12	212	7313	93	4594

Background concentrations recommended by Rosenbaum et al. (*5*) were added to the modeled concentrations for chloroform (0.083 μg/m^3), dichloromethane (0.15 μg/m^3), tetrachloroethylene (0.14 μg/m^3), trichloroethylene (0.081 μg/m^3), and xylenes (0.17 μg/m^3). The recommended background concentration for benzene (0.48 μg/m^3) was not used. Background is defined as that part of the total concentration not accounted for explicitly in the modeling analysis and includes long-range transport, persistent historical emissions, and nonanthropogenic emissions. A variety of types of observations (taken from the literature) were used to estimate background concentrations, including the midrange of observations specified as background, the lower end of the range specified as the Northern Hemisphere average, the lower end of the range specified as the global average, and the lower end of the range specified as remote or rural. The dates of the studies used for background were benzene, 1985; chloroform, 1990; dichloromethane, 1990; tetrachloroethylene, 1994; trichloroethylene, 1990; and xylenes, 1990. The benzene estimate is the oldest and perhaps the most tenuous because benzene emissions decreased nationally during the 1990s due to changes in gasoline formulation, meaning that the background concentration had decreased by the time of our study. Accurate benzene emissions data have only recently begun to be collected in the metro area, but the recent data illustrate the decrease in benzene emissions. Hennepin County (the largest metropolitan area county) benzene emissions decreased from 1.9 million lb in 1997 to 1.5 million lb in 1999. For these reasons, the benzene background concentration used by Rosenbaum et al. (*5*) was considered out of date and was not used.

Results and Discussion

Table 2 shows the percentages of modeled concentrations attributable to point, area, and mobile source emissions categories. In many cases, the percentage of model-predicted pollutant concentrations from a source category was different from that category's percentage of emissions. For example, the percentage of predicted benzene concentrations attributable to mobile sources was higher than the percentage of benzene emissions from mobile sources. The percentage of predicted benzene concentrations attributable to point sources was smaller than the percentage of emissions from point sources. Similar patterns were seen for other pollutants as well. These results are likely due to the model release characteristics for mobile source emissions (e.g. as a model area source with ground-level emissions). In contrast, point source emissions were simulated as a release from an elevated stack, usually with thermal and mechanical buoyancy. This finding suggests that ground-level sources such as mobile sources may contribute more to local concentrations per mass of emissions than traditional, elevated-release point sources.

Concentrations of all substances were low as compared to measurements and model predictions in other large urban areas. Both measured and modeled concentrations of nine VOCs were lowest on average in BCK and higher in the other communities (Table 4), although the differences between communities were small for all substances. The variability in modeled and measured concentrations and in model–monitor differences was greatest in PHI and least in BCK for most pollutants. The high variability in PHI is likely due to the steeper emissions gradients in that community, especially in mobile source emissions (shown for benzene by the shading in Figure 1).

The aim of this analysis was to compare the results from a regulatory type air dispersion modeling analysis with measurements. We collected measurements using two methods, canisters and OVMs. On the basis of the fact that the canister method is the U.S. EPA Federal Reference Method, we assume greater confidence in that method. Nevertheless, we believe it is useful to compare the model predictions with results from each method until further studies systematically comparing the two methods can be done. In most cases, we have also analyzed the results separately by community due to the differences in both

TABLE 4. Monitored Mean Concentrations (μg/m^3), Mean Differences between Model Predictions and Measurements (μg/m^3), Root Mean Squared Error, and Fractional Bias[a]

		monitor site canisters			monitor site OVMs			home outdoor OVMs		
pollutant	metric	BCK ($n = 54$)	ESP ($n = 55$)	PHI ($n = 51$)	BCK ($n = 49$)	ESP ($n = 50$)	PHI ($n = 46$)	BCK ($n = 26$)	ESP ($n = 18$)	PHI ($n = 34$)
benzene	mon. mean	0.9	1.9	1.5	1.0	2.1	1.8	0.9	1.7	2.0
	mean diff.	0.3	−0.6	1.4	0.2	−0.9	1.3	0.4	−0.2	−0.1
	RMSE	0.5	1.1	1.8	0.5	1.3	1.8	0.6	0.7	1.0
	Fx-Bias	−0.3	0.4	−0.6	−0.1	0.5	−0.5	−0.4	0.1	0.0
chloroform	mon. mean	0.1	0.1	0.3	0.1	0.1	0.1	0.0	0.0	0.1
	mean diff.	0.0	0.0	−0.2	0.0	0.0	0.0	0.1	0.1	0.0
	RMSE	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1
	Fx-Bias	−1.2	−0.1	1.0	−0.4	−0.3	−0.1	−0.8	−0.8	0.6
dichloromethane	mon. mean	0.4	0.5	0.6	0.2	0.5	0.4	0.2	0.4	0.6
	mean diff.	−0.3	−0.4	−0.4	−0.1	−0.3	−0.2	−0.1	−0.2	−0.5
	RMSE	0.8	0.6	1.4	0.3	0.8	0.4	0.3	0.4	0.7
	Fx-Bias	0.9	1.0	1.2	0.3	1.0	0.8	0.3	0.7	1.2
ethylbenzene	mon. mean	0.4	1.1	0.8	0.3	1.0	0.8	0.3	0.7	0.9
	mean diff.	−0.0	−0.6	0.2	0.0	−0.6	0.2	0.1	−0.3	−0.3
	RMSE	0.2	0.9	0.5	0.2	0.8	0.5	0.2	0.5	0.5
	Fx-Bias	0.1	0.9	−0.3	0.0	0.8	−0.2	−0.3	0.5	0.4
styrene	mon. mean	0.3	0.4	0.4	0.1	0.2	0.2	0.1	0.1	0.2
	mean diff.	−0.2	−0.2	0.1	0.1	0.0	0.3	0.1	0.1	0.1
	RMSE	0.2	0.3	0.3	0.1	0.2	0.4	0.2	0.2	0.2
	Fx-Bias	0.6	0.7	−0.2	−0.5	0.0	−0.9	−0.6	−0.7	−0.4
tetrachloroethylene	mon. mean	na[b]	na	na	0.2	0.4	0.4	0.7	0.4	0.8
	mean diff.	na	na	na	−0.0	−0.2	−0.2	−0.5	−0.2	−0.6
	RMSE	na	na	na	0.2	0.3	0.3	0.9	0.2	1.0
	Fx-Bias	na	na	na	0.2	0.6	0.5	1.1	0.6	1.1
toluene	mon. mean	2.0	8.4	3.9	2.5	9.4	3.6	2.1	3.8	4.4
	mean diff.	1.7	−2.3	4.0	1.0	−3.7	4.8	1.8	0.8	1.9
	RMSE	2.1	7.1	5.0	4.7	14.3	6.5	4.1	3.1	4.4
	Fx-Bias	−0.6	0.3	−0.7	−0.3	0.5	−0.8	−0.6	−0.2	−0.4
trichloroethylene	mon. mean	0.3	0.4	0.6	0.1	0.1	0.2	0.1	0.1	0.3
	mean diff.	−0.1	−0.3	−0.4	−0.0	−0.0	−0.0	0.0	0.0	−0.1
	RMSE	0.3	0.4	0.5	0.1	0.2	0.2	0.1	0.1	0.2
	Fx-Bias	0.7	0.9	1.0	0.1	0.1	0.1	−0.2	−0.4	0.3
xylenes	mon. mean	1.8	5.1	3.7	1.6	4.8	3.9	1.4	3.5	4.6
	mean diff.	0.8	−1.9	2.4	1.0	−1.8	2.5	1.4	−0.4	0.5
	RMSE	1.2	3.3	3.3	1.3	3.0	3.5	1.6	1.9	2.2
	Fx-Bias	−0.4	0.5	−0.5	−0.5	0.5	−0.5	−0.7	0.1	−0.1

[a] See text for definition. Negative mean differences are indicative of model underprediction. [b] na, not available.

measured and modeled concentrations between the communities.

One way to compare modeled concentrations to measurements is to plot the modeled values against the measured values and to calculate linear regression statistics. Figure 2 shows scatterplots of model predictions versus measurements (along with regression statistics) for benzene, ethylbenzene, and xylenes. All sites and sample types within each community are included; however, the following interpretations are not changed when the regressions are done separately by sample type and community. Model performance was best in BCK and worst in PHI. It was best for for benzene, ethylbenzene, and xylenes and worst for chloroform, dichloromethane, and tetrachloroethylene with intermediate performance for styrene, toluene, and trichloroethylene. Two points should be made about these comparisons. First, the model appears to perform the best for pollutants emitted predominantly from mobile sources, worst for pollutants emitted predominantly from area sources, and intermediate for pollutants mainly from point sources. Second, the model performance is best for pollutants whose measurements are always or nearly always above detection limits and worst for pollutants that are often below detection. The treatment of values below detection is often a difficult matter. There is useful information in values below detection, and excluding such data may introduce errors into an analysis. We do not know the exact value for a particular below detection measurement, but we do know that the value is low, and it is within a specific range. If, for example, the model prediction is low at the same time that the measurement is below detection, then we have some (albeit imperfect) information about model performance.

It can be seen in Figure 2 that the slopes of the regression lines for the relationships between model predictions and measurements are less than 1. Furthermore, at high measured concentrations the model predictions tended to fall below the 1:1 line, while at low measured concentrations, the model predictions tended to fall above the 1:1 line. This result was true for all pollutants and indicates that there was a tendency for the model to overpredict when the monitored concentrations are low and to underpredict when the monitored concentrations are high. Since the model used annual average emissions, this finding is not surprising. We infer that the model did not capture the full temporal variability due to emissions variations on seasonal, weekly, and diurnal scales.

Although linear regression is a common method for comparing the association between two variables, it is not a complete characterization of the association. For example, R^2 values can be high in the case where two variables are strongly related but there is a consistent bias (e.g., if a model prediction is always half of the measured value). In addition, it can be shown that potentially different values of R^2 could result from different distributions of the measured variables.

Another way of visualizing modeled versus measured results is shown in Figure 3 as boxplots of the differences between model-predicted benzene concentrations and mea-

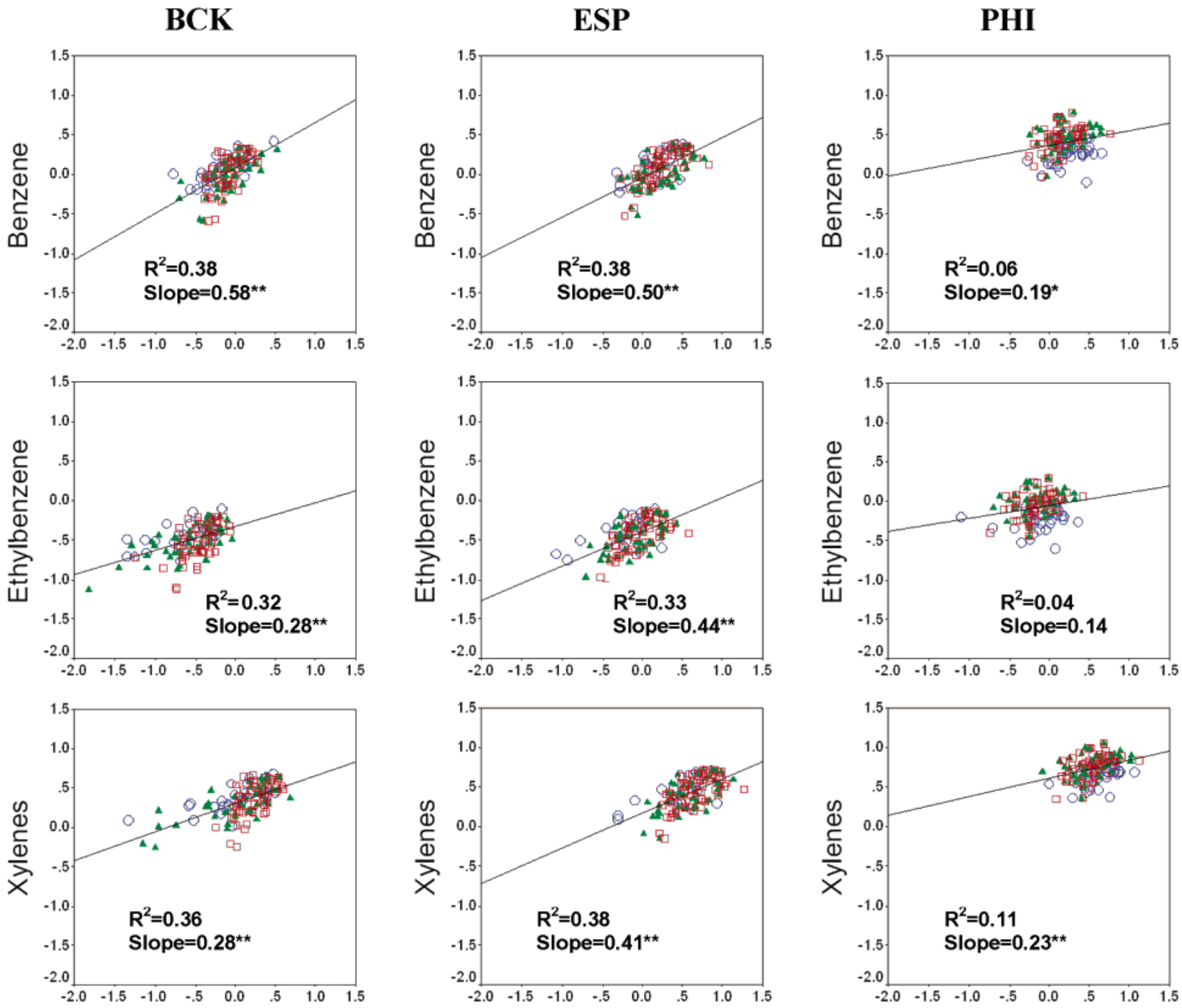

FIGURE 2. Scatterplots of log-modeled concentrations (*y*-axis) vs log-monitored concentrations (*x*-axis) matched in time and space for three pollutants in each of the three communities. The linear regression lines are shown. The points represent all samples taken in the community, both at the community monitoring site and at study participants' homes, by canister and by OVM. The 1:1 line (not shown) would extend diagonally from lower left to upper right of each plot. The regression coefficient and the slope are given. Asterisks indicate the *p*-value of the slope ($<$0.01 and $<$0.001). Squares indicate canister measurements, circles represent OVMs at participants' homes, and triangles indicate OVMs at community monitoring sites.

sured values. Each bar shows the distribution of model−monitor differences for one sampling location, with the number of observations at that site shown beside the bar. In general, there was a tendency for the model to overpredict benzene concentrations at the BCK sites and to underpredict at ESP sites, and there was a range from underprediction to overprediction at PHI sites. A similar pattern of overprediction in BCK, underprediction in ESP, and mixed results in PHI was also found for ethylbenzene, toluene, and xylenes. For the other pollutants, this pattern of over- versus underprediction did not apply.

The overprediction at PHI was most noticeable and occurred mainly at the community monitoring site (see bars labeled PHI91OVM and PHI92CAN). Predictions at the other PHI sites were closer to measurements. The reason for the overprediction at this location appears to lie in the spatial representation of mobile source emissions. The north end of PHI abuts a major interstate highway exchange (I94−I35W commons) that is one of the most heavily trafficked road sections in the metropolitan area. Seventy-three percent of estimated benzene emissions were from mobile sources (Table 1), and these emissions were represented as census tract-sized polygonal area sources in the model (Figure 1). Census tracts spanning the major highway at the north end of PHI had some of the highest estimated emissions densities of mobile source pollutants in the metro area. One census tract (27053006000) was unusually shaped and encompassed a large section of the busiest part of the highway while also extending a considerable distance (600 m) south of the highway. The PHI community monitoring site was located at the far southern extreme of tract 27053006000, at the point most distant from the highway. Other nearby census tracts that spanned the highway extended only a short distance (280 m) from the highway.

The ISCST3 model algorithm for representing area source emissions (the representation we used for area and mobile sources) is based on a numerical integration over the area in the upwind and crosswind directions. Since the algorithm estimates the integral over the area upwind of the receptor location, receptors may be located within the area itself. This is an improvement over area source representations in previous air toxics modeling (*3*−*5*) in which area/mobile source emissions were characterized as pseudo-stacks. The ISCST3 area source algorithm assumes a uniform emissions density within an area source, and the model predictions at a receptor located within the area source are strongly

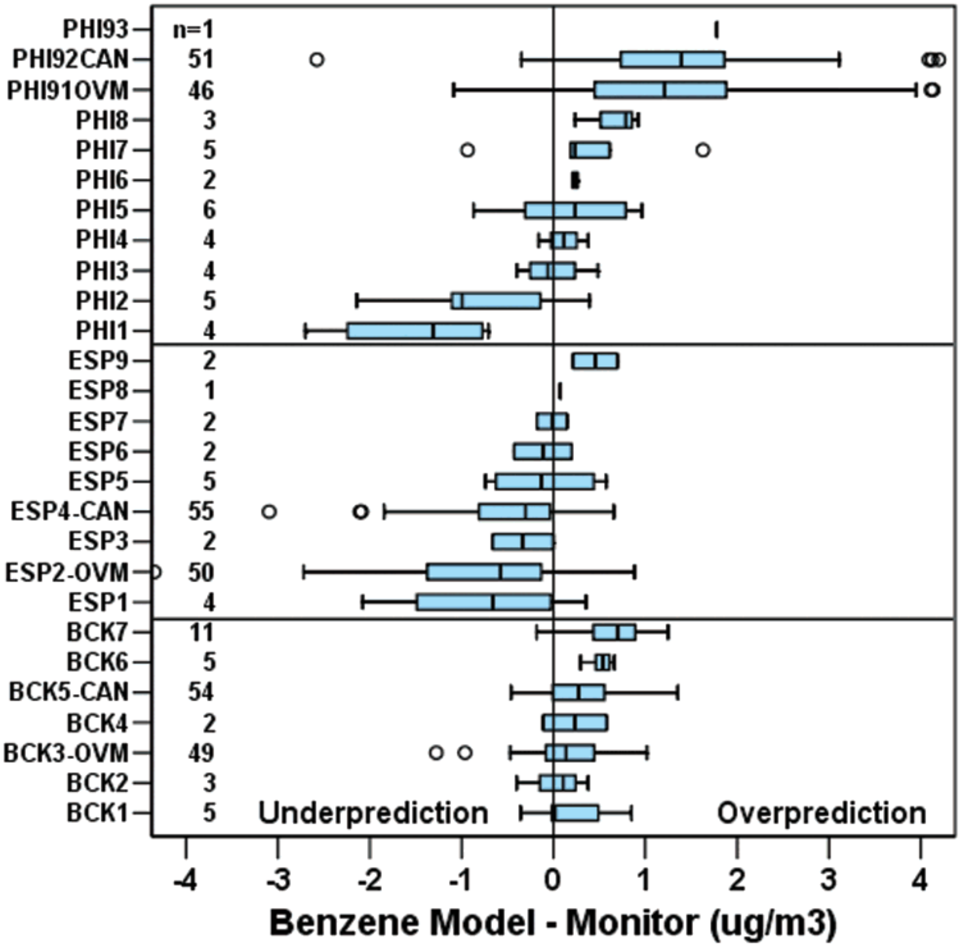

FIGURE 3. Boxplot of differences between modeled and monitored ambient concentrations of benzene at each of the receptors (i.e., community monitoring sites and study homes). Community monitoring sites are designated by the last three letters "CAN" to designate a canister sample or "OVM" to designate an OVM sample. All sites without designation were OVM samples. Each bar represents the results for all monitoring days (including canister and OVM measurements) at one receptor and shows the 25th−75th percentile values. The center line within the bar is the median, and the arms extend to encompass all values not considered outliers (outliers defined as values greater than 1.5 box lengths from the edge of the box are shown as circles).

influenced by the emissions from that source. If the assumption of uniform emissions is violated, the model will likely overpredict at points of low emissions and underpredict at points of high emissions within the area source.

In the present case, the combined factors of the high emissions at one end of census tract 2705300600, the sharp gradient in emissions, the shape of the tract, the location of the monitoring site at the extreme end of the tract, and the nature of ISCST3 area source algorithm probably caused the model overprediction for benzene at the PHI community monitoring site. This overprediction at the Phillips community monitoring site was also seen for the other pollutants emitted predominantly from mobile sources (i.e., ethylbenzene, toluene, and xylenes). We infer from these results that the spatial representation of emissions, when characterized as model area sources, can be especially important in an area of sharp emissions gradients, and care must be taken to capture the relevant scale of spatial variability in such cases. Our analysis at the census tract resolution did not adequately capture the gradient in mobile source emissions near a major roadway. An alternative approach, such as representing major highways as line sources, would likely improve model performance in this regard.

Linear regression is a useful tool for judging model performance, but it is only one metric, and it may not be useful when a large percentage of measurements are below detection. For example, consider a case where all the measurements are below detection. The model could, in one scenario, predict values across a wide range, all above the measurement detection level. In another scenario, the model could always predict values below the measurement detection level. In both cases, given the lack of detection in the measurements, the linear regressions will be insignificant, but we think it is clear that the model is more useful in the second scenario.

The final way in which we have considered model−monitor comparisons is by characterizing the bias. We have chosen to present three metrics relating to the bias, the mean error (e.g., the mean difference between the model and the measurement), the root mean-squared error (RMSE), and the fractional bias (Table 4). Each of these metrics has pros and cons. The mean error is easily understood and preserves the sign of the bias. The RMSE is a measure of the deviations from the 1:1 relationship and preserves the scale of the original measurements. It is derived from the mean squared error, which is comprised of bias (the extent of over or under estimation) and variance (precision). The fractional bias is presented because it is the statistic recommended by U.S. EPA for model−monitor comparisons.

There were a total of 383 samples taken, 160 canisters and 223 OVMs. Since each sample was analyzed for nine pollutants, there were 9 × 383 = 3447 possible model−monitor comparisons. Overall, the model results were more likely to underpredict than to overpredict (e.g., there were

1235/3447 cases (36%) where the mean error was positive (34% for canisters and 37% for OVMs)). By pollutant, the percentages of positive mean errors (overprediction) were benzene (61%), chloroform (<1%), dichloromethane (<1%), ethylbenzene (40%), styrene (56%), tetrachloroethylene (3%), toluene (69%), trichloroethylene (14%), and xylenes (62%). From this breakdown by pollutant, it is clear that for chloroform, dichloromethane, and tetrachloroethylene the model almost always predicts lower than the measurement. However, these pollutants were often below the ADL, and the measurement values often consisted of one-half ADL replacement values. By community (excluding chloroform, dichloromethane, and tetrachloroethylene), the percentages of positive mean errors (overprediction) were BCK (59%), ESP (23%), and PHI (68%). Thus, it appears that there was a tendency for model overprediction at BCK and especially at PHI and a tendency for model underprediction at ESP.

With a few exceptions, the RMSE was lowest in BCK, intermediate in ESP, and highest in PHI (Table 4). Since the RMSE measures the deviation from the 1:1 relationship, these results indicate that the model performed best in BCK and worst in PHI. Comparing RMSE between different pollutants does not provide useful information because the magnitude of the measurements affects the RMSE, and the concentrations of the different pollutants have different magnitudes. Thus, we cannot use the RMSE to evaluate the model performance across pollutants. There was no difference in the RMSE between sampling methods.

U.S. EPA guidance for selecting the best performing air dispersion model (*15*) states, “Although a completely objective basis for choosing a minimum level of performance is lacking, accumulated results from a number of model evaluation studies suggests that a factor-of-two is a reasonable performance target a model should achieve before it is used for refined regulatory analyses”. The guidance goes on to recommend the fractional bias as a screening tool for evaluating whether a model should be eliminated from consideration. The fractional bias is calculated as

$$\mathrm{FB} = 2\left[\frac{\mathrm{OB} - \mathrm{PR}}{\mathrm{OB} + \mathrm{PR}}\right]$$

where OB and PR refer to the average observed and predicted values. The EPA guidance suggests that the fractional bias calculation be made using the highest 25 concentrations (unmatched in space and time), but due to the relatively small *n* for a given pollutant by community comparison, we used all available data. Furthermore, we used data matched in space and time as a stricter test of model performance. The fractional bias was chosen by the EPA because it is symmetrical and bounded, with values ranging between +2 (extreme underprediction) and −2 (extreme overprediction). In addition, the fractional bias is dimensionless, allowing comparisons among different pollutants and concentration levels. Fractional bias values between +0.67 and −0.67 are equivalent to model predictions within the factor of 2 criterion, an EPA criterion for acceptance.

Table 4 shows the fractional bias values for each pollutant by sample type and community. The EPA fractional bias criterion for model acceptance (between +0.67 and −0.67) was met for benzene across all communities and sample types. For other pollutants, the acceptance criterion was generally met. Notable exceptions were dichloromethane and both tetrachloroethylene and trichloroethylene in canisters. Based on the fractional bias criterion, the model usually performed better in BCK than in the other two communities, and there was a tendency for better model performance as compared to OVM measurements than as compared to canister measurements.

The modeling was done such that the predictions were matched in space with the measurements. In addition, the predictions were matched in time as nearly as possible (e.g., the meteorological data were matched in time, but the emissions data were taken as annual average values since these were the only data available). The comparisons discussed previously were done with the data matched in space and time. However, it is widely recognized that Gaussian plume models are not formulated to predict a given space and time matched event, rather the model results represent the ensemble average of a population of events that could occur under a given set of meteorological conditions. Figure 4 shows model−monitor comparisons for benzene unmatched in space and time, matched in space, matched in time, and matched in both space and time. It is clear from the figure that when the space and time matching criterion is removed, the model predictions improved dramatically—in going from comparisons matched in time and space to unmatched comparisons, the RMSE decreased from 1.19 to 0.32 and the regression coefficient increased from 0.27 to 0.97. Similar improvements were seen for other pollutants.

The sources of error in model predictions include model formulation, model inputs (met data, terrain data, source physical representation, emissions data), and stochastic nature of the atmosphere. Sources of measurement error include a range of factors such as sample handling, methodological uncertainties, analytical equipment performance, and accuracy of analytical standards. On the basis of measurement precision and accuracy (*8*) and on estimates of model uncertainty (*16*), we believe that the measurement error to be small as compared to model error.

In general, the differences in model performance among pollutants were greater than among communities. A key question is why the model performed better for some pollutants and some locations than others. Possible reasons include the model formulation itself and the inputs to the model such as meteorological data, terrain data (or the lack of it), source characterization data, and emissions data. The model algorithms generally treat all pollutants alike; however, they may not account for, or may treat incorrectly, some atmospheric processes that are more important for some pollutants than others. For example, the model (as it was run) did not account for processes that remove a pollutant from the atmosphere, such as deposition or chemical reaction. In general, the substances considered in this study are volatile and not thought to be removed from the atmosphere substantially by deposition. With regard to chemical transformation, the substances for which the model performed most poorly are the substances with the longest atmospheric half-lives of the pollutants in our study (i.e., those that are least susceptible to degradation). The model performed better for substances with shorter half-lives (those that are more susceptible to degradation). It may be inferred from such evidence that deposition and chemical transformation are not likely to be the source of the difference in model performance for the compounds considered here, and we reject the hypothesis that the model formulation is responsible for differences in performance among pollutants and locations. We also reject meteorological data and terrain data as sources of discrepancies in model performance among pollutants since these were identical for all pollutants.

Two remaining factors that might account for the difference in model performance for certain compounds are the physical representation of the sources and the quantification of the emissions data. It is true that some sources were not adequately represented by treating the emissions as a polygonal area source with the dimensions of the census tract. However, the model generally performed better for pollutants emitted primarily from mobile sources. These are

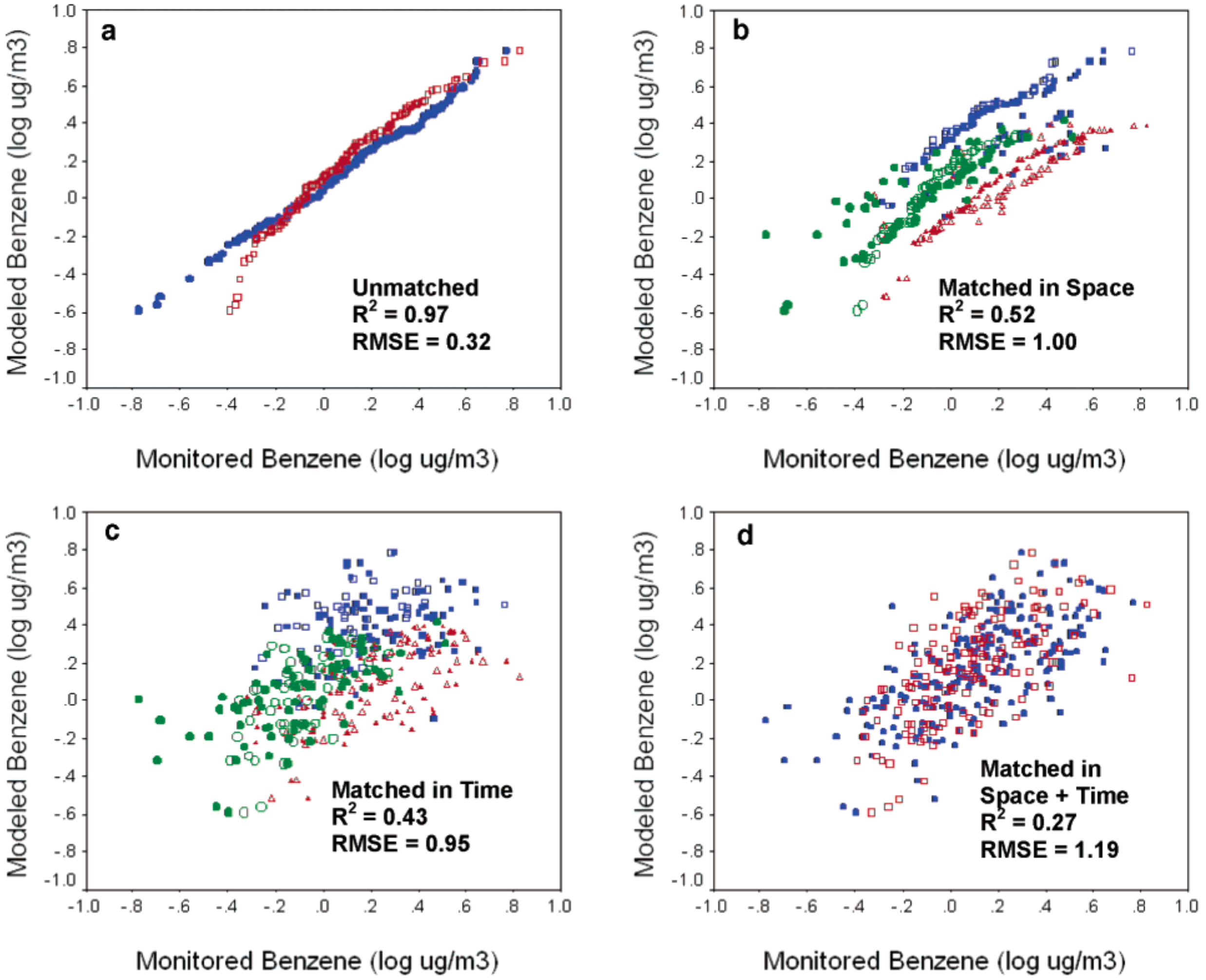

FIGURE 4. Scatterplots of log-modeled benzene concentrations (*y*-axis) vs log-monitored benzene concentrations (*x*-axis). The 1:1 line (not shown) would extend diagonally from lower left to upper right of each plot. Data from all communities and sample types are plotted. Panel a shows values unmatched in space and time. Panel b shows values matched in space but not in time. For both panels a and b, open squares represent canister measurements, and closed circles represent OVM measurements. Panel c shows values matched in time but not in space. Panel d shows values that are matched in both space and time. For both panels c and d, squares represent PHI, triangles represent ESP, circles represent BCK, open symbols represent canisters, and closed symbols represent OVMs.

ubiquitous ground-level sources in the metro area, so representing them as model area sources is probably reasonable except where strong gradients exist, such as near major highways. Mobile source emissions from major highways would likely be better characterized as line sources.

In contrast to mobile sources, point and area source emissions usually occur at discrete locations, and precise locational data and release characteristics (stack parameters) are important for their characterization. In our study, we used locational data for point sources, although in most cases we did not have exact stack locations or stack parameters. Nevertheless, the point source locational data and stack data were better than for area sources that were modeled as census tract-sized area sources. These issues of locational data and stack parameters appear to be important factors in the differences in model performance among pollutants.

The quantification of emission rates is also an important source of error. Sax and Isakov (*16*) studied regulatory model uncertainty and found that emissions estimation was the primary source of uncertainty. We used the best available emissions data, which were annual average emissions. Clearly, the temporal variations in emissions need elucidation to improve predictions. Finally, we note that, as a general rule, air dispersion model performance improves as averaging time increases. The 48-h average predictions in this study are relatively short-term as compared to other VOC modeling studies (*3*−*5*). Despite the limitations of this modeling analysis and the short averaging time, the ISCST3 regulatory model results were generally within a factor of 2 of measurements for most VOCs, meeting the EPA model acceptance criteria.

Acknowledgments

We are indebted to the air quality monitoring staff of the Minnesota Pollution Control Agency. This research was supported by grants from the U.S. Environmental Protection Agency's Science to Achieve Results (STAR) Program. We also thank 3M Corporation for providing organic vapor monitors. Although the research described in this paper was funded in part by the U.S. Environmental Protection Agency's STAR Program through Grants GR825241-010 and R827928-010, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Literature Cited

(1) South Coast Air Quality Management District C. *Multiple Air Toxics Exposure Study in the South Coast Air Basin*, MATES-II, Final Report; available at http://www.aqmd.gov/news1/MATES_II_results.htm (accessed August 18, 2003)

(2) Pratt, G. C.; Palmer, K.; Wu, C. Y.; Oliaei, G.; Hollerbach, C.; Fenske, M. J. *Environ. Health Perspect.* **2000**, *108*, 815−825.

(3) U.S. Environmental Protection Agency, OAR. *National Air Toxics Assessment*; available at: http://www.epa.gov/ttn/atw/nata/ (accessed August 8, 2003).

(4) Woodruff, T. J.; Axelrad, D. A.; Caldwell, J.; Morello-Frosch, R.; Rosenbaum, A. *Environ. Health Perspect.* **1998**, *106*, 245−251.

(5) Rosenbaum, A. S.; Axelrad, D. A.; Woodruff, T. J.; Wei, Y. H.; Ligocki, M. P.; Cohen, J. P. *J. Air Waste Manage. Assoc.* **1999**, *49*, 1138−1152.

(6) Guideline on Air Quality Models. *Code of Federal Regulations*, Vol. 40, Part 51, Appendix W.

(7) Lorber, M.; Eschenroeder, A.; Robinson, R. *Atmos. Environ.* **2000**, *34*, 3995−4010.

(8) Sexton, K.; Adgate, J. L.; Ramachandran, G.; Pratt, G. C.; Mongin, S. J.; Stock, T. H.; Morandi, M. T. *Environ. Sci. Technol.* **2004**, *38*, 423−430.

(9) Adgate, J. L.; Ramachandran, G.; Pratt, G. C.; Waller, L. A.; Sexton, K. *Atmos. Environ.* **2002**, *36*, 3255−3265.

(10) Adgate, J. L.; Ramachandran, G.; Pratt, G. C.; Waller, L. A.; Sexton, K. *Atmos. Environ.* **2003**, *37*, 993−1002.

(11) Pratt, G. C.; McCourtney, M.; Wu, C. Y.; Bock, D.; Sexton, K.; Adgate, J.; Ramachandran, G. Measurement and source apportionment of human exposures to toxic air pollutants in the Minneapolis−St. Paul metropolitan area. In *Proceedings of a Specialty Conference Co-sponsored by the Air and Waste Management Association and the U.S. Environmental Protection Agency*, Cary, NC, 1998; pp 64−72.

(12) Chung, D.; Morandi, M. T.; Stock, T. H.; Afshar, M. *Environ. Sci. Technol.* **1999**, *33*, 3661−3665.

(13) Stock, T. H.; Morandi, M. T.; Pratt, G. C.; Bock, D.; Cross, J. H. Comparison of the results of VOC monitoring with diffusive samplers and canisters. In *Proceedings of the 8th International Conference on Indoor Air Quality and Climate, Indoor Air '99*, Edinburgh, Scotland, 1999; Raw, G., Aizlewood, C., Warren, P., Eds.; Edinburgh, Scotland, 1999.

(14) Great Lakes Commission. *Overview of the RAPIDS System*; available at http://www.glc.org/air/rapids/rpdsover.html #rpdsdbs (accessed August 18, 2003)

(15) U.S. Environmental Protection Agency, OAQPS. *Protocol for determining the best performing model*; EPA-454/R-92-025; U.S. Government Printing Office: Washington, DC, 1992.

(16) Sax, T.; Isakov, V. *Atmos. Environ.* **2003**, *37*, 3481−3489.

Received for review September 22, 2003. Revised manuscript received January 8, 2004. Accepted January 20, 2004.

ES030638L