NUATRC

NUMBER 15
2009

RESEARCH REPORT

A Pilot Study using EPA's CMAQ Model and Hospital Admission Data to Identify Multipollutant "Hot Spots" of Concern in Harris County, Texas

Winifred J. Hamilton, Daewon W. Byun, Wenyaw Chan, Jason K.S. Ching, Younghun Han, Ricardo A. Lopez, Violeta F. Coarfa, and DaeGyun Lee

MICKEY LELAND NATIONAL URBAN AIR TOXICS RESEARCH CENTER

ABOUT THE NUATRC

The Mickey Leland National Urban Air Toxics Research Center (NUATRC or the Leland Center) was established in 1991 to develop and support research into potential human health effects of exposure to air toxics in urban communities. Authorized under the Clean Air Act Amendments (CAAA) of 1990, the Center released its first Request for Applications in 1993. The aim of the Leland Center since its inception has been to build a research program structured to investigate and assess the risks to public health that may be attributed to air toxics. Projects sponsored by the Leland Center are designed to provide sound scientific data useful for researchers and for those charged with formulating environmental regulations.

The Leland Center is a public-private partnership, in that it receives support from government sources and from the private sector. Thus, government funding is leveraged by funds contributed by organizations and businesses, enhancing the effectiveness of the funding from both of these stakeholder groups. The U.S. Environmental Protection Agency (EPA) has provided the major portion of the Center's government funding to date, and a number of corporate sponsors, primarily in the chemical and petrochemical fields, have also supported the program.

A nine-member Board of Directors oversees the management and activities of the Leland Center. The Board also appoints the thirteen members of a Scientific Advisory Panel (SAP) who are drawn from the fields of government, academia and industry. These members represent such scientific disciplines as epidemiology, biostatistics, toxicology and medicine. The SAP provides guidance in the formulation of the Center's research program and conducts peer review of research results of the Center's completed projects.

The Leland Center is named for the late United States Congressman George Thomas "Mickey" Leland from Texas who sponsored and supported legislation to reduce the problems of pollution, hunger, and poor housing that unduly affect residents of low-income urban communities.

This project has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement X83234601. The contents of this document do not necessarily reflect the views and policies of the Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

PREFACE

The Clean Air Act Amendments of 1990 established a control program for sources of 188 "hazardous air pollutants, or air toxics" that may pose a risk to public health. With the passage of these Amendments, Congress established the Mickey Leland National Urban Air Toxics Research Center (NUATRC) to develop and direct and environmental health research program that would promote a better understanding of the risks posed to human health by the presence of these toxic chemicals in urban air.

Established as a public/private research organization, the Center's research program is developed with guidance and direction from scientific experts from academia, industry, and government and seeks to fill gaps in scientific data. These research results are intended to assist policy makers in reaching sound environmental health decisions. The NUATRC accomplishes its research mission by sponsoring research on human health effects of air toxics at universities and research institutions and by publishing research findings in its "NUATRC Research Reports," thereby contributing meaningful and relevant data to the per-reviewed literature.

In December 2003, the Center released RFA 2003-01, "NUATRC Small Grants Program: Exposures and Health Effects of Urban Air Toxics," which encouraged investigators to develop and conduct short-term research projects dealing with non-carcinogenic health effects of air toxics in human subjects. These projects were envisioned as pilot projects that could serve as a basis for more extended future research. Projects that have an interdisciplinary approach, projects that test new techniques, and innovative or high-risk projects were strongly encouraged.

The Center was particularly interested in innovative projects in the area of exposure and health effects of urban air toxics in human subjects. The studies were to be hypothesis-driven and designed to test the relationship between exposures to air toxics under environmentally relevant conditions and health effects in urban communities. Dr. Winifred Hamilton of Baylor College of Medicine and co-investigators at Baylor, University of Houston, University of Texas Health Science Center - Houston, and the U.S. Environmental Protection Agency were awarded a two-year contract to conduct the research project, "A Pilot Geospatial Analysis of Exposure to Air Pollutants (with Special Attention to Air Toxics) and Hospital Admissions in Harris County, Texas."

The study investigators' research in this pilot study was organized around the hypothesis that the rate of Harris County residents hospitalized during the study period differs geographically among 337 4x4 kilometer domains, and correlates with exposure to air pollutants estimated using the EPA's Community Multiscale Air Quality with Air Toxics (CMAQ-AT) model. They used the CMAQ model and model inputs (meteorology and emissions) developed by the University of Houston as part of previous ozone modeling for Houston's TexAQS field study to estimate pollutant concentrations; this project then used ArcGIS (ESRI, Redlands, CA) geospatial modeling software to extract and/or combine the estimated exposure, admissions, and demographic data for each of the domains for subsequent analysis. The investigators hypothesized that the CMAQ model would allow air toxic concentrations to be determined to a more comprehensive extent than can be accomplished using fixed site monitors.

According to the investigators, the primary goal of this pilot study was to advance the identification of potential "hot spots" of disproportionate exposure and health effects by (1) using an air pollution simulation model to estimate exposure fields, and (2) utilizing actual health endpoints to predict risk. The investigators used simulated hourly values from MM5 and CMAQ models for a 90-day period in the year 2000 for 25 meteorological and air pollutant variables including 16 air toxics for 337 4 X 4-km cells in Harris County, Texas. Each cell was characterized by the simulated output, hospital admissions rates by discharge diagnosis (cardiovascular, respiratory, or either), and seven demographic variables. The investigators explored principal components analysis and various averaging schema. They used a linear mixed effects regression model to generate regression coefficients to calculate predicted admission rates and used geospatial techniques to visualize potential "hot spots."

Despite generating the best emissions and meteorological data that could be produced for the study time period, the model-predicted pollutant concentrations proved not to be adequate for the objectives of this health effects study. Thus, the investigators did not provide results relating model-predicted exposure to health endpoints. Instead, the report focuses on methodology and delineates areas that need further development.

A "Statement by the NUATRC Scientific Advisory Panel" (or the "Statement") precedes this report, and highlights some of the strengths and limitations of this research so that, once refined, this approach might be used for areas of the country with significant pollutant gradients, such as Houston, TX with localized areas of high pollution or "hot spots."

Each time a NUATRC-funded study is completed, the Investigators submit a draft final research report. The draft final report undergoes an extensive evaluation procedure that asses the strengths and limitations of the study and comments on clarity of the presentation, data quality, appropriateness of study design, data analysis, and interpretation of the study findings. The objective of the review process is to ensure that

the Investigator's report is complete, accurate, and clear.

The evaluation first involves an external review by a team of three reviewers, including a biostatistician. The reviewers' comments are then considered by members of the Scientific Advisory Panel (SAP). The comments of the external reviewers and the SAP members are then provided to the Investigator. In its communication with the Investigator, the SAP may suggest alternative interpretations for the results, and also discuss new insights that the study may offer to the scientific literature. The Investigator has the opportunity to exchange comments with the SAP and, if necessary, revise the draft report. The SAP may also publish its own comments regarding the strengths of the study as well as alternate interpretations of the study results in the "Statement by the NUATRC SAP" (or the "Statement"). In accordance with NUATRC policy, the Board of Directors approves the publication of the revised final report after the recommendation of the SAP. The research presented in the NUATRC Research Reports represents the work of its investigators; the comments presented in the "Statement" represent the views of the SAP.

The NUATRC appreciates hearing comments from its readers from industry, academic institutions, government agencies, and the public about the usefulness of the information contained in these reports, and about other ways that the NUATRC may effectively serve the needs of these groups. The NUATRC wishes to express its sincere appreciation to Dr. Winifred Hamilton and her research team, the SAP, and external peer-reviewers whose expertise, diligence, and patience have facilitated the successful completion of this report.

STATEMENT OF THE NUATRC SCIENTIFIC ADVISORY PANEL

Statement by the Scientific Advisory Panel of the Mickey Leland National Urban Air Toxics Research Center Regarding the Research Report, "A Pilot Study Using EPA's CMAQ Model and Hospital Admission Data to Identify Multipollutant 'Hot Spots' of Concern in Harris County, Texas."

BACKGROUND

In 2003, the Center released RFA 2003-01, "NUATRC Small Grants Program: Exposures and Health Effects of Urban Air Toxics." The primary objective of the RFA was to encourage investigators to develop and conduct innovative short-term research projects dealing with non-carcinogenic health effects of air toxics in human subjects. The projects were envisioned as pilot projects that could serve as a basis for more extended future research. Dr. Winifred Hamilton of Baylor College of Medicine and co-investigators at Baylor, University of Houston, University of Texas Health Science Center - Houston, and the U.S. Environmental Protection Agency were awarded a two-year contract to conduct the research project, "A Pilot Geospatial Analysis of Exposure to Air Pollutants (with Special Attention to Air Toxics) and Hospital Admissions in Harris County, Texas."

RESEARCH APPROACH

The study investigators' research in this pilot study was organized around the stated hypothesis that the rate of Harris County residents hospitalized during the study period differs geographically among 337 4x4 kilometer domains, and correlates with exposure to air pollutants estimated using the EPA's Community Multiscale Air Quality with Air Toxics (CMAQ-AT) model and model inputs (meteorology and emissions) developed by the University of Houston as part of previous ozone modeling for Houston's TexAQS field study. This project then used ArcGIS (ESRI, Redlands, CA) geospatial modeling software to extract and/or combine the estimated exposure, admissions, and demographic data for each of the domains for subsequent analysis. The investigators hypothesized that the CMAQ model would allow air toxic concentrations to be determined to a more comprehensive extent than can be accomplished using fixed site monitors.

According to the investigators, the primary goal of this pilot study was to advance the identification of potential "hot spots" of disproportionate exposure and health effects by (1) using an air pollution simulation model to estimate exposure fields, and (2) utilizing actual health endpoints to predict risk. The investigators used simulated hourly values from MM5 and CMAQ models for a 90-day period in the year 2000 for 25 meteorological and air pollutant variables including 16 air toxics for 337 4 X 4-km cells in Harris County, Texas. Each cell was characterized by the simulated output, hospital admissions rates by discharge diagnosis (cardiovascular, respiratory, or either), and seven demographic variables. The investigators explored principal components analysis and various averaging schema. They used a linear mixed effects regression model to generate regression coefficients to calculate predicted admission rates and used geospatial techniques to visualize potential "hot spots."

Despite generating the best emissions and meteorological data that could be produced for the study time period, the model-predicted pollutant concentrations proved not to be

adequate for the objectives of this health effects study. Thus, the investigators did not provide results relating model-predicted exposure to health endpoints. Instead, the report focuses on methodology and delineates areas that need further development.

The comments of the NUATRC Scientific Advisory Panel below highlights some of the strengths and limitations of this research so that, once refined, this approach might be used for areas of the country with significant pollutant gradients, such as Houston, TX with localized areas of high pollution or "hot spots."

COMMENTS OF THE NUATRC SCIENTIFIC ADVISORY PANEL

The authors are to be commended for taking on the challenge of applying a large-scale Eulerian air quality model (CMAQ) to the problem of variable health effects incidence in a complex urban area (Houston, TX).

Strengths:

- The development of hospital admissions data as a measure of variable health effects was a particular strong point of the study.
- The investigators provide a thorough explanation for how these data were generated and describe the issues and problems of organizing these data in a system for spatial and temporal analyses.

Limitations:

The conclusions drawn from the study are fraught with methodological and interpretation challenges as they relate to evaluation of the model to develop the needed exposure estimates. The authors repeat in several places that their chief numerical conclusions are not included in the report because of "concerns" about the exposure estimates computed from the CMAQ output and the instability of the resultant risk estimates. However, the authors have not adequately addressed these issues, leaving the reader to evaluate its conclusions and interpretation based only on the summary of methods and application.

Specifically,

- The modeling system (formulation and inputs) was not evaluated for the use intended in this application: predicting the atmospheric concentrations needed to analyze the relationship between exposure and hospital admissions. The authors attempt to use a long-term average (90-day) concentration of air pollutants as the estimate for chronic exposures that could trigger hospital admissions. The authors acknowledge that taking such an average in each 4x4 cell in the modeled domain for Harris County enfolds substantial error in the model predictions for reasons like the well-known and characterized under-performance of CMAQ for ozone at night. But the more useful response by the authors would have been to obtain an appropriate exposure time from the model, by looking throughout the domain at finer spatial scales for patterns and trends which would then be traceable at that scale to the model-predicted pollutant concentrations.
- The authors claim, without attribution to any literature, that lags of up to 7 days are important in air pollution health effects studies (the usual lag structures are 0-3 days), but do not explain why a 90-day mean is significant for their proposed 7-day lag.
- Using a 90-day average, the authors are attempting to predict hospital admissions by comparing model estimates to only four ambient VOC measurement sites and two $PM_{2.5}$ measurements in the entire domain. Additionally, the predictions versus the measurements are compared with observations of only one of the hazardous air pollutants, benzene.
- Overall, the 90-day averages were biased sufficiently that the model output could not reliably estimate spatial locations of "hot spots." This pilot study did not provide an adequate test of the CMAQ model for predicting atmospheric concentrations with the characteristics needed for studying relationships between air toxics exposures and health effect differences spatially across a modeled domain.

Conclusions:

Despite its limitations, this work makes useful contributions to the field:

- It demonstrates the promise and pitfalls of using CMAQ in an epidemiological study, and provides valuable lessons for other researchers interested in this approach. CMAQ is a complex model, and atmospheric chemistry - particularly in Houston - is extremely complicated. There are problems with bringing CMAQ down to a 4-km grid; other research teams need to be cautious about taking larger grid models down to a finer scale.
- The study showed that investigators interested in this type of work need to consider the following issues related to the use of models:

1. The most important environmental parameter(s) the model must predict and the consequences if the model is not adequate in predicting this parameter;
2. The requirements for operating the model;
3. The source and quality of inputs;
4. The formulation and completeness of model components;
5. The personnel, computational, and storage demands of the model;
6. The duration of total model effort;
7. The identification of criteria of performance that modeling system needs to meet to be adequate for the work; and
8. The establishment of fitness of the air quality management system's output for use in the effort (e.g. health effects assessment).

The investigators showed that using this model in an epidemiological application is expensive and time consuming, and leads to relatively uncertain results. Furthermore, this study suggests that for this type of research to succeed, there must be detailed discussions between the modeler, who is responsible for producing estimates of airborne concentrations and exposures, and epidemiologists to ensure that the accuracy of the estimates will be consistent with and sufficient for the intended objective of the study.

In conclusion, this pilot study has not produced reliable results concerning the relationship between pollution hot spots and health effects. Though the report follows, we caution the reader that interpretation of results concerning the hot spots and health effects relationship is unwarranted.

A Pilot Study using EPA's CMAQ Model and Hospital Admission Data to Identify Multipollutant "Hot Spots" of Concern in Harris County, Texas

A Final Report to the

Mickey Leland National Urban Air Toxics Research Center

Winifred J. Hamilton[1], Daewon W. Byun[2,3], Wenyaw Chan[4], Jason K.S. Ching[5], Younghun Han[4,6], Ricardo A. Lopez[1,7], Violeta F. Coarfa[2] , and DaeGyun Lee[2]

[1] Baylor College of Medicine
[2] University of Houston
[3] National Oceanic and Atmospheric Administration
[4] The University of Texas Health Science Center - Houston
[5] U.S. Environmental Protection Agency
[6] The University of Texas M.D. Anderson Cancer Center
[7] IRC Risk and Safety LLC

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

ABSTRACT

Residents in certain geographical locations, such as near freeways or industrial facilities, are likely to be exposed to heavier loads of multiple air pollutants than those farther from such sources. Such "hot spots" are poorly delineated by monitor-based studies, and few attempts have been made to examine potential associations between modeled output and actual health effects. In this pilot geospatial study we used simulated hourly values from the MM5 and Community Multiscale Air Quality (CMAQ) models for a 90-day period in 2000 for 25 meteorological and air pollution variables, including 16 air toxics, for 337 4 x 4-km cells in Harris County, Texas. Each cell was characterized by the simulated output; hospital admission rates by discharge diagnosis (cardiovascular, respiratory, or either); and seven demographic variables. Principal components analysis and various averaging schema were explored. A linear mixed-effects regression model was used to generate regression coefficients from which to create predictive "hot spot" maps. Although preliminary findings suggested spatial differences in predicted hospitalization rates, the pilot study raised significant concerns about the ability of the current model to usefully estimate exposure and other methodological issues that need to be addressed. The methodology, problems encountered, and potential next steps are discussed.

INTRODUCTION

OVERVIEW

A number of monitoring and/or exposure assessment studies have documented that some neighborhoods are regularly exposed to higher levels of multiple air pollutants than are other neighborhoods. In many instances, other environmental and demographic stressors such as contaminated water, lead-based paint, poverty, and poor nutrition may add to the vulnerability of residents in these neighborhoods to health problems associated with poor air quality. This awareness led to, in 1994, Executive Order 12898 that, among other provisions, required (1) achieving environmental justice to be part of the mission of all federal agencies and (2) "identifying and addressing, as appropriate, disproportionately high and adverse human health or environmental effects . . . [in] minority and low-income populations" (U.S. Office of the President, 1994).

A number of challenges are faced by such "hot spot" efforts including lack of geographical coverage in monitor-based studies, concern about the use of monitored or modeled data as surrogates for personal exposure, the problem of establishing guidelines for individual pollutants when exposure and health effects exist in a multipollutant environment, and numerous statistical challenges including ecologic bias, uncertainty, and collinearity, especially in multipollutant spatial models. These challenges are compounded when actual health effects, such as hospital admissions for cardiovascular or other diseases, are included in the study design.

In this pilot study, we examined the use of simulated concentrations of multiple air pollutants of particular concern in the Houston area, including five criteria air pollutants (CAPs) and 16 hazardous air pollutants (HAPs), as well as two meteorological variables (temperature and relative humidity), to assess the potential effects of spatial differences in the level of chronic exposure to these variables on hospital admissions for cardiovascular and respiratory causes, controlling for seven demographic variables. A multivariate linear mixed-effects model (LMM) was used to generate regression coefficients that, although unstable for assessing the contribution of individual meteorological and pollutant variables due to correlation between the multiple pollutants in the final models, were used to calculate predicted rates of admission for respiratory and cardiovascular disease, adjusted for significant meteorological, demographic, and pollutant variables. These predicted rates were then used to create "hot spot" maps identifying areas of potential concern. For this pilot study, we used a 4 x 4-km grid that resulted in 337 areas for analysis, and we used MM5 and several adaptations of the U.S. EPA's three-dimensional photochemical air quality model, the Community Multi-Scale Air Quality (CMAQ) model, to simulate the meteorological and pollutant values.

Although we are encouraged by numerous aspects of this pilot study, the effort also generated a number of significant concerns and challenges that must be addressed in future work. For example, are the current inventories sufficient for this level of resolution? What is the appropriate averaging time for the meteorological and pollutant variables, especially in a multipollutant model in which most of the pollutants have significantly different daily temporal profiles? Is principal components analysis (PCA) useful in reducing the number of variables and problems with collinearity in the models? Is CMAQ, although developed as a multipollutant model but to some degree optimized to predict ozone (O_3) formation for regulatory decision-making, capable of simulating human exposure with sufficient accuracy to be useful in such a health-based

model? Can the differing levels of uncertainty in a grid-based model with significant population gradients be adequately addressed statistically?

These questions and various difficulties encountered during this pilot study have led us to focus in this report primarily on the methods used, problems encountered and potential next steps, choosing not to include the results of the regression models except in a general discussion of the chronology of the steps we took to develop maps of predicted rates of hospitalization based on simulated exposure to ambient air pollution and various demographic factors that influence vulnerability. In addition to the methodology, including data collection and preparation for the analyses, this report provides an overview of the use of models in air pollution and health effects studies, as well as a number of approaches that might be used to improve subsequent efforts.

URBAN POLLUTION AND HEALTH EFFECTS STUDIES

Time-Series and Case-Crossover Studies

Many air pollution and health effects studies done in urban areas have used a retrospective time-series or case-crossover design. Most of these studies have focused on CAPs, particularly O_3 and particulate matter (PM), although nitrogen oxides (NO_x) and carbon monoxide (CO) are included in some of the study designs, some of which use innovative statistical approaches to reduce the effect of collinearity introduced by additional pollutants on the effect estimates.

For measuring exposure, retrospective time-series or case-crossover studies generally use a single daily mean concentration of the pollutants of interest that is obtained from a centrally placed monitor or calculated from a number of fixed-site monitors in the community (Alberdi Odriozola et al., 1998; Barnett et al., 2006; Burnett et al., 1999; Dominici et al., 2003; Dominici et al., 2004; Ko et al., 2007; Lee et al., 2007; Lee et al., 2006; Lin et al., 2003; Magas et al., 2007; Medina-Ramon et al., 2006; Morgan et al., 1998; Poloniecki et al., 1997; Saez et al., 1999; Wellenius et al., 2006; Yang and Chen, 2007; Yang et al., 2004; Zanobetti et al., 2003). Hospital admissions, emergency room visits, or death are the most commonly used health endpoints, in part because of the availability of relatively large datasets that are needed to assess day-to-day variability in exposure and response. These studies effectively use each "case" as its own control, largely eliminating personal confounders such as smoking and education. For this reason, such studies are being increasingly refined and are particularly useful as a measure of acute effects although some delayed effects can also be captured by integrating longer lag times into the design (Martins et al., 2006; Neuberger et al., 2007; Zanobetti and Schwartz, 2008). However, using such a global measure of exposure largely ignores the likelihood of significant spatial variations in pollutant levels and the disproportionate burden of such "hot spots" of exposure and/or susceptibility within urban areas. This is especially important for HAPs, also referred to as air toxics, which tend to be localized.

Zanobetti, Bell, Dominici and others have more recently conducted comparisons between cities using time-series methodology that attempted to look at community characteristics (Bell and Dominici, 2008; Jerrett et al., 2007; Zanobetti et al., 2000), responding in part to questions about environmental justice issues. Bell and Dominici's study published in 2008, for example, examined effect modification by community characteristics on the short-term effects of ozone on mortality in 98 U.S. communities. For exposure, the investigators used daily average weather data for each community from the National Climatic Data Center and concentrations of O_3 and PM from the U.S. Environmental Protection Agency Aerometric Information Retrieval Service (EPA AIRS). Data from multiple monitors within each community were averaged, using a 10% trim to avoid the influence of outliers, for a single mean value for each city (Bell and Dominici, 2008). Bell and Dominici found that community-level characteristics modified the relation between O_3 and mortality, with higher effect estimates associated with higher unemployment, being Black, increased use of public transportation, and a lower percentage of households with central air conditioning. However, as the authors note, this study design does not address intra-community "hot spots," which is where most environmental justice inequities exist-at a local level.

Prospective Cohort Studies

Several major prospective studies, including the six-city (Dockery et al., 1993; Krewski et al., 2003; Laden et al., 2006) and American Cancer Society (Krewski et al., 2003; Pope et al., 2002; Pope et al., 2004) studies, have examined intercity differences and are able to control for various confounders, such as education, smoking, and body-mass index. Because of the potential importance of these studies for establishing National Ambient Air Quality Standards (NAAQS) for PM, the data collected for these two studies were subsequently re-analyzed by another team of researchers (Health Effects Institute, 2000). The re-analysis confirmed the findings of the original studies, i.e., intercity differences in mortality correlate with different levels of pollution. Both studies focused primarily on particulate

matter < 2.5 microns in diameter ($PM_{2.5}$) as the primary pollutant of interest. These studies, however, also do not address intra-city "hot spots," relying in the first study primarily on a centrally located fixed-site monitor and in the second on the mean of all monitors for each city.

Spatial Studies

Findings from the previously discussed study designs have been reproduced in other studies and, collectively, have played key roles in the regulatory process, often supporting a tightening of the NAAQS to better protect public health. Nevertheless, none of them provides information that directly enables researchers, clinicians, or policy makers to identify particular geographical areas of disproportionate exposure and health effects within urban areas for which, perhaps, specific local interventions might make a measurable difference. In addition, these studies seldom include any HAPs. The reasons for this include limited data, low concentrations that increase the likelihood of measurement error, the lack of health-based standards to drive research, and the fact that ambient HAP concentrations tend to be more localized than do CAP concentrations and are therefore less amenable to city-wide averaging.

Except for occasional limited investigations into disease clusters, few spatial exposure-disease analyses within urban areas exist, in part because of the difficulty in generalizing from group data (Greenland, 2001). However, a growing awareness that certain segments of many urban populations are disproportionately exposed to environmental stressors (American Lung Association, 2001; Bullard, 1983; Maantay, 2007; Morello-Frosch et al., 2002; Woodruff et al., 2003; Zanobetti and Schwartz, 2000) and improvements in geospatial and statistical techniques have recently led to increased research in this arena (Briggs, 2005; Chen et al., 2007; Dolinoy and Miranda, 2004; Jerrett et al., 2005; Liao et al., 2006; Maheswaran et al., 2006; Nuckols et al., 2004; Scoggins et al., 2004; Yanosky et al., 2008; Zandbergen and Chakraborty, 2006; Zhang, 2006; Zhou and Levy, 2007). In their spatial study of intra-city air pollution and mortality in Los Angeles using available monitoring data and various spatial interpolation techniques, Jerrett and associates, for example, observed within-city health effects to be nearly three times greater than suggested by models that compared metropolitan areas and relied on community averages as the exposure estimate (Jerrett et al., 2005). Again, this suggests that more localized exposure assessments are needed.

One immediate problem confronting spatial studies in most urban areas is a lack of exposure data, as fixed-site monitors seldom provide sufficient geographical coverage. Mathematical techniques can be used to expand the usefulness of data from fixed site monitors (Andria et al., 2008; Brown et al., 1994). For example, application of the Kalman filter to measured concentrations of a single pollutant at a fixed-site monitor can improve the quality of the data for exposure estimation by reducing the contribution of noise and reconstructing missing data points. Kriging, often in combination with the Kalman filter, is a geostatistical interpolation technique that can be used to predict concentrations of a single pollutant across a limited geographical territory based on measurements obtained from a set of fixed monitors. The spatial estimations produced by the kriging technique can further be adjusted by a number of factors, such as meteorological conditions, land-use topology, and even characteristics of the emission sources (e.g., industrial or vehicular), but are nevertheless dependent on sufficient and well placed monitors, each of which measures a meaningful set of pollutants, to produce useful interpolated concentrations for estimating exposure. For financial and other reasons, this is seldom the case.

More comprehensive exposure assessment, such as personal monitoring (Payne-Sturges et al., 2004) or community-based saturation monitoring (Zhu et al., 2008), is generally done at a small scale and is usually not easily generalizable to other communities, although the growing body of such information is particularly valuable for future efforts to refine hybrid approaches to better estimate local-scale exposure. Currently many community efforts that attempt to define potential "hot spots" utilize industry-reported levels of emissions, such as are provided by the Toxic Release Inventory (TRI), or composite inventories of emissions prepared by state or national agencies, such as the Texas Emissions Inventory (TEI) or National Emissions Inventory (NEI). These latter inventories include, with varying degrees of resolution, industrial, mobile, and area emissions, and are regularly updated. They are often used in models to better understand CAP emissions and pollutant transport, to develop pollution-reduction strategies for areas in nonattainment for one or more of the CAPs, and to help define areas of potential concern near major emission sources.

One spatial design-the proximity analysis-compares some measure of health effects, such as hospital admissions or respiratory symptoms, in areas near and distant to a known pollution source, such as a major highway or large industrial emitter (Blumenstock et al., 2000; Maheswaran and Elliott, 2003; Perlin et al., 2001; Sheppard et al., 1999). Another design compares health effects in defined

geographical areas, such as census tracts, with some measure of exposure, such as nearest monitor or number of sources (Buckeridge et al., 2002; Elliott et al., 2000). These designs, however, tend to be limited to a single pollutant or source category and therefore are unlikely to well represent health risk, although they can be very useful for initially identifying key exposure or health disparities, which can then be targeted for additional scrutiny.

In addition and as noted earlier, most studies have focused on CAPs, especially O_3, $PM_{2.5}$, CO, NO_2, and sulfur dioxide (SO_2), for which there are NAAQS. However, air toxics are also associated with health effects, including acute effects such as eye and skin irritation, nausea, headache, and difficulty in breathing, (Brunekreef and Holgate, 2002; Burnett et al., 1999; Leikauf, 2002; Moolgavkar, 2000; Morello-Frosch et al., 2000; Morgan et al., 1998; Morris, 2001; South Coast Air Quality Management District, 1999; Weisel, 2002), as well as long-term effects, including cardiovascular damage, respiratory scarring, immune dysfunction, asthma, cancer, and Parkinson's disease (Herbert et al., 2006; Jacquez and Greiling, 2003; Morello-Frosch et al., 2000; South Coast Air Quality Management District, 1999; Woodruff et al., 1998).

The inclusion of a larger number of pollutants, including CAPs and HAPs, raises special challenges in epidemiological models in part because of correlation between many pollutants, which can create problems in interpreting results. At the same time exposure to environmental hazards and vulnerability to health effects are complexly multifactorial and most current study designs are limited in their ability to address this complexity. It is interesting to note, for example, that the color-coded U.S. Air Quality Index (AQI), which is used to issue health-based advisories based on modeled air quality forecasts and/or measured concentrations, is capable of only addressing a single NAAQS pollutant at a time. Although there has been discussion of a multipollutant U.S. AQI, this has not been developed, in part because of the complexity of multipollutant exposure. In Canada, government researchers have recently developed a multipollutant air quality health index in response to this very concern, that our current methods of assessing air quality for health advisories do not capture additive effects of multiple pollutants or reflect the apparent no-threshold concentration-response relationship between air pollution and health (Stieb et al., 2008). Noting the uncertainty of how best to reflect the mix of pollutants, these researchers currently recommend, based on extensive sensitivity analyses and mortality data, use of a 10-point scale based on continuous trailing 3-hour concentrations of three pollutants: NO_2, O_3 and $PM_{2.5}$. Some of the findings and challenges of multipollutant health-effects research are addressed in the Discussion of this report.

AIR QUALITY MODELING

To address some of the shortcomings of monitor-based and proximity studies, increasingly complex models are being developed to estimate exposure using, variously and often in combination, meteorological, emissions, chemical reactivity, transport, geographical, time-activity, and other data. Such models have the advantage of estimating pollutant concentrations in areas where there are no monitors and of estimating concentrations of pollutants that are not being measured by existing monitors or are measured poorly by existing instrumentation. The complexity of these evolving simulation models, however, may lead to erroneous simulated concentrations. These models, for example, are dependent not only on the quality of emissions data and meteorological input but also, in the case of photochemical models, on the ability of the chemical algorithms to accurately estimate atmospheric chemistry. Nevertheless, a growing sophistication in air quality modeling, which is now increasingly used for source identification and for developing effective pollution reduction strategies, coupled with a growing awareness of the limitations of many of the current epidemiologic study designs especially with regard to "hot spot" analysis, has led to recent efforts to extend pollutant modeling to estimate exposure and, increasingly, to better understand the potential role of air pollutants in causing or exacerbating adverse health effects. Our pilot study uses one of these models, the U.S. EPA's CMAQ model.

Until relatively recently, air quality models typically addressed individual pollutants separately, primarily to address regulatory issues. For example, a dispersion model might be used to estimate the spatial distribution of and population exposure to benzene from a large industrial source under various meteorological conditions and to determine whether additional permits might be allocated to expand the operation. However, pollutants do not exist in isolation and control strategies that address one set of problems may aggravate other related pollutant issues. Similarly, populations are exposed to multipollutants, some of which are synergistic, and different populations have different vulnerabilities to adverse effects associated with exposure.

The growing appreciation of the complexity of air quality and human exposure has encouraged the development of more sophisticated modeling systems. For example,

pollutants in the atmosphere are subject to numerous transport processes and transformation pathways that control their composition and levels. Also, pollutant concentration fields are sensitive to the type and history of the atmospheric mixtures of different chemical compounds. Thus, modeled abatement strategies of pollutant precursors, such as volatile organic compounds (VOCs) and NO_x, to reduce O_3 levels may under a variety of conditions cause an exacerbation of other air pollutants, such as PM, or create problems with acidic deposition. In addition, although early attention in the U.S. tended to focus on O_3, NO_2, CO, SO_2, and PM, more recent efforts have begun to recognize the 187 defined HAPs, diesel exhaust and other species, as their roles in CAP formation and health effects are better appreciated.

An overview of the development of comprehensive air quality models is beyond the scope of this report, as is more than a brief description of several models currently being used, directly or indirectly, in exposure and/or health effects studies. The U.S. EPA web site offers an overview of and links to most of the air quality models in use today, including dispersion, photochemical and receptor models (U.S. Environmental Protection Agency, 2008). In general, air quality modeling techniques can be divided into three categories:

- Gaussian dispersion modeling, in which the distribution of nonreactive species are described with a Gaussian plume dispersion equation or one of its variations using a steady state assumption. AERMOD and ASPEN, briefly described in the next section, are examples of Gaussian dispersion models.

- Lagrangian modeling, in which movements of air parcels are followed to simulate mixing and chemical transformations. HYSPLIT, briefly described in the next section, is an example of a Lagrangian model.

- Eulerian modeling, in which a fixed three-dimensional grid system is used to represent comprehensive atmospheric processes, such as transport, emissions, physical and chemical transformations, and deposition. CAMx and CMAQ, briefly described in the next section, are examples of Eulerian models.

As noted earlier, air quality models are among the primary tools used to evaluate the impacts from emissions changes and therefore play a major role in the development of regulatory policy. Increasingly, as the models become more robust and accurate, the output is also being used as a measure of human exposure or of health benefits associated with emissions reductions. In addition to models that simulate pollutant concentration fields, there are a number of exposure models that can help refine these fields to better estimate human exposure. Among such models are the Hazardous Air Pollutant Exposure Model (HAPEM), Air Pollutant Exposure Model (APEX), and Stochastic Human Exposure and Dose Simulation (SHEDS) model. These models use simulated ambient air pollutant concentrations as base input and then modify that input with added information on activity patterns, indoor/outdoor exchanges, toxicity factors, and microenvironments. Paramount to the use of these exposure models, however, is the need for high-quality meteorological data and pollutant concentrations for baseline input. Brief descriptions of several air quality simulation models that have been applied to the Houston region, including CMAQ, follow.

AERMOD

The American Meteorological Society/Environmental Protection Agency Regulatory Model Improvement Committee Model (AERMOD) atmospheric dispersion modeling system (U.S. Environmental Protection Agency, 2008) is an integrated model that includes three modules: (1) a steady-state model designed to compute ground-level air pollutant concentrations for short-range (up to 50 kilometers) dispersion of air pollutant emissions from stationary industrial sources; (2) a meteorological data preprocessor (AERMET) that accepts surface meteorological data, upper air soundings, and optionally, data from on-site instrument towers; and (3) a terrain preprocessor (AERMAP) whose main purpose is to provide a physical relationship between terrain features and the behavior of air pollution plumes. The AERMOD system, which is particularly well-suited for neighborhood-level assessments, has recently been used in conjunction with CMAQ in hybrid models-including a study in Houston-to improve local-scale exposure estimates (Isakov et al., 2007; Stein et al., 2007).

ASPEN

The Assessment System for Population Exposure Nationwide (ASPEN) model consists of a dispersion and a mapping module. The dispersion module is a Gaussian formulation, based on the Industrial Source Complex Short Term 3 (ISCST3) dispersion model, for estimating ambient annual average concentrations at a set of fixed receptors within the vicinity of an emission source. The mapping

module produces a concentration for each census tract. Input data needed are emissions data, meteorological data and census tract data. The Emissions Modeling System for Hazardous Pollutants (EMS-HAP) is used to process the emissions inputs. The U.S. EPA used ASPEN and the National Emissions Inventory (NEI) final version 3 to simulate annual 1999 concentrations of 177 air pollutants (a subset of the 187 HAPs plus diesel PM) in U.S. census tracts for the most recent National-Scale Air Toxic Assessment (NATA) (U.S. Environmental Protection Agency, 2007), the results of which were released in 2006. In the Houston area, Whitworth and associates recently used the NATA ASPEN output for benzene and 1,3-butadiene to examine the spatial distribution of childhood lymphohematopoetic cancer, generally finding higher levels of cancer in census tracts with higher levels of thee pollutants (Whitworth et al., 2008). The NATA exposure and risk assessment also used an exposure model, HAPEM5, to improve the exposure metric. As part of the performance evaluation of the simulations used in our study, we compared ASPEN and our output for selected HAPs in Harris County.

HYSPLIT

The Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model computes simple air parcel trajectories to complex dispersion and deposition simulations. The dispersion of a pollutant is calculated by assuming it to be either in a puff or particle state. HYSPLIT has been used in various applications including apportionment of species such as PM and mercury to their source locations (Martello et al., 2008; Sunderland et al., 2008). Stein and associates have tested the feasibility of a hybrid model approach that uses CMAQ, HYSPLIT, and/or AERMOD to simulate with high spatial resolution benzene concentrations in Houston for August 18 to September 4, 2000 (Stein et al., 2007). In this simulation, HYSPLIT was used to model concentration variability from different sources or pathways. The HYSPLIT concentrations, along with the more spatially resolved concentrations from AERMOD, were then added to the CMAQ-calculated background concentration to estimate the total mean benzene concentration.

CAMx

The Comprehensive Air Quality Model with extensions (CAMx) model is an Eulerian photochemical dispersion model that allows for integrated "one-atmosphere" assessments of gaseous and particulate air pollution over scales ranging from sub-urban to continental. CAMx is maintained and distributed by ENVIRON. The Texas Commission on Environmental Quality (TCEQ) has used CAMx for modeling its base-case ozone scenarios to develop its State Implementation Plans (SIP) to reduce ambient ozone concentrations in nonattainment areas such as Houston. More recently, the University of Houston's Institute for Multi-dimensional Air Quality Studies (UH-IMAQS) and others have run parallel CMAQ and CAMx simulations to better understand the strengths and weaknesses of each model. The primary difference between the two models is different handling of plumes, which causes different vertical distributions of emissions. Chang and Allen used CAMx to study the role of chlorine radicals on ozone formation (Chang and Allen, 2006) in East Texas, and the South Coast Air Quality Management District in California recently used CAMx to estimate the health benefits of decreased $PM_{2.5}$ levels associated with proposed marine vessel control measures (South Coast Air Quality Management District, 2007).

CMAQ

The Community Multiscale Air Quality (CMAQ) model is an Eulerian multipollutant, multiscale air quality modeling system that simulates various photochemical and physical processes that are thought to be important for understanding atmospheric trace gas transformations and distributions. The CMAQ system uses a three-dimensional nested grid system. For our investigation, the CMAQ modeling used nested domains with grid sizes of 36, 12, 4, and 1 km, and 23 vertical layers. For our health-based model we used output from the 4-km grid and layer 1, the ground-level layer. The CMAQ modeling system contains three types of modeling components: (1) a meteorological modeling system (e.g., MM5) for the description of atmospheric states and motions, (2) emissions inventories (e.g., NEI) and software (e.g., SMOKE) for apportionment of anthropogenic and natural emissions into the grid, and (3) a chemistry-transport modeling system (e.g., SAPRC) for simulation of air pollutant concentrations. The CMAQ model was developed through a partnership between U.S. EPA and the National Oceanic and Atmospheric Agency (NOAA), and is available from the Community Modeling & Analysis System (CMAS; www.cmascenter.org). It is designed as a holistic modeling tool for handling all the major pollutant issues, including photochemical oxidants, particulate matter, and acid and nutrient deposition. As such, it lends itself to modifications, for example, to explicitly simulate HAPs of particular interest. A detailed

explanation of the CMAQ system is beyond the scope of this report; for additional information, visit www.epa.gov/asmdnerl/CMAQ or www.cmaq-model.org. The physical options and modifications to CMAQ used for the simulations in our study are discussed in the Methods section. Briefly, the version of CMAQ used in our study, CMAQ4.4 (SAPRC99-ARO), uses a modification of the SAPRC99 chemical mechanism to explicitly simulate a number of aromatic and other HAP species. We use the term CMAQ4.4 to refer to this modification.

CMAQ has been recently used to examine changes in O_3 concentrations based on changes in NO_x emissions from power companies in the Eastern U.S. (Gego et al., 2008), and has been used in a number of studies to examine human exposure to environmental pollutants, often including risk assessment or assessment of dose across multiple environments and activity patterns (Furtaw, 2001; Georgopoulos et al., 2005; Isakov et al., 2007; Marshall et al., 2008; Sanhueza et al., 2003; Sokhi et al., 2006). In the Houston-Galveston area (HGA), Ching and associates used CMAQ to investigate subgrid variability by simulating key photochemical species, including CO, O_3, NO_x, and acetaldehyde, at four grid sizes (36, 12, 4, and 1 km) for August 30, 2000 (Ching et al., 2006).

CMAQ-HAP

The CMAQ-HAP model is a simplified version of our SAPRC99-ARO-modifed CMAQ4.4 model. CMAQ-HAP, which was developed by Dr. Violeta F. Coarfa while with the UH-IMAQS team, is an engineering model that utilizes oxidant fields produced by the CMAQ4.4 model to generate additional HAP species, greatly reducing computational time for these HAPs. The CMAQ-HAP model, for example, performs a one-day simulation in 1 hour and 40 minutes, 8-9 times faster than the CMAQ4.4 model.

Using CMAQ to study HAPs explicitly is a relatively new application of the CMAQ system. Although Eulerian air quality modeling has been used to study ozone nonattainment problems in the HGA for many years, it has not generally been used for air toxics modeling. Most of the air toxics modeling for the region has been performed using a Gaussian dispersion model, such as the Gaussian-Plume Multiple Source Air Quality Algorithm (RAM) (Radian Corporation, 1995; U.S. Environmental Protection Agency, 1987) or the ISCST model (U.S. Environmental Protection Agency, 1995; U.S. Environmental Protection Agency, 2002). The EPA applied the ISCST model to Houston for the model year 1996 as a demonstration project, "Example Application of Modeling Toxic Air Pollutants in Urban Areas" (U.S. Environmental Protection Agency, 2002). As noted earlier, however, dispersion modeling has a number of limitations despite good local-scale resolution for many species. It does not, for example, account for wind shear, track plumes beyond about 50 km, handle chemistry correctly, or include biogenic sources. For many HAPs, ignoring photochemical production is of particular concern as the secondary creation of HAPs in the atmosphere can account for a significant portion of their ambient concentrations.

Luecken and associates simulated concentrations for five HAPs-formaldehyde, acetaldehyde, benzene, 1,3-butadiene, and acrolein-across the continental U.S. for year 2001 at a resolution of 36 km using an adaptation of CMAQ (Luecken et al., 2006). Seigneur and co-workers modified CMAQ to perform regional modeling of the atmospheric fate and transport of benzene and diesel particles (Seigneur et al., 2002) and have more recently reviewed the status of air toxics modeling, emphasizing the potential of Eulerian models (Seigneur, 2005). Ching and associates have created an air toxics version of CMAQ by incorporating modifications to version 4 of the Carbon Bond (CB4) chemical mechanism. This version of CMAQ, called CMAQ-AT, has been used to simulate concentrations of 1,3-butadiene, formaldehyde, acetaldehyde, and benzene at 4-km resolution in central Philadelphia (Ching et al., 2004). In addition, Ching and associates used the CMAQ-AT output to calculate average exposure levels using HAPEM, which they compared with NATA estimates for Philadelphia.

Comparison of Models for Health Studies

The use of simulated air pollutant concentrations from an air quality model for health effects studies requires careful attention to the nature of the pollutants studied (e.g., are secondary photochemical pollutants important? Is transport?), the geographical resolution, the probable roles of topography and meteorology, the time scale, and the health endpoints (e.g., out-of-hospital cardiac arrest or cancer). Each model has various strengths and weaknesses. In addition, depending on the goals of the study, study design, and availability of data, it may be appropriate to consider using several models (a hybrid approach) or using a combination of modeled and measured pollutant concentrations to better approximate exposure. Among models, the CMAQ system is particularly well designed for secondary photochemically produced pollutants and biogenic emissions, neither of which is included in dispersion or Lagrangian models. CMAQ is also one of the few systems that can accommodate CAP and HAP

emissions in a "one atmosphere" system. The current maximum resolution capability of the CMAQ system at ground level is around 1 km^2, but more detailed emissions input, chemistry, terrain information, and transport mechanisms are needed to improve simulated output at this resolution. Increased computational efficiency and infrastructure will also be needed to reasonably run the CMAQ model at this resolution and for longer study periods for future local-scale health-based studies.

Although Lagrangian models handle transport efficiently and with high resolution, the Lagrangian model is essentially a one-pollutant model, as are most diffusion models. In addition, Lagrangian and dispersion models generally do not include secondary pollutants (or do so only to a limited degree) and so are best suited for relatively inert species, such as benzene, PM, and some metals. On the other hand, local-scale dispersion modeling can provide improved spatial characterization of mobile and industrial emissions unattainable with CMAQ. For this reason, several researchers have recently proposed the use of a hybrid modeling approach that uses, for example, input from CMAQ to model the background concentrations augmented with concentrations from local-scale sources generated by ASPEN, AERMOD, and/or HYSPLIT (Cook et al., 2008; Isakov et al., 2007; Stein et al., 2007). Others are working to develop methods for weighting concentrations simulated by CMAQ or other models with time-activity, commuting, and other exposure factors; personal monitoring; population density; and observed ambient concentrations, as appropriate (Georgopoulos et al., 2005; Ozkaynak et al., 2009). Efforts to combine modeled and measured pollutant output utilizing space-time modeling are especially promising (McMillan et al., in press). These modifications and improved computational capabilities are also critical for improving simulated air quality output for regulatory decision-making, as well as for linked health effects research (Isakov et al., 2007; Riccio et al., 2006; Xie and Berkowitz, 2007).

In the preliminary work presented here, we use a combination of simulated output from MM5 and two CMAQ versions with modified chemistry to better represent selected HAPs, along with the CAPs. Although local-scale "hybrid" approaches offer a means for improving exposure estimates, they were unavailable to be implemented during this phase of our study and generally beyond the scope of our pilot study. It will be of interest in the future to re-examine the results of the current study with outputs from approaches that provide additional spatial detail and/or weighting from measured concentrations.

HOT SPOTS

The phrase "hot spot" is used in the environmental arena to mean slightly different things. Most commonly it is used either to (1) define a relatively small geographical area in which an individual pollutant exceeds some regulatory guideline or threshold, or (2) define a neighborhood where one or sometimes several pollutant concentrations are elevated relative to some guideline or levels elsewhere such that there is reasonable expectation that these elevated concentrations present an increased risk of adverse health effects to those who reside or spend significant time in the area or neighborhood.

The U.S. EPA and several other federal agencies have programs that designate geographical areas of concern for exposing residents to elevated levels of pollutants that are likely to be associated with adverse health effects, especially among susceptible populations. For individual CAPs, the U.S. EPA has health-based standards and designates as nonattainment areas those areas that do not meet these standards, also ranking these areas by severity. Although localized "hot spots" can and do exist for CAPS, by design for regulatory compliance and because of the nature of the criteria pollutants-which tend to be urban-based, present in relatively large quantities, and often travel significant distances-nonattainment areas generally encompass metropolitan areas. In the HGA the nonattainment area for O_3 consists of eight counties. Within the 1990 amended federal Clean Air Act (CAA), the only specific reference to pollution "hot spots" is under transportation conformity, which requires "hot-spot" analyses for many new transportation projects to assess air quality impacts on a smaller scale, such as at a congested roadway intersection, than an entire nonattainment area. Transportation conformity "hot-spot" analyses use a dispersion model and focus on future localized CO or PM concentrations that would likely increase if the transportation project were implemented. The Agency for Toxic Substances and Disease Registry (ATSDR) Superfund program is another example of a "hot-spot" designation, and is somewhat distinctive in that most Superfund sites address multiple pollutants as well as multiple media through which exposure can occur.

In general, however, the term "hot spot" has been used to describe relatively small geographical areas where residents are exposed to elevated levels of one or more of the 187 designated HAPs plus diesel exhaust, in part because of their toxicity at low levels and propensity to remain relatively localized. By definition, HAPs are "those pollutants that are known or suspected to cause cancer or

other serious health effects, such as reproductive effects or birth defects" (U.S. Environmental Protection Agency, 2008). The U.S. EPA's NATA program ranks census tracks in the U.S. by public health risk to selected HAPs. In 2006 it released its most recent assessment, which addressed cancer and noncancer risk to 133 pollutants (U.S. Environmental Protection Agency, 2006). NATA uses the NEI, a dispersion model, an exposure model, and species-specific toxicologic factors. Although census tracks are often larger than many "hot spot" designations and NATA does not use the phrase per se, the elements and intent of the program are very similar to, for example, the California "hot spot" program.

The California Air Resources Board (CARB) defines an air pollution "hot spot" as a "location where emissions from specific sources may expose individuals and population groups to elevated risks of adverse health effects-including but not limited to cancer-and contribute to the cumulative health risks of emissions from other sources in the area" (California Air Resources Board, 2006). The goals of California's AB 2588 Air Toxics "Hot Spots" Program are to collect emission data, identify facilities having localized impacts, ascertain health risks, notify nearby residents of significant risks, and reduce those significant risks to acceptable levels. The program uses cancer potency factors and acute and chronic noncancer Reference Exposure Levels (RELs) approved by California's Office of Environmental Health Hazard Assessment and CARB to assess health risk in areas of suspected impact.

In Texas, the TCEQ maintains an Air Pollutant Watch List (APWL) (Texas Commission on Environmental Quality, 2008), which is defined as a list of areas in Texas where specific pollutants were measured at levels of concern, generally using the state's Effects Screening Levels (ESLs) that have been developed for short-term and long-term exposure for approximately 5,000 substances (Texas Commission on Environmental Quality, 2008). In Harris County, for example, an area near the Lynchburg Ferry is on the APWL for benzene and styrene, and an area near Manchester is listed for 1,3-butadiene. In general, measurements from fixed-site monitors determine inclusion on the list.

For the specific purposes of the pilot study presented here, a "hot spot" is a 4 x 4-km cell with elevated predicted risk for admission to an area hospital for either cardiovascular or respiratory disease and that is associated with elevated levels of one or more simulated pollutants as determined by multivariate statistical analysis, controlling for various demographic factors.

RATIONALE

The primary rationale for this pilot study is to help advance the identification of potential "hot spots" of disproportionate exposure and health effects by (1) using an air pollution simulation model for more comprehensive geographical coverage than can be accomplished using monitors, and (2) utilizing actual measurable health endpoints to predict risk. The emphasis of this effort is on developing methods for the above and on delineating various problems that may limit its usefulness and/or that need further development.

The decision to study Harris County, Texas, for this preliminary study is the result of several factors. First, Harris County has some of the highest O_3 and HAP concentrations in the country. The county is a designated severe nonattainment area for the eight-hour O_3 standard, and ranked 1st in 2003 among U.S. counties for exposure to elevated 1-hour concentrations of O_3 and 7th in exposure to all NAAQS pollutants. Among HAPs, Harris County ranks 1st in the nation for total emissions of the 187 HAPs (Green Media Toolshed, 2008; U.S. Environmental Protection Agency, 2008).

In addition, Harris County has significant geographic pollutant gradients that make it particularly well suited for a "hot spot" study. The county, which includes most of Houston and is located in Southeast Texas along the Gulf Coast, has a mixture of densely populated and rural areas, complicated meteorology that results in high levels of photochemical species and a tendency for geographically localized areas of high pollution, often associated with emissions from its large concentration of petrochemical facilities (Webster et al., 2007). In addition, its lack of zoning likely amplifies the exposure gradient (Maantay, 2001).

The unusually large number of hospitals, many in or connected with institutions in the Texas Medical Center, provides excellent geographical coverage for the area and likely results in more consistent use of the hospital system and better data than in areas with fewer facilities. In our study, data from 95 hospitals in Harris County and the contiguous counties were used.

In addition, the county is exceptionally demographically diverse, with a population of more than 3.4 million, of whom approximately 42.1% are White, 32.9% are Hispanic or Latino, 18.5% are Black, and 5.1% are Asian (U.S. Census Bureau, 2000). Other demographic indicators, such as income, have wide gradients across the county with, for example, average median household income by census tract ranging from $6,673 to $170,121. Such demographic variability is important for teasing out the role of these

factors in vulnerability to adverse health events.

Harris County also has a large number of fixed-site air-pollution monitors, most of which collect hourly data, and has recently been the focus of several intensive air quality studies, including the Texas Air Quality Study 2000 (TexAQS-I) and, more recently, the 2006 TexAQS-II (Texas Commission on Environmental Quality, 2007; The University of Texas at Austin, 2000). The information and findings from these studies are available for our study and are being used to refine model simulations, inventories and, more generally, to better understand pollutant formation in the region. The UH-IMAQS is a central participant in these regional modeling and measuring efforts and is currently running daily near-real time CMAQ forecasts for the region, which are available online, along with statistical and visual measures of model performance (University of Houston Institute for Multi-dimensional Air Quality Studies, 2008).

HYPOTHESIS

The hypothesis that we are testing is that the age-adjusted rates, by discharge diagnosis, of Harris County residents hospitalized during the study period differ geographically among the 337 4 x 4-km cells that overlay the county and correlate with MM5-, CMAQ4.4- and CMAQ-HAP-simulated meteorological values and/or concentrations of ambient air pollutants, controlling for other variables.

OBJECTIVE AND SPECIFIC AIMS

The primary objective of this pilot study was to test, evaluate, and improve the methodology for conducting multipollutant "hot spot" analyses, using simulated pollutant concentrations and actual health effects. The hypothesis focused our initial efforts.

The specific aims of this exploratory study included:

- Simulate hourly meteorological and pollutant concentrations using the MM5, CMAQ4.4, and CMAQ-HAP models for 337 4 x 4-km cells across Harris County for July 1 through September 28, 2000.

- Conduct selected model performance evaluations and explorations of the CMAQ output, such as by comparing simulated and observed concentrations at available monitors within Harris County and developing different averaging schema or other representations of the simulated data, with an emphasis on evaluating the capability of CMAQ output to reasonably estimate human exposure.

- Obtain hospital admissions data for the study area; assess the quality of the data and clean as appropriate; extract and describe the relevant records; develop methods for characterizing individual-level information; and geocode residential addresses of Harris County patients for the study period, noting any difficulties.

- Obtain demographic data from Census 2000 for the study area; assess the quality of the data and clean as appropriate; extract the relevant records; and develop methods for apportioning the aggregate data to each of the 337 CMAQ 4 x 4-km cells, noting any difficulties.

- Use ArcGIS and statistical software to characterize each 4 x 4-km cell by the simulated meteorological and pollutant variables chosen, hospital admissions data by cardiovascular respiratory and cardiopulmonary (either) discharge diagnosis, and demographic information.

- Develop, evaluate, and refine statistical methods for assessing potential associations between simulated meteorological and pollutant variables and hospital admission rates, controlling for selected individual and group demographic variables, with the goal of developing "hot-spot" rankings and predictive maps of risk (i.e., probability of hospitalization by diagnosis category) by 4 x 4-km cell for the various outcomes studied.

- Use ArcGIS, SigmaPlot, SAS, and other software to create maps, graphs, and other visual representations as appropriate for the independent and dependent variables to aid in the exploration and communication of the data and of the analyses.

- Evaluate and discuss the methodology developed, problems encountered, and future efforts to be explored as a result of this pilot study.

- Share the methodology, problems, outcomes, and other aspects of the study through manuscripts, presentations, and other venues to stimulate discussion and approaches for improving the model.

METHODS

In this section, we discuss briefly the evolution of some of the methods used, followed by detailed descriptions of the actual methods used in the final study, including the data sources used.

PRELIMINARY WORK

Before attempting the analysis that forms the core of this final report, we explored, tested, and analyzed a single month of data, August 2000, using this first phase to explore a limited set of data, identify problems, and scrutinize and refine the methodology. It was during this first phase that we adapted the grid for the health data, scrutinized and developed statistical scripts for the hospital admissions data, and made numerous decisions, such as the decision to include the cells that overlapped the edges of Harris County, that we would subsequently apply to the expanded three-month data set. This initial exploratory model used the 1999 NEI (NEI99). Output was explored temporally, looking at four 6-hour averages (midnight to 6 am; 6 am to noon; noon to 6 pm; and 6 pm to midnight), as well as spatially, by examining 30-day averages across the 337 cells. It was also during our exploration of one month of data that we developed customized software and statistical scripts, described later in the Methods section, to reduce the computational and personnel time needed to run some of the CMAQ simulations, reduce the time required and potential for error while apportioning demographic data to the grid cells, and develop systematic methods for geoaddressing patient addresses to increase the accuracy and percentage of records geoaddressed within the constraints of time and resources. In addition, unanticipated problems associated with the use of a grid-based system for health effects data resulted in re-thinking in various ways the statistical methods for the multivariate regressions.

After various adjustments, we expanded the study to include three months of data (July-September 2000). During this phase, two models were developed, one with the regular year 2000 TEI and the other with the imputed TEI. The imputed inventory adds additional reactive VOCs to the regular inventory, based on observational data from the TexAQS studies, to better approximate peak ozone levels. The effect of this change is less well understood for some of the other species (Kim et al., 2006). Because there was no meteorological data for September 29 and 30 (CST), we initially ran the regular TEI analysis using meteorological input from other similar days in order to simulate the entire three months. Using the 92-day model, we examined four pollutant averaging schema temporally and spatially, and incorporated principal components analysis (PCA) and analysis of residual spatial autocorrelation into our statistical methods. The statistical team also developed methods to reduce uncertainty from variations in the cell populations. The statistical model used was a linear mixed-effects model (LMM), as was used in the final analyses described in this report. Thus for the 92-day analysis we developed 24 multivariate LMM, using four averaging schema, three outcomes, and for each outcome used either the PCA factors or the 16 HAPs. Lessons learned were incorporated into the analysis reported in this final report.

For the analyses reported here, we chose to drop the two days that used surrogate meteorological input, use the imputed TEI, and focus on a single averaging schema. The latter decision is discussed separately under Averaging Schema. We also added two meteorological variables, temperature and relative humidity, as independent variables, and chose to use NO_x in the model rather than NO_2 to better capture this pollutant class. We also examined the issue of multicollinearity in our model, which is the result of correlation between multiple pollutants in multivariate models. Because our primary study objective was to define "hot spots" of disproportionate exposure and health effects, we chose to accept the collinearity in the statistical model and to focus on using the full multivariate model to predict conditional hospitalization rates for each cell for each outcome. This decision, however, acknowledges that the interpollutant collinearity distorts the multivariate regression effect estimates such that they cannot be used to assess the individual contributions of the variables, a distinct disadvantage. On the other hand, the potential advantage is a more comprehensive and useful delineation of areas of concern, areas which then can be targeted for appraisal, additional data collection, and more rigorous statistical analyses. Some of the issues concerning multipollutant models are addressed in the Discussion section.

In the next sections we discuss the data sources and preparation of the data for use in the statistical model.

MODELING AND DATA COLLECTION

Specific data-related problems and challenges encountered are noted in this section as appropriate, with a more rigorous discussion of problems of particular concern in the Discussion.

Grid

The unit of spatial analysis chosen for this pilot study was the 4 x 4-km cell. This level of resolution is within the capabilities of the CMAQ model and is being used for forecasting in the HGA. Higher resolution at the 1 x 1-km resolution creates much higher computational demands and is not generally supported by the resolution of

emissions data at this time. In addition, a much longer study period would be required to have sufficient hospital admissions to support this resolution and was not feasible for this pilot study. Nevertheless, efforts are on-going to work at a higher resolution, which would be more capable of capturing meaningful neighborhood-level "hot spots." See Future Efforts.

Because data sources of different resolution needed to be apportioned at the cell level, accurate cell boundaries were particularly important. Creating the grid in ArcGIS-the software we used for creating and merging different geospatial data layers for subsequent analysis-was more challenging and time-consuming than first anticipated. Because CMAQ is designed to flexibly accommodate a wide range of input, its governing equations are expressed in a generalized coordinate system that determines the necessary grid and coordinate transformations, and it can accommodate various vertical coordinates and map projections (Byun and Schere, 2006). This relative flexibility is one of CMAQ's strengths, but it unexpectedly led to some challenges when apportioning specific patient addresses and demographic variables to the 4 x 4-km cells.

Geographically, there was a 30-meter difference between seconds to two decimal places and seconds to four decimal places-a sufficient difference to incorrectly place numerous patient addresses into the wrong cell. The first grids developed in ArcGIS from coordinates used for CMAQ's 4-km grid were inadequate for the demands of the health portion of this pilot study. We eventually used the highest level of precision available in the CMAQ system, using the longitude and latitude coordinates of the 1 x 1-km CMAQ grid in degrees, minutes, and seconds, with seconds to four decimal places. Working at this precision, we subsequently created ArcGIS-based 4-km and 12-km grids that co-localized with the 1-km grid. This not only produced the precision needed for accurately apportioning other health-related data layers to the cells, but allowed us to develop databases that can be utilized at higher grid resolutions for future work.

The coordinate system used was State Plane Coordinate System (SPCS), Texas South Central region (Federal Information Processing Standard (FIPS) 4204), feet, North American datum (NAD) 1983. This system uses the Lambert Conformal Conic projection. From the 4 x 4-km grid for East Texas (5,395 cells), we extracted the 337 cells that intersect with Harris County (Figure 1). These 337 cells, which are numbered by column and row as part of the larger East Texas grid, are used for all subsequent characterizations and analyses in this study.

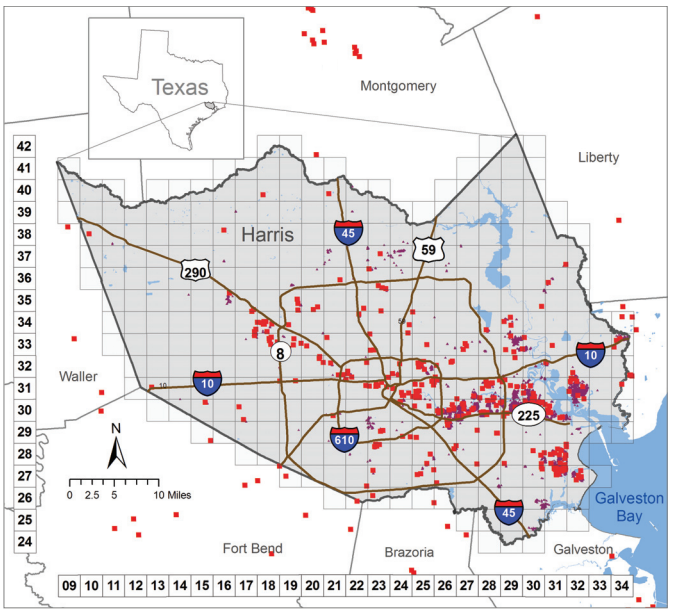


Figure 1: The study area, Harris County, Texas. Shown is the 4 x 4-km grid used for the meteorological and pollutant simulations, geospatial apportionment of residential addresses and demographic characteristics, and statistical analyses. Column (left) and row (bottom) numbers correspond to the regional CMAQ grid. Each cell is designated by a four-digit column-row name. The red squares and smaller purple triangles represent Toxic Release Inventory (TRI) and Harris County permitted facilities, respectively.

Meteorological and Pollutant Data

The exposures of primary interest in this multipollutant model included two meteorological variables (temperature and relative humidity), five CAPs (CO, O_3, NO_x, $PM_{2.5}$, SO_2) and 16 HAPs (acetaldehyde, acetone, acrolein, benzene, 1,3-butadiene, chloroform, cresols, ethylene dibromide, ethylene dichloride, ethylene oxide, formaldehyde, methylene chloride, perchloroethylene, phenols, trichloroethylene, and vinyl chloride). The selection of HAPs was primarily based on the health effects and toxicological literature, although a desire to include HAPs of particular interest to the HGA and to include representative chemical species also played a role. The hourly output for the final model was simulated for July 1 through September 29, 2000, as meteorological simulations were not available for September 30 and October 1. The additional day was necessary for subsequent conversion of the time base for all simulated species from Coordinated Universal Time (UTC) to Central Standard Time (CST), a conversion factor of -6 hours. Although our study period falls within Central Daylight Time (CDT), CST was used for comparisons with TCEQ monitors, which are reported in CST year-wide, and for easier expansion of the study time base in future work.

Modeled temperatures were extrapolated to a vertical height of 1.5 meters; relative humidity and pollutant concentrations were estimated at the middle of layer one, i.e., at approximately 17 meters. Hourly simulated concentrations were obtained on the hour and reflect centered mean averages of the period 30 minutes before and 30 minutes after the hour.

Averaging Schema

For this pilot spatial study, an averaging schema needed to be chosen to represent chronic ambient exposure in each of the 337 cells across the study period. This was complicated by a number of factors, including daily temporal variations in the maximal and minimal hourly concentrations across the pollutants, activity patterns that may make some windows of exposure more important than others, and variations among the pollutants in the penetrance of outdoor pollution indoors. Assessment of simulated concentrations against measurements from available monitors using the chosen average schema was another consideration.

Before choosing an averaging schema for the final pilot analysis, we examined in the 92-day model (see Preliminary Work) four averaging schema that had exposure relevance: (1) mean of the daily maximum one-hour concentrations, (2) mean of the daily maximum moving six-hour means, (3) mean of the daily six highest hourly concentrations, and (4) mean of the daily 24-hour means. Although there were subtle differences in the regression output based on the averaging schema chosen, for the preliminary development and refinement of our methodology we chose to use the 90-day mean of the 24-hour daily means for the simulated meteorological and pollutant values for each of the 337 cells. This was not the best averaging schema for assessing CMAQ O_3 model performance, the pollutant for which we had the most observed values. In our 92-day simulation, the mean simulated-to-observed ratio was 0.91 for the mean of the daily maximum moving six-hour means and 1.58 for the mean of the daily 24-hour means. The simulated values for the longer averaging period were approximately 58% higher than the observed values. This bias is discussed in the model performance section. The simulated-to-observed ratio for mean of the 24-hour means, however, for most of the other pollutants was generally as good as or better than other averaging times.

There were, however, other reasons that led us to choose the 90-day mean of the 24-hour daily means for the pilot study. Among these was a desire for a measure of chronic exposure. Our pilot study is fundamentally a study of the effect of geographical differences in on-going chronic exposure to multiple pollutants on vulnerability to adverse health effects, in our study measured by hospital admissions and especially by admissions for cardiovascular disease. Averaging shorter spans of simulated output (e.g., mean of the maximum daily one-hour concentrations or mean of the daily maximum moving six-hour means) may be generally more appropriate for acute effects. Biological plausibility is an important component of selecting an averaging schema.

Also, these shorter maximal spans, in a model with up to 23 meteorological and pollutant values in the final regression, occur at different times of the 24-hour day for different pollutants. As discussed under Temporal and Spatial Variability, examination of different averaging spans and/or time of day demonstrated significant variations in the time at which maximal values and/or moving averages occurred. Thus the peak value of one pollutant could be during the afternoon and another shortly after midnight. Because activity patterns and outdoor-to-indoor penetration ratios vary significantly across the pollutant species, we felt that using an averaging schema based on a relatively short but temporally variable daily maximum mean value introduced too many other variables into the model. Although a model could potentially be developed with different windows and averaging schema for each pollutant and meteorological variable, this was beyond the scope of this pilot study and could also introduce some inadvertent misclassification.

Another factor in our choice was comparison with other pollution and health effect studies, which most commonly use a 24-hour averaging schema. In the case of time-series and case-crossover designs such as the recent study by Dominici and associates of the effect of $PM_{2.5}$ and O_3 on hospital admissions for cardiovascular and respiratory outcomes (Dominici et al., 2006), the 24-hour average was used with various lags, whereas several prospective cohort studies that have addressed spatial differences, such as the six-city study (Dockery et al., 1993; Laden et al., 2006) and a recent geospatial study of chronic PM_{10} exposure in the Nurses' Health Study (Yanosky et al., 2008), used 24-hour means averaged over the study period.

Last, use of 24-hour means allowed more comparisons with observed values from area monitors, which were sometimes only available as 24-hour samples or had considerable missing or invalid hourly data that often did not allow us to identify one-hour or moving six-hour maximum concentrations but was usually adequate to calculate a daily 24-hour mean. The importance of the

averaging schema in teasing out the best exposure metric is addressed again in the Discussion.

MM5 Model

Temperature and relative humidity were among the independent variables included in the health-based statistical model. Our initial proposal did not include meteorological variables as potential predictors of outcome as our preliminary work suggested that the spatial variation across the 337 cell in Harris County, given the relatively short modeling run, would not warrant inclusion. After analysis of the 92-day dataset, we decided to add these to the final model. In addition to their potential value as independent variables, they were useful in assessing CMAQ model performance as high-quality simulation of meteorology is a critical component of the ability of the model to subsequently simulate pollutant concentrations adequately.

Hourly values for temperature and relative humidity were simulated by the UH-IMAQS researchers using version 3.6.1 of the fifth generation Pennsylvania State/National Center for Atmospheric Research mesoscale model (MM5). The MM5 is a limited-area nonhydrostatic, terrain-following sigma-coordinate model designed to simulate or predict mesoscale atmospheric circulation. MM5 is widely used for providing meteorological characterizations throughout the air quality modeling community.

The physical options for the MM5 simulation, the output of which was used both directly in the health-based model and as input for the CMAQ4.4 and CMAQ-HAP simulations, were as follows:

- Grell cumulus at 36 km and 12 km; no cumulus schema at 4 km
- Analysis nudging for domain 1 (36 km) and domain 2 (12 km)
- Continuous one-way nesting for domain 1, domain 2, and domain 3 (4 km)
- Medium-Range Forecast Planetary Boundary Layer (MRF PBL) parameterization
- Rapid Radiative Transfer Model (RRTM) radiation scheme
- Modified National Oceanic and Atmospheric Administration Land-Surface Model (NOAA LSM) with Texas Forest Service year 2000 Land Use/Land Cover (TFS LULC 2000) data; addition of the TFS data significantly improves land-cover resolution and categorization for the HGA (Cheng and Byun, 2008)

CMAQ4.4 and CMAQ-HAP Models

The air pollution data for this study were simulated by the UH-IMAQS team on a Linux cluster (parallel mode) using the CMAQ4.4 and CMAQ-HAP models, briefly described earlier in the Air Quality Modeling section. The simulations were run at 36-km (133 x 91 cells), 12-km (89 x 89 cells), and 4-km (83 x 65 cells) resolutions. The simulations at 36 km and 12 km were used to provide the boundary conditions for the 12-km and 4-km simulations, respectively.

The chemical mechanism used in the CMAQ4.4 model was a refined version of the SAPRC99 mechanism, i.e., SAPRC99-ARO, in which 18 aromatics (including benzene, acrolein, 1,3-butadiene, propene, and styrene) were explicitly represented. Thus the CMAQ4.4 model used for this study is in effect an air toxics modification of CMAQ (see also the earlier more general description the CMAQ model). The selection of gas-phase HAPs was based on several factors, including regional levels and published literature (Mayor's Task Force on the Health Effects of Air Pollution, 2006). For the gas-phase chemistry, a computationally efficient numerical method, the Euler Backward Iterative (EBI) solver, was used. The Piecewise Parabolic Method (PPM) was used for the horizontal and vertical advections, and AERO3 was used for the aerosols. Version 2.1 of the Sparse Matrix Operator Kernal Emissions Modeling (SMOKE) system was used to apportion spatially, chemically and temporally the source emissions. MM5 was used to simulate the meteorological conditions for all three grid resolutions, using the parameters listed in the previous section. For the final 90-day CMAQ4.4 simulation the following emission inventories were used:

- 36-km grid resolution
  - 1999 National Emissions Inventory (NEI99) final version 3 (U.S. Environmental Protection Agency, 2001)
- 12- and 4-km grid resolutions
  - 2000 Texas Emissions Inventory (TEI 2000 base5b imputed) + NEI99 final version 3 + MOBILE6 vehicle emissions
  - Hybrid version of Global Biosphere Emissions and Interactions System (GloBEIS) and the EPA's Biogenic Emissions Inventory System (BEIS), using TCEQ LULC

input and Meteorology-Chemistry Interface Processor (MCIP) output

For the CMAQ-HAP simulation, which uses CMAQ4.4 output for its primary input, only NEI99 final version 3 was used. Although we present here output using the TEI 2000 base5b imputed inventory, in which additional highly reactive volatile organic compounds (HRVOCs) were added to the regular TEI to compensate for recognized shortcomings in the inventory, the UH-IMAQS also ran the simulation with the regular TEI as noted earlier. Both outputs were evaluated and explored in the health model. The imputed inventory generally improves model simulations of O_3 on high ozone days (Kim et al., 2006), and was used in our final 90-day model.

Thus MM5, SMOKE, CMAQ4.4, and CMAQ-HAP were utilized to generate the simulated output, which included temperature, relative humidity, CO, O_3, NO_2, NO, $PM_{2.5}$, PM_{10}, SO_2, and 16 HAPs. For our final model, NO_2 and NO were combined to form NO_x, which was thought to better represent the nitrogen oxide species. In addition, PM_{10} was not included in the model as, based on the health literature, $PM_{2.5}$ appears to be a better measure of PM-related health effects. CMAQ4.4, using the SAPRC99-ARO modification, was used to simulate eight of the HAPs (acetaldehyde, acetone, acrolein, benzene, 1,3-butadiene, cresols, formaldehyde, and phenols). CMAQ-HAP was used to simulate the other eight HAPs of interest (chloroform, ethylene dibromide, ethylene dichloride, ethylene oxide, methylene chloride, perchloroethylene, trichloroethylene, and vinyl chloride).

Descriptive statistics of the MM5- and CMAQ4.4- and CMAQ-HAP-simulated output for July 1-September 28, 2000 are shown in Table 1, using the averaging period chosen for this spatial chronic exposure study, i.e., the 90-day mean of the 24-hour means.

Table 1: Descriptive statistics of simulated output. For this study, the following 26 ground-layer (vertical layer 1) air pollutant and meteorological values were simulated by the MM5, CMAQ4.4 and CMAQ-HAP models for 337 4 x 4-km cells in or intersecting the boundary of Harris County, Texas, for July 1 through September 28, 2000. The meteorological and pollutant measurements shown here are expressed as the 90-day mean of the 24-hour daily means for each of the 337 cells, the averaging schema used for this pilot study.

SPECIES	ABBR	UNIT	N	MEAN 90-DAY 24-HOUR MEANS	STD DEV	MINIMUM 90-DAY 24-HOUR MEAN	MAXIMUM 90-DAY 24-HOUR MEAN
Relative Humidity	RH	%	337	62.00	1.38	58.62	65.43
Temperature	TEMP	C	337	28.69	0.66	27.63	30.15
Carbon monoxide	CO	ppm	337	0.19	0.05	0.11	0.40
Nitrogen dioxide	NO_2	ppb	337	10.25	3.92	3.37	21.40
Nitrogen oxide	NO	ppb	337	1.36	1.12	0.16	7.25
Nitrogen oxides	NO_x	ppb	337	11.62	4.97	3.53	28.15
Ozone	O_3	ppb	337	38.09	2.73	29.02	42.41
Particulate matter < 2.5 µm	$PM_{2.5}$	µg/m3	337	9.63	2.48	5.10	26.08
Particulate matter < 10 µm	PM_{10}	µg/m3	337	30.14	10.55	10.60	85.78
Sulfur dioxide	SO_2	ppb	337	2.97	2.28	1.16	28.37
Acetaldehyde	CCHO	ppb	337	0.64	0.07	0.50	0.98
Acetone	ACET	ppb	337	1.89	0.07	1.69	2.11
Acrolein	ACROL	ppb	337	0.01	0.02	0.00	0.16
Benzene	BENZ	ppb	337	0.18	0.10	0.04	0.79
1,3-Butadiene	BUTA	ppb	337	0.13	0.22	0.01	2.61
Chloroform	CHCL	ppb	337	0.01	0.03	0.00	0.39
Cresols	CRES	ppb	337	0.01	0.00	0.00	0.03
Ethylene dibromide	BRC	ppb	337	0.00	0.00	0.00	0.00
Ethylene dichloride	CLC	ppb	337	0.01	0.02	0.00	0.21
Ethylene oxide	ETOX	ppb	337	0.01	0.01	0.00	0.16
Formaldehyde	HCHO	ppb	337	2.48	0.39	1.83	4.00
Methylene chloride	CLME	ppb	337	0.04	0.02	0.01	0.20
Perchloroethylene	CL4ET	ppb	337	0.01	0.01	0.00	0.05
Phenols	PHEN	ppb	337	0.00	0.01	0.00	0.09
Trichloroethylene	CL3ET	ppb	337	0.00	0.01	0.00	0.08
Vinyl chloride	CLET	ppb	337	0.01	0.02	0.00	0.25

Abbreviations: ABBR = abbreviation; CMAQ = Community Multiscale Air Quality; CMAQ-HAP = CMAQ adapted for selected gas-phase HAPs; HAP = hazardous air pollutant; ppb = parts per billion volume; ppm = parts per million volume; MM5 = Fifth-Generation National Center for Atmospheric Research / Penn State Mesoscale Model; $\mu g/m^3$ = micrograms per cubic meter; µm = micrometers (microns); N = number; ppm = parts per million volume; STD DEV = standard deviation

Temporal and Spatial Variability

Daily temporal variations in the simulated CMAQ output were explored by plotting the mean concentrations for three-, six- and 12-hour spans (not shown), beginning at midnight CST, for the 90-day interval. Photochemically active species such as formaldehyde, ozone, and acrolein tended to demonstrate significant variations during the sunlight hours, either increasing or decreasing depending on their photochemistry. Less photochemically active species, such as PM, CO, and perchloroethylene, demonstrated much smaller differences during the 24-hour day. As noted in our discussion of the averaging schema chosen for this pilot study, future refinements of the exposure metric may wish to utilize these daily temporal variations to weight estimated exposure based on activity patterns and indices of toxicity.

Another measure of interest was within-cell variability. Although we did not include any measure of within-cell variability in this pilot study, we speculate that subgrid variability in conjunction with intercell differences in chronic baseline concentrations may be important in triggering health effects, such as hospital admissions. In Figure 2, we plotted the standard deviations for the hourly simulated concentrations for four pollutants, CO, O_3, NO_x, and $PM_{2.5}$, with the 4 x 4-km cell with the highest variability shown in black. In future refinements, subgrid variability can be used to refine exposure and/or vulnerability to health effects that may be particularly affected by large variations between the minimum and maximum concentrations. See also Future Efforts.

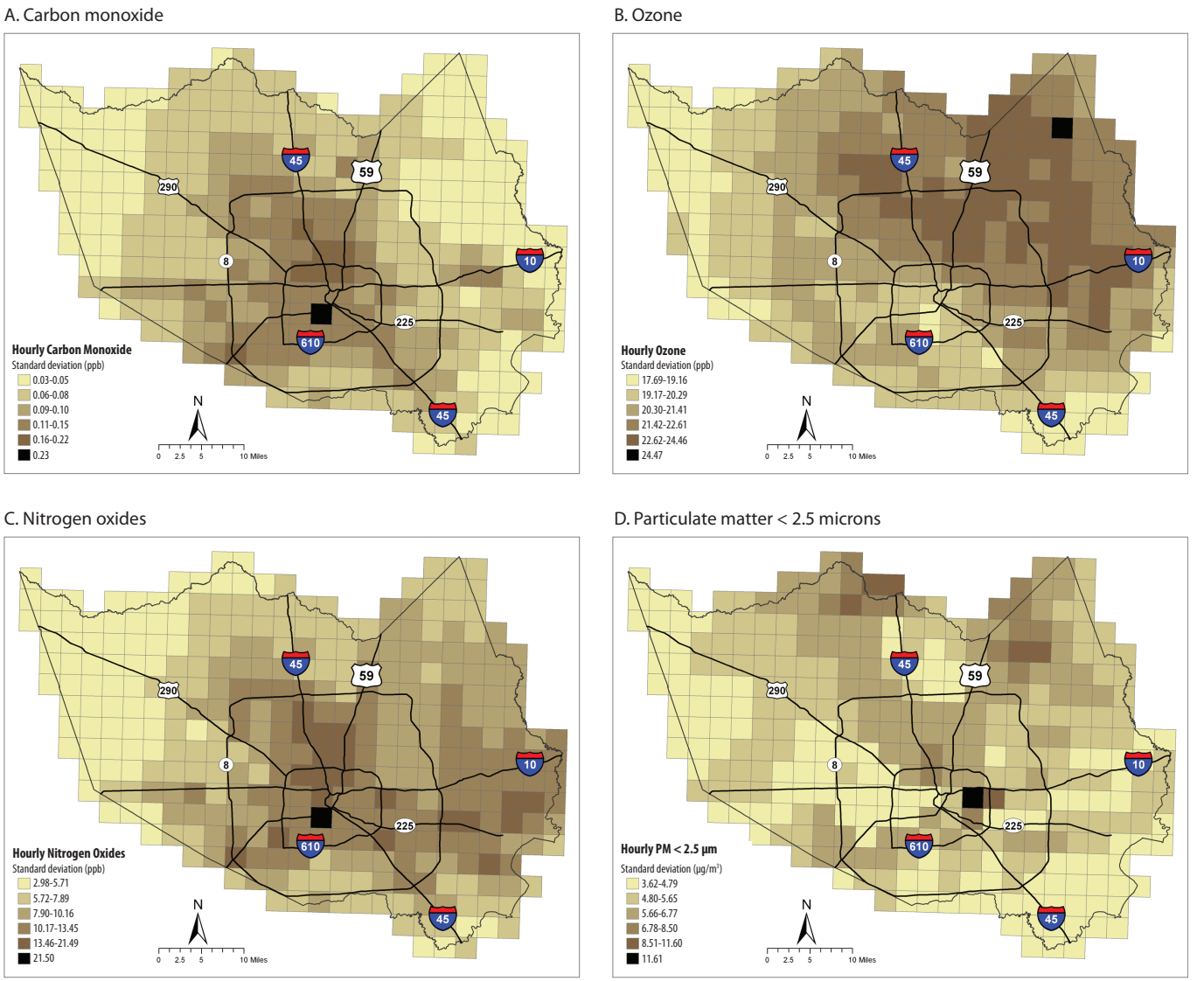


Figure 2: Within-cell variability by standard deviation. Shown, by cell, is the standard deviation for the hourly concentrations (N = 2,160). A. Carbon monoxide; B. Ozone; C. Nitrogen oxides; and D. Particulate matter < 2.5 microns. Intracell variability, along with baseline concentrations, may be a factor in hospital admissions.

In this "hot-spot" chronic exposure study, spatial variations across the 337 cells are of particular interest. To help us explore the spatial pollutant gradients across the 337 cells in Harris County, we created maps to demonstrate the spatial distribution of the meteorological and pollutant variables simulated by the MM5, CMAQ4.4, and CMAQ-HAP models, using the 90-day mean of the 24-hour daily means. In Figure 3 (A-K), three-dimensional maps of concentrations in all 337 cells, as well as two-dimensional color-coded orientation maps, are displayed, with the cell with the highest concentration for that particular meteorological or pollutant variable shown in black. For purposes of this study, only those variables that tended to remain in the multivariate LMMs, along with NO_x and two HAPs of special concern in the HGA, are shown. In general, there was considerable heterogeneity in the concentration levels across the cells.

Model Performance

In this pilot study we examined potential associations between gridded simulated meteorological values and

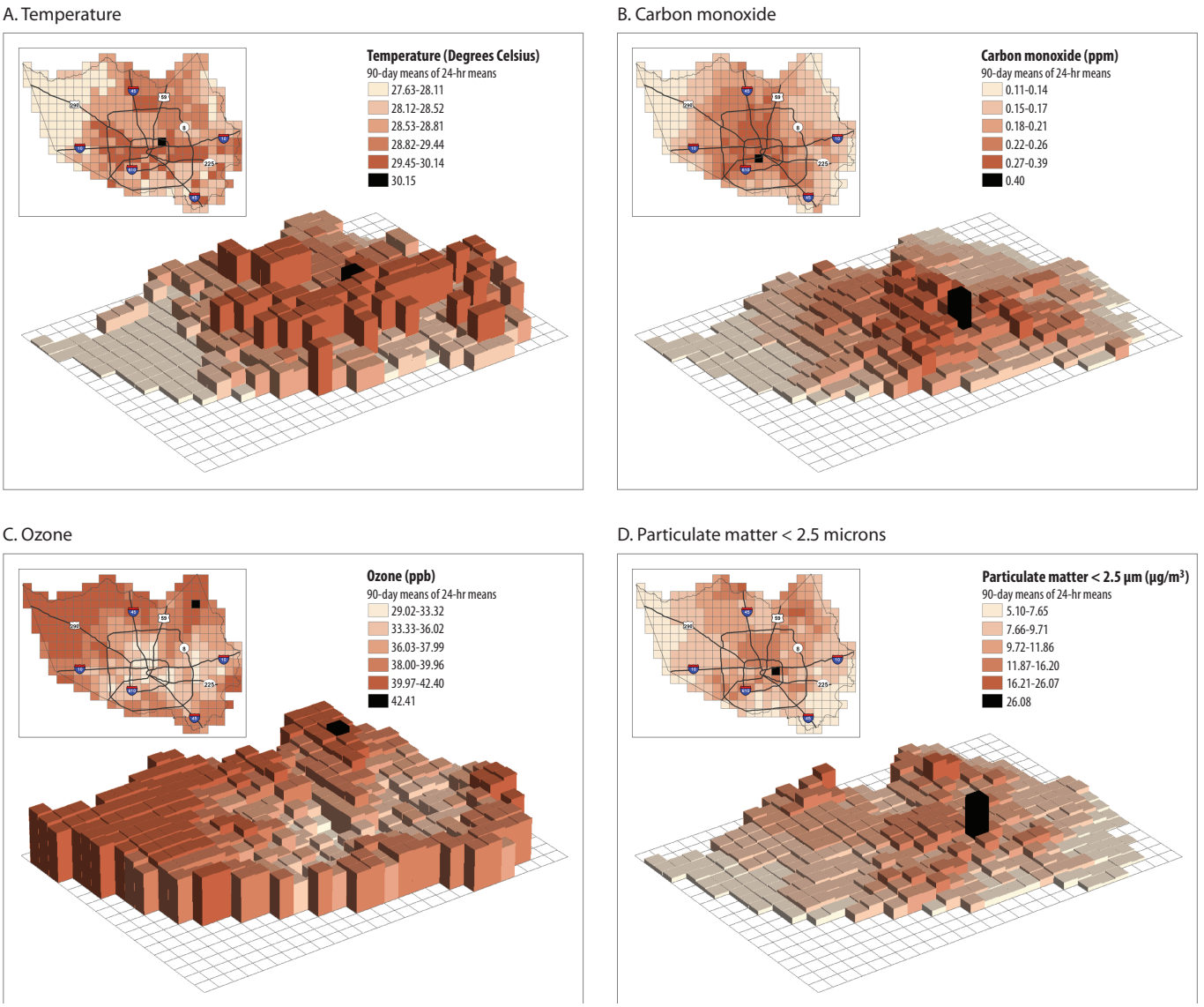


Figure 3: Spatial variations, by cell, across Harris County. Shown are 10 MM5-, CMAQ4.4-, or CMAQ-HAP-simulated variables based on the 90-day mean of the 24-hour daily means for July 1 through September 28, 2000. In both the three-dimensional and two-dimensional maps, the black cell marks the cell with the highest concentration for that pollutant. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Acetaldehyde; F. Acrolein; G. Formaldehyde; H. Methylene chloride; I. Nitrogen oxides, J. Benzene; and K. 1,3-Butadiene.

pollutant concentrations and health effects. Whether or not the simulated values are adequate surrogates for exposure is of key concern. A comprehensive assessment of model performance by MM5, CMAQ4.4, and CMAQ-HAP is beyond the scope of this study. Numerous investigators in the U.S. and elsewhere are, however, evaluating the performance of CMAQ using increasingly sophisticated analytic tools (Irwin et al., 2008; Napelenok et al., 2008; Stein et al., 2007; Zhang et al., 2007).

In the HGA, the CMAQ4.4 model is currently being run at 4-km resolution daily on multiple processors at the UH-IMAQS for near-real-time forecasting. Performance evaluations (spatial and time-series plots and statistics comparing CMAQ output and observed concentrations from area monitors) are made available each day online (www.imaqs.uh.edu/aqfmain.htm) for the 36-km U.S., 12-km East Texas, and 4-km HGA domains for O_3, NO, $PM_{2.5}$, temperature, and wind speed. The CMAQ4.4 parameters for the current simulations are largely identical to those used in our study, except that our study specifically extracts a

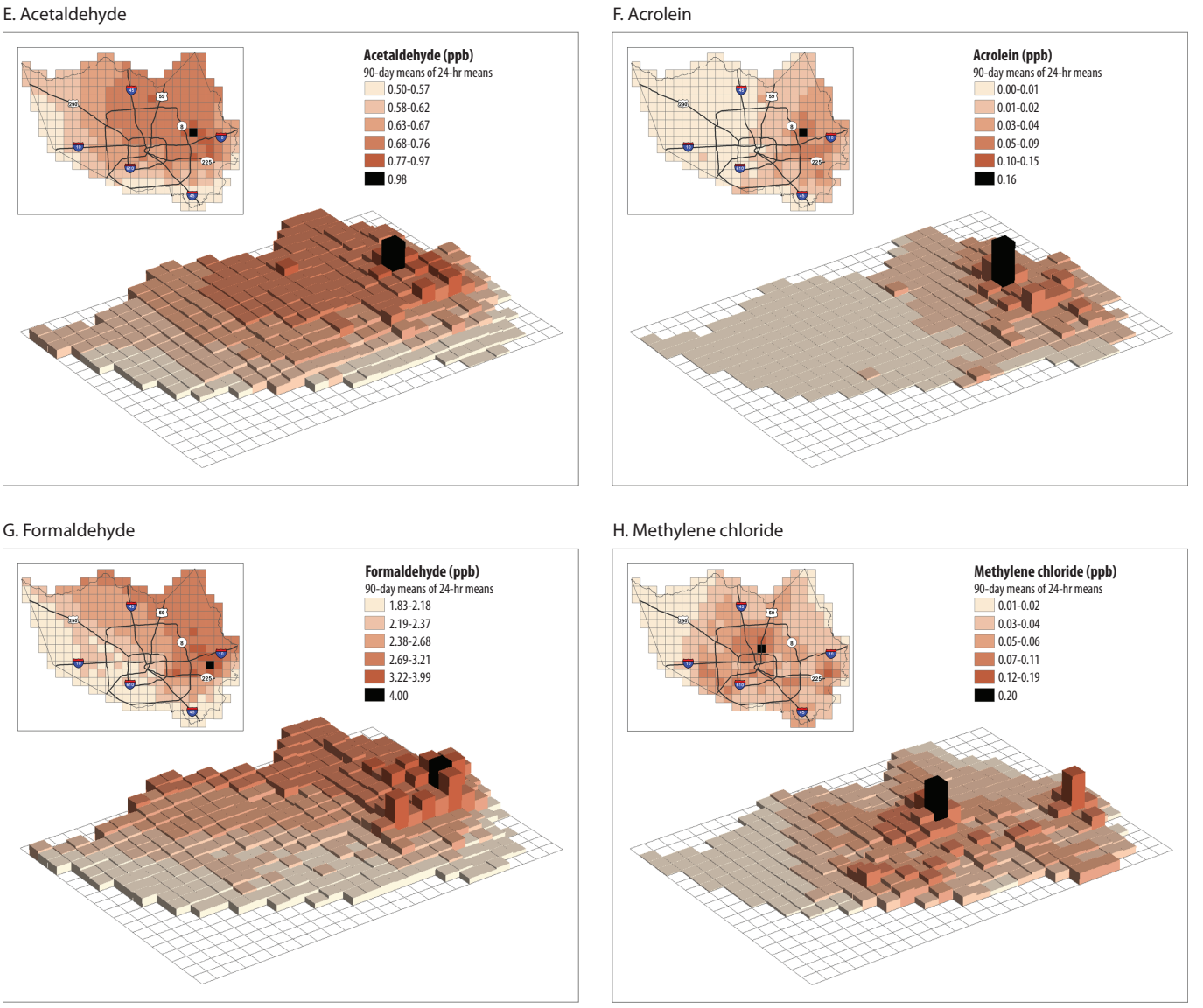


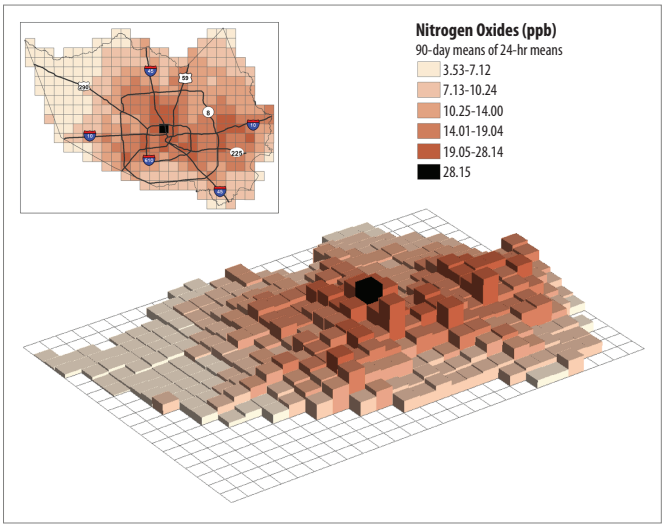
Figure 3 (cont.): Spatial variations, by cell, across Harris County. Shown are 10 MM5-, CMAQ4.4-, or CMAQ-HAP-simulated variables based on the 90-day mean of the 24-hour daily means for July 1 through September 28, 2000. In both the three-dimensional and two-dimensional maps, the black cell marks the cell with the highest concentration for that pollutant. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Acetaldehyde; F. Acrolein; G. Formaldehyde; H. Methylene chloride; I. Nitrogen oxides, J. Benzene; and K. 1,3-Butadiene.

number of HAP species. Nevertheless, the daily comparisons of the CMAQ4.4 simulations and concentrations measured at regional CAMS provide useful information for assessing the some of the strengths and weaknesses of the year 2000 simulated output used in our study.

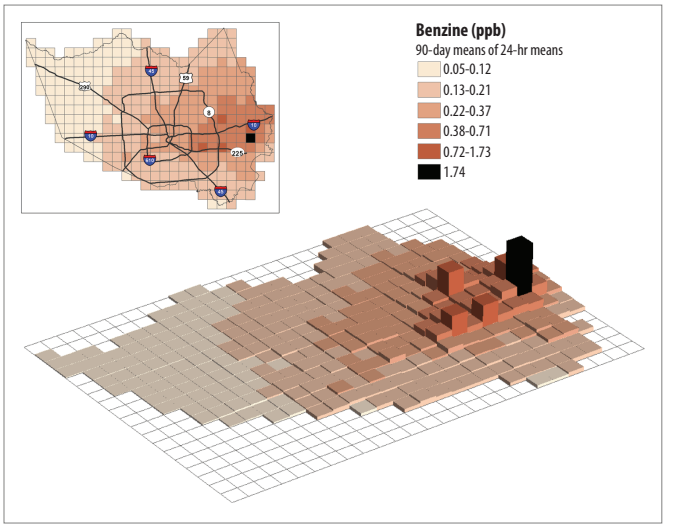
Comparison of Simulated and Observed Concentrations

We received extensive pollutant and meteorological data for 2000 from researchers at TCEQ. Depending on the species and monitor, the databases have undergone varying degrees of validation since 2000. Working with TCEQ, we accessed online or received as text files from a number of divisions within TCEQ all of the available hourly and intermittent measured ambient pollutant and

I. Nitrogen oxides



J. Benzene



K. 1,3-Butadiene

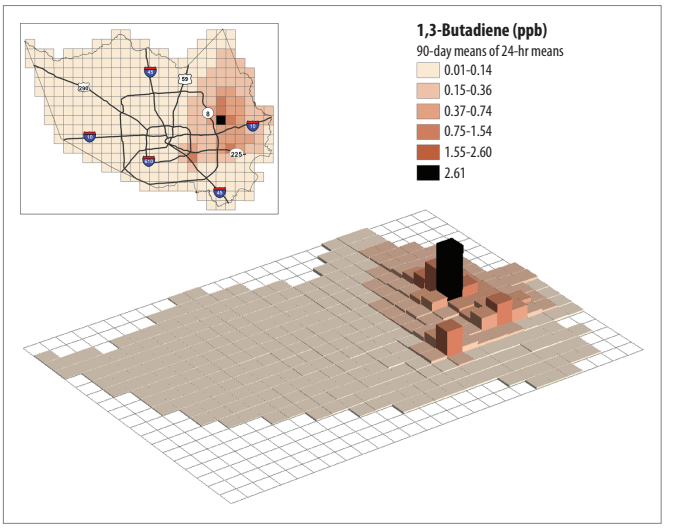


Figure 3 (cont.): Spatial variations, by cell, across Harris County. Shown are 10 MM5-, CMAQ4.4-, or CMAQ-HAP-simulated variables based on the 90-day mean of the 24-hour daily means for July 1 through September 28, 2000. In both the three-dimensional and two-dimensional maps, the black cell marks the cell with the highest concentration for that pollutant. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Acetaldehyde; F. Acrolein; G. Formaldehyde; H. Methylene chloride; I. Nitrogen oxides, J. Benzene; and K. 1,3-Butadiene.

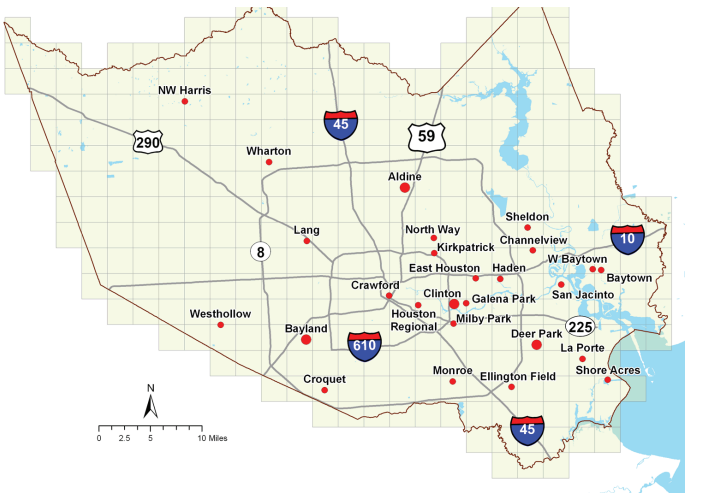


Figure 4: Continuous air monitoring stations (CAMS). Shown are active CAMS in Harris County during the study period, July 1-September 28, 2000. See Table 2 for a list of variables measured at each monitor. The larger red circles denote the four monitor locations that measure the most chemical species.

meteorological data. For the purposes of this study, we extracted only those species we simulated, and we calculated 90-day means of 24-hour daily means whenever possible for comparison with the averaging schema used in this pilot study. Twenty-four hour means were only calculated if there were a minimum of 20 hourly measurements. Twenty-six monitors in Harris County collected data during the study period; however, the amount of data collected at each varied considerably (Table 2; Figure 4). Descriptive statistics of the pollutant concentrations measured at Harris County monitors during the study period are given in Table 3.

Comparisons between the MM5-, CMAQ4.4-, and CMAQ-HAP-simulated output with concentrations measured at monitors focused on four monitors because of their location and the availability of data. Of the HAP species, only

Table 2: Harris County monitors. The following monitors (N = 26) were operational for the 90-day study period, July 1-September 28, 2000. Observed data were received directly from the Texas Commission on Environmental Quality (TCEQ) as well as downloaded from TCEQ online databases.

CAMS ID #	EPA AIRS ID #	COL ROW	NAME	CONTINUOUS HOURLY DATA	INTERMITTENT 1-HR and 24-HR SAMPLES (#)
1	48-201-1034	2731	HOUSTON EAST	MET, NO_x, O_3, $PM_{2.5}$	
8	48-201-0024	2435	ALDINE	MET, CO, NO_x, NO_y, O_3, BZ, BUT	CARB24 (7), CATMN (15), PM_{10} (16)
15	48-201-0026	2932	CHANNELVIEW	MET, $PM_{2.5}$	CATMN (12)
26	48-201-0029	1638	NW HARRIS	NO_x, O_3, SO_2	CATMN (15), PM10 (15)
35	48-201-1039	2929	DEER PARK	MET, CO, NO_y, NO_x, O_3, $PM_{2.5}$, BZ, BUT	CARB1 (33), CARB24 (18), CATMN (13), PM_{10} (11)
48	48-201-0068	1936	WHARTON	MET	
51	48-201-1040	2827	ELLINGTON	MET	
53	48-201-0055	2029	BAYLAND PARK	MET, NO_x, O_3, BZ, BUT	CARB1 (7), CATMN (12)
81	48-201-0070	2530	HOUSTON REG	MET, SO_2, O_3	
145	48-201-0061	3227	SHORE ACRES	MET	CATMN (15)
148	48-201-0058	3132	BAYTOWN	MET	CATMN (14)
166	48-201-1041	3031	SAN JACINTO	MET	CATMN (13)
167	48-201-0057	2630	GALENA PARK	MET	CATMN (15)
169	48-201-0069	2630	MILBY PARK	MET	CATMN (13)
403	48-201-1035	2630	CLINTON	MET, CO, NO_x, O_3, SO_2, BZ, BUT	CARB1HR (37), CARB24 (13), CATMN (11), PM_{10} (28)
404	48-201-0060	2532	KIRKPATRICK	MET	
405	48-201-0046	2533	N WAYSIDE	SO_2, O_3	
406	48-201-0062	2627	MONROE	MET, SO_2, O_3	PM_{10} (14)
407	48-201-1037	2331	CRAWFORD	CO, NO_x, O_3	PM_{10} (14)
408	48-201-0047	2033	LANG	CO, NO_x, O_3	PM_{10} (15)
409	48-201-0051	2127	CROQUET	MET, SO_2, O_3	
410	48-201-0066	1729	WESTHOLLOW	MET, O_3	PM_{10} (12)
603	48-201-0803	2831	HADEN	MET, NO_x, O_3	CATMN (15)
604	48-201-0804	2933	SHELDON	MET, NO_x, O_3	
607	48-201-0807	3232	BAYTOWN	MET, NO_x, O_3	
608	48-201-0808	3128	LA PORTE	MET, NO_x, O_3	

Abbreviations: BUT = 1,3-butadiene; BZ = benzene; CARB1 and CARB24 = 1-hour and 24-hour samples of carbonyls, including acetaldehyde, acetone and formaldehyde; CATMN = Community Air Toxics Monitoring Network, 24-hour samples of ~ 107 hazardous air pollutants; CO = carbon monoxide; MET = meteorological variables including wind speed, wind direction, outdoor temperature, dew point temperature, relative humidity, solar radiation, net radiation, and/or precipitation; NOx = nitrogen oxides, including nitrogen oxide and nitrogen dioxide; NO_y = total reactive nitrogen; O_3 = ozone; PM_{10} = particulate matter < 10 microns (24-hour samples); $PM_{2.5}$ = particulate matter < 2.5 microns; SO_2 = sulfur dioxide

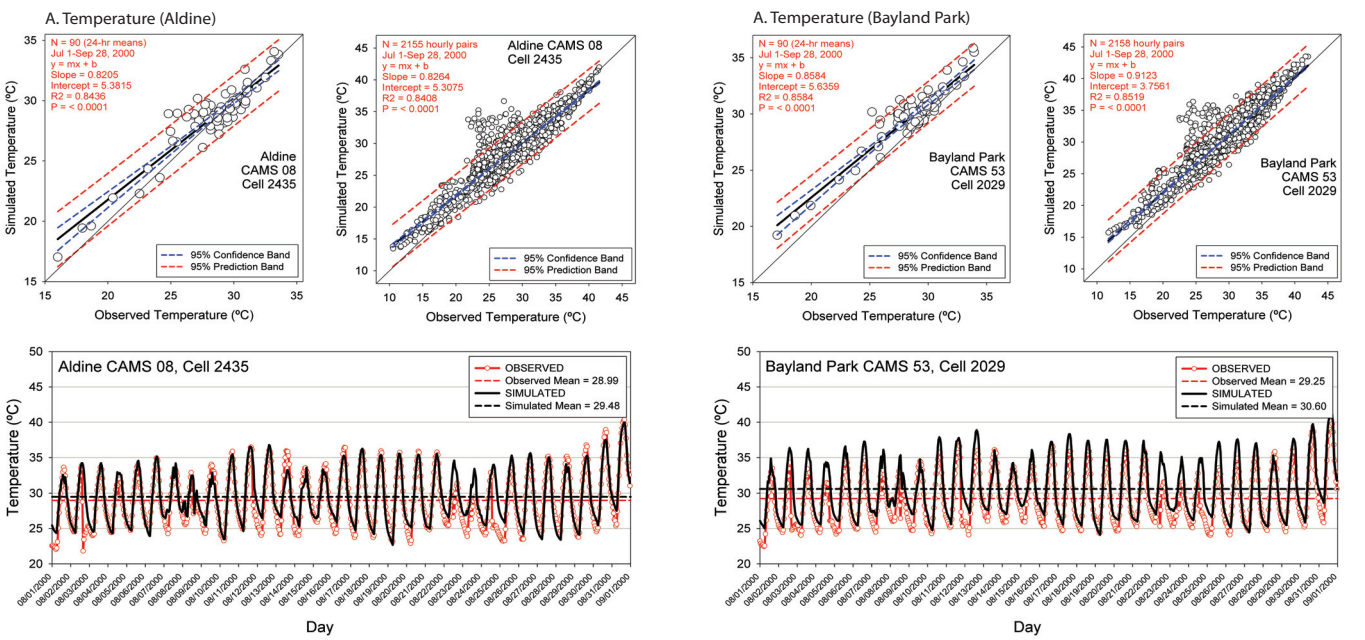


Figure 5: Comparison of simulated and observed values. Shown are meteorological and pollutant variables that tended to remain in the multivariate linear mixed-level regression, as well as nitrogen oxides and benzene (a pollutant of special concern in the Houston-Galveston area). Of the air toxic species, only benzene had sufficient hourly observed measurements for comparison. For each, scatter plots of the 24-hour daily means and of the hourly pairs for the 90-day study period, as well as a time-series plot for August are shown. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Nitrogen oxides; and F. Benzene.

Table 3: Descriptive statistics of observed pollutant concentrations. Data are from all available data at all monitors in Harris County that were operational between July 1 and September 28, 2000. Data are expressed as 90-day means of 24-hour daily means, based on available collection methods and sufficient data. Only CAMS data, and to a lesser extent AutoGC data, have sufficient hourly data for reasonable comparison with Table 1. For purposes of this table, only those pollutants that were simulated by the MM5, CMAQ4.4, and CMAQ-HAP models for this study are included. Data are from the Texas Commission on Environmental Quality.

SPECIES	ABBR	UNIT	N OBS	N MON	MEAN OF 90-DAY* 24-HR MEANS	STD DEV	MIN 90-DAY* 24-HR MEAN*	MAX 90-DAY* 24-HR MEAN*	TCEQ SOURCE
90-Day Means from 24-Hour Means Computed from Hourly Measurements									
Relative humidity	RH	%	8,588	4	72.78	13.45	67.20	86.21	CAMS
Temperature	TEMP	C	46,547	22	28.78	2.02	27.88	30.74	
Carbon monoxide	CO	ppm	9,875	5	0.43	0.06	0.38	0.53	
Nitrogen oxide	NO	ppb	16,127	8	6.607	3.80	1.66	12.14	
Nitrogen dioxide	NO_2	ppb	15,684	8	12.46	3.46	6.74	16.73	
Nitrogen oxides	NO_X	ppb	19,254	11	17.12	7.75	8.18	25.49	
Ozone	O_3	ppb	34,668	17	27.51	10.22	22.10	36.56	
PM < 2.5 µm	$PM_{2.5}$	µg/m^3	6,352	3	12.84	1.28	11.77	13.59	
Sulfur dioxide	SO_2	ppb	8,195	4	2.78	1.97	1.54	5.72	
Acetaldehyde	CCHO	ppb	100	2	ID	ID	ID	ID	CARB 1HR
Acetone	ACET	ppb	153	3	ID	ID	ID	ID	
Formaldehyde	HCHO	ppb	100	2	ID	ID	ID	ID	
1,3-Butadiene	BUTA	ppb	2,748	4	0.30	0.31	0.08	4.00	AutoGC
Benzene	BENZ	ppb	3,747	4	0.44	0.24	0.20	4.00	
90-Day Means from Intermittent 24-Hour Canister Measurements									
Acetaldehyde	CCHO	ppb	31	2	0.98	0.31	0.77	1.20	CARB 24HR
Acetone	ACET	ppb	31	2	0.14	0.02	0.13	0.15	
Formaldehyde	HCHO	ppb	31	2	9.82	3.00	7.70	11.94	
1,3-Butadiene	BUTA	ppb	163	12	0.13	0.22	0.00	0.81	CATMN
Benzene	BENZ	ppb	163	12	0.13	0.10	0.05	0.44	
Chloroform	CHCL	ppb	163	12	0.01	0.00	0.01	0.02	
Ethylene dichloride	CLC	ppb	163	12	0.00	0.00	0.00	0.00	
Methylene chloride	CLME	ppb	163	12	0.07	0.08	0.01	0.26	
Trichloroethylene	CL3ET	ppb	163	12	0.02	0.01	0.00	0.04	
Vinyl chloride	CLET	ppb	163	12	0.01	0.01	0.00	0.02	
PM < 10 µm	PM10	µg/m^3	118	8	29.80	10.77	19.45	54.65	hn PM

Abbreviations: ABBR = abbreviation; AutoGC = automated 1-hour averaging gas chromatography; CAMS = continuous air monitoring sites; CARB 1HR = carbonyl 1-hour averaging monitors; CARB 24HR = carbonyl 24-hour averaging monitors; CATMN = Community Air Toxics Monitoring Network; CMAQ = Community Multi-scale Air Quality; CMAQ-HAP = CMAQ adapted for selected gas-phase HAPs; MM5 = Fifth-Generation National Center for Atmospheric Research / Penn State Mesoscale Model; µg/m^3 = micrograms per cubic meter; µm = micrometers (microns); ID = insufficient data; MAX = maximum; MIN = minimum; N OBS = total number of measurements; N MON = number of monitors; PM = particulate matter; ppb = parts per billion volume; ppm = parts per million volume; STD DEV = standard deviation; * Actual number varies based on sufficient data for calculating each 24-hour mean

benzene had sufficient hourly observed measurements for comparison. Of the Harris County monitors, the monitors at Aldine (EPA 48-201-0024), Bayland Park (EPA 48-201-0055), Clinton (EPA 48-201-1035), and Deer Park (EPA 48-201-1039) are among the monitors that collect the most comprehensive data, currently and in 2000. Figure 5 (A-F) displays scatter plots of the simulated and observed 90-day mean of the 24-hour daily means and of the hourly pairs, as well as time-series plots for August 2000, for temperature, CO, O_3, $PM_{2.5}$, NO_X, and benzene. Temperature, CO, O_3, and $PM_{2.5}$ are plotted in part because they tended to remain in our multivariate models but also because of their suspected role in pollution-related cardiopulmonary disease. Nitrogen oxides and benzene were included based on recent literature that is suggestive of the role of NO_x in

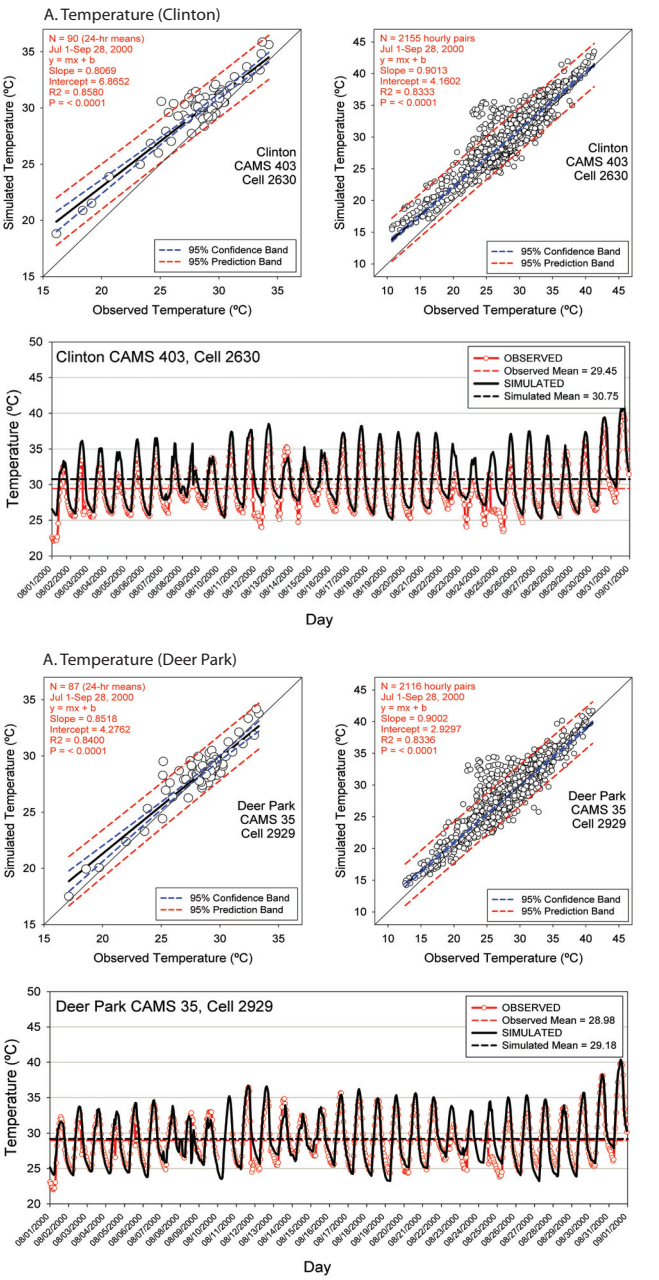


Figure 5 (cont.): Comparison of simulated and observed values. Shown are meteorological and pollutant variables that tended to remain in the multivariate linear mixed-level regression, as well as nitrogen oxides and benzene (a pollutant of special concern in the Houston-Galveston area). Of the air toxic species, only benzene had sufficient hourly observed measurements for comparison. For each, scatter plots of the 24-hour daily means and of the hourly pairs for the 90-day study period, as well as a time-series plot for August are shown. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Nitrogen oxides; and F. Benzene.

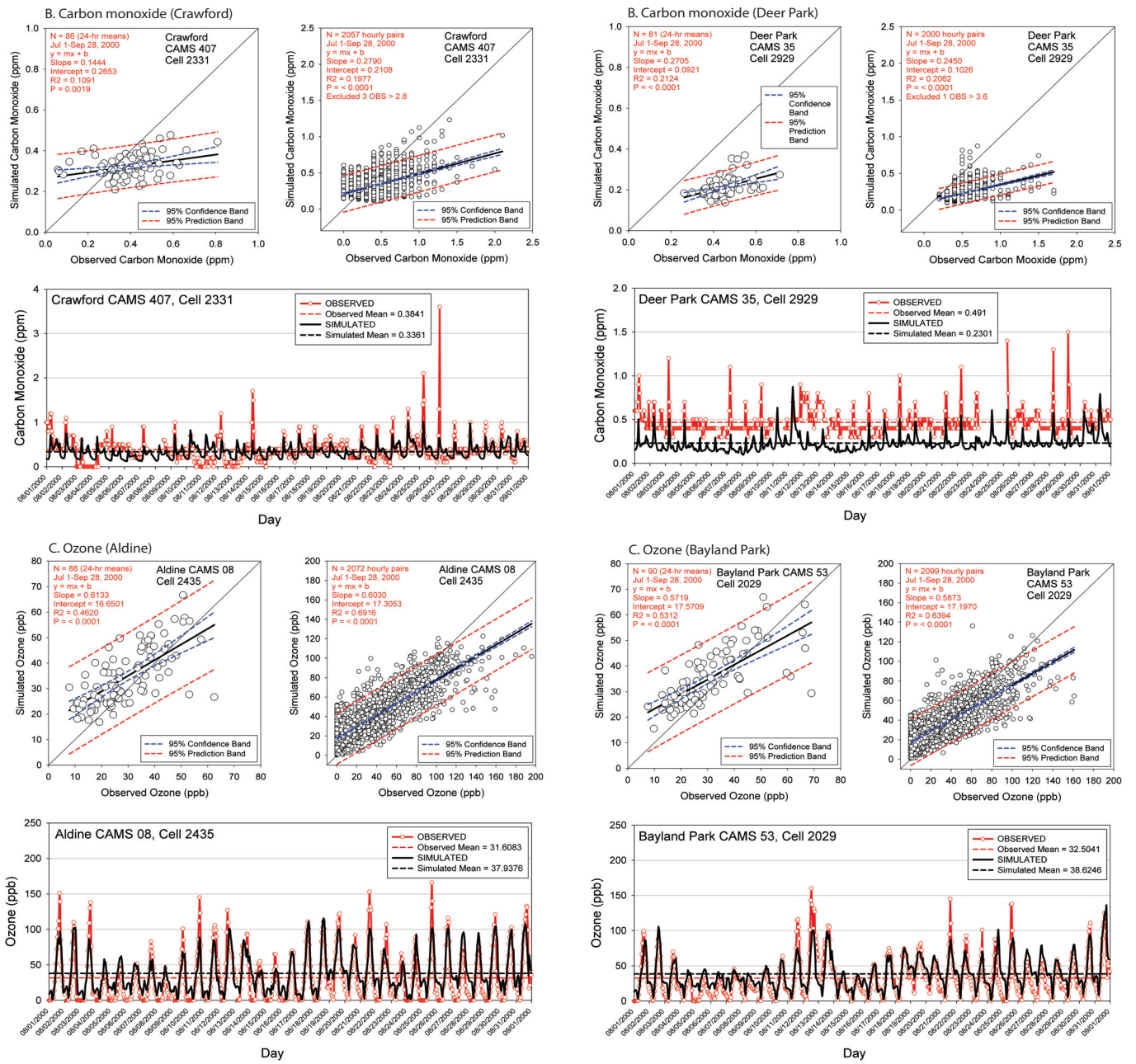


Figure 5 (cont.): Comparison of simulated and observed values. Shown are meteorological and pollutant variables that tended to remain in the multivariate linear mixed-level regression, as well as nitrogen oxides and benzene (a pollutant of special concern in the Houston-Galveston area). Of the air toxic species, only benzene had sufficient hourly observed measurements for comparison. For each, scatter plots of the 24-hour daily means and of the hourly pairs for the 90-day study period, as well as a time-series plot for August are shown. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Nitrogen oxides; and F. Benzene.

health effects studies, possibly as a surrogate, and ongoing concern about elevated benzene levels in the region. Because of the importance of good correlation between simulated and measured wind speed and direction in simulating pollutant species, we also compared simulated and observed wind speed and direction at individual Harris County monitors and for the region (Figure 6).

In Harris County, more monitors measure O_3 than any other pollutant. This is due in part to the fact that the HGA is in nonattainment for O_3. In addition to health-based

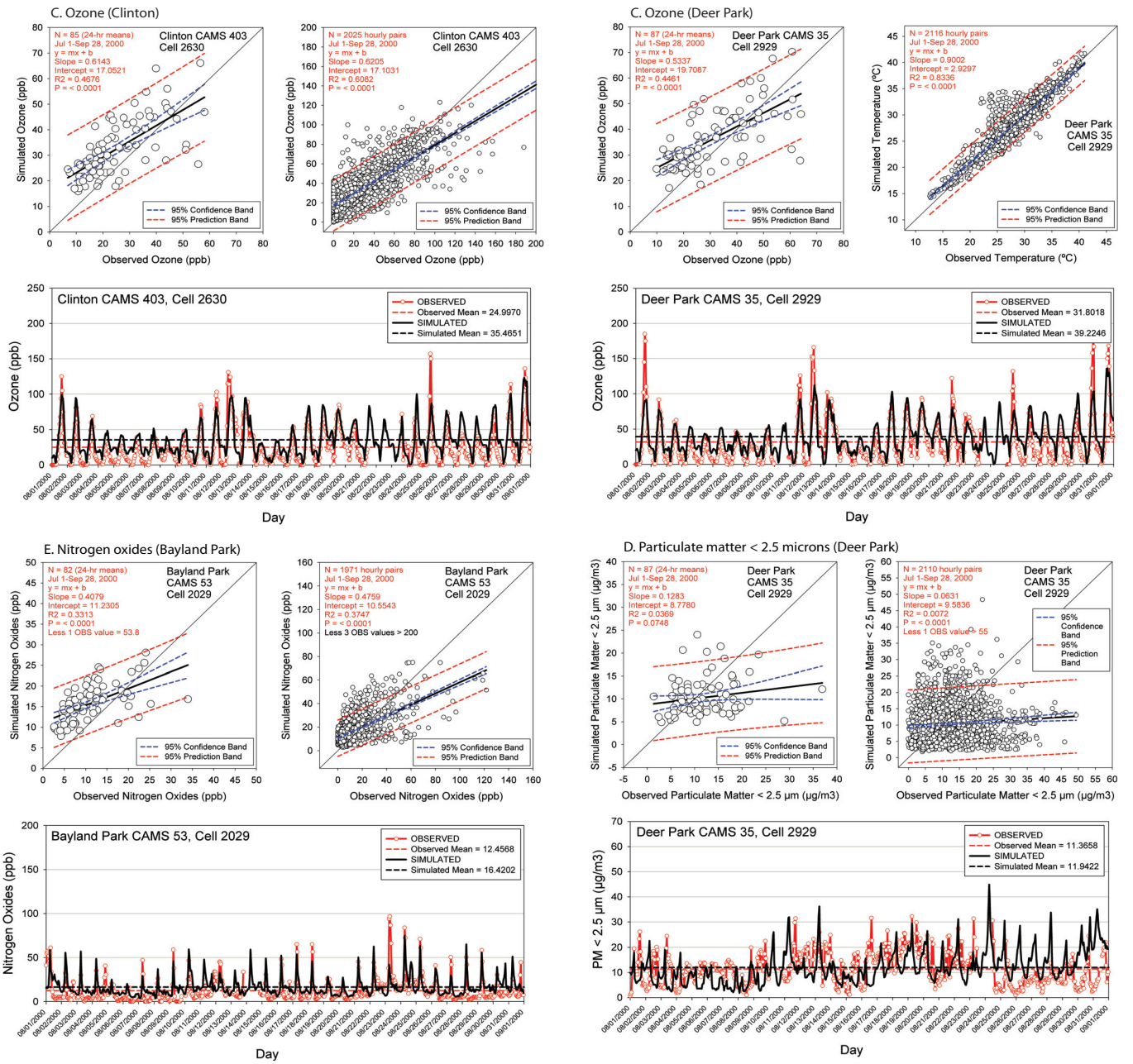


Figure 5 (cont.): Comparison of simulated and observed values. Shown are meteorological and pollutant variables that tended to remain in the multivariate linear mixed-level regression, as well as nitrogen oxides and benzene (a pollutant of special concern in the Houston-Galveston area). Of the air toxic species, only benzene had sufficient hourly observed measurements for comparison. For each, scatter plots of the 24-hour daily means and of the hourly pairs for the 90-day study period, as well as a time-series plot for August are shown. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Nitrogen oxides; and F. Benzene.

warnings, the monitors play an important role in testing O_3 simulations generated by CAMx and CMAQ. These simulations are critical in developing effective control strategies to achieve attainment of the O_3 standard. During the study period, 17 monitors in Harris County measured hourly concentrations of O_3. Figure 7 compares the simulated and observed 90-day means of the 24-hour daily means at these monitors and with the concentration in the underlying CMAQ cells. In general, and as is observed in the time-series and hourly scatted plots for ozone (Figure

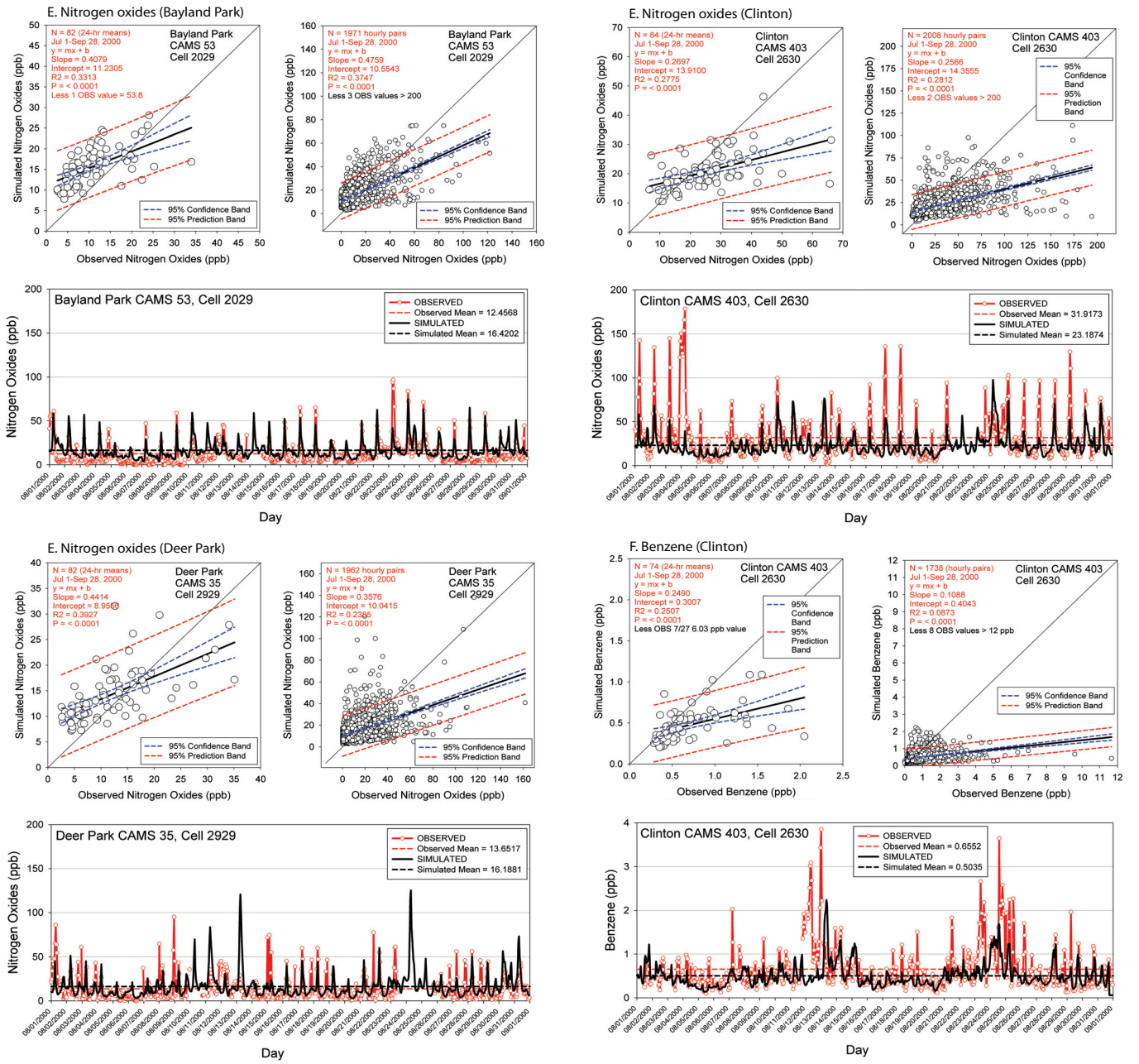


Figure 5 (cont.): Comparison of simulated and observed values. Shown are meteorological and pollutant variables that tended to remain in the multivariate linear mixed-level regression, as well as nitrogen oxides and benzene (a pollutant of special concern in the Houston-Galveston area). Of the air toxic species, only benzene had sufficient hourly observed measurements for comparison. For each, scatter plots of the 24-hour daily means and of the hourly pairs for the 90-day study period, as well as a time-series plot for August are shown. A. Temperature; B. Carbon monoxide; C. Ozone; D. Particulate matter < 2.5 microns in diameter; E. Nitrogen oxides; and F. Benzene.

5C), the agreement between hourly values is considerably higher than for the 24-hour means or for the 90-day mean of the 24-hour daily means. As is also addressed in the Discussion, CMAQ4.4 consistently demonstrates a nighttime bias with higher O_3 levels at night than are measured at the monitors, which largely accounts for the higher mean CMAQ values. In the spatial display shown in Figure 7, the observed mean O_3 values over the 90-day study period were, in general, consistently lower than the simulated concentrations. Only five of the 17 monitors had

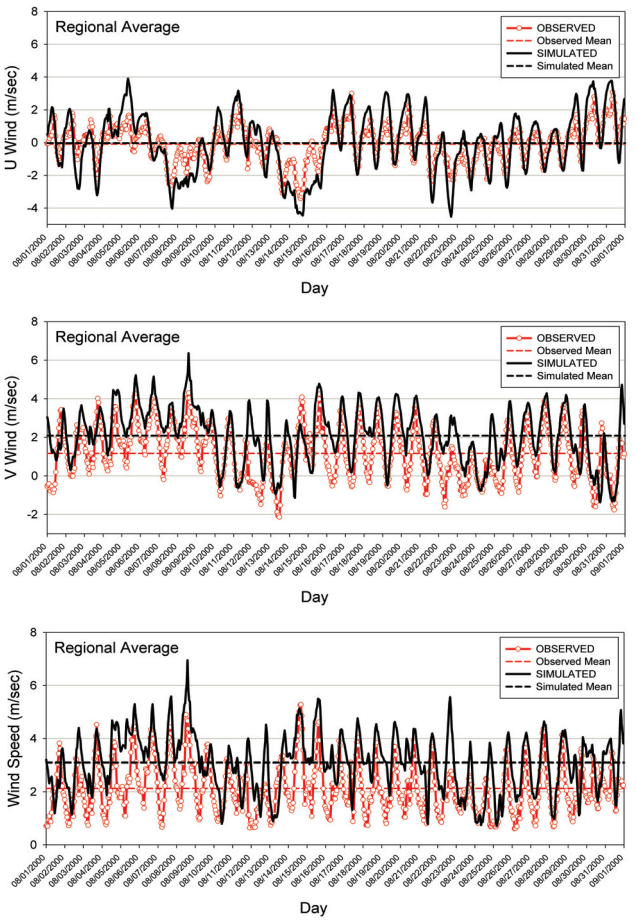


Figure 6: Comparison of regional wind direction and speed. Shown are time-series plots of simulated and observed values for wind direction (UWIND, VWIND) and wind speed (WS) for the region for August 2000.

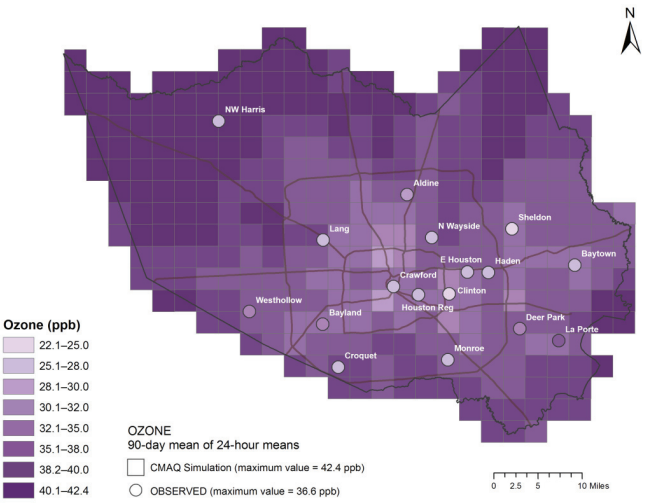


Figure 7: Comparison of CMAQ4.4-simulated and observed ozone concentrations. Shown are 90-day means of the 24-hour daily means of ozone during the study period, July 1 through September 28, 2000 CST, for all Harris County monitors (N = 17) with hourly ozone data during this period, and the same averaging for each of the 337 4 x 4-km cells.

sufficient validated data to compute daily means for the entire 90 days. The mapped comparison might be improved by adjusting the CMAQ means to match the days for which there was sufficient observed data, although that misrepresents the CMAQ spatial distribution used in the model. The use of a best fit model, which compares the observed value to the actual cell or any of the contiguous eight cells, would likely demonstrate better correlation. We also examined the ranking of the 17 areas of the two sets of values, which were not significantly correlated. A different averaging scheme (e.g., daytime only or maximum eight-hour running average) for O_3 may be warranted but, as noted earlier, introduces other problems as our knowledge about different lengths and windows of elevated exposures to different species is rudimentary. In general, except for temperature, relative humidity, and O_3, the observed data were insufficient to assess hourly performance or spatial variations in concentrations. Additional work is needed in this area, possibly using data mining tools, spatial interpolation techniques, and/or exposure modules to better assess meaningful exposure patterns and model performance.

Comparison with NATA ASPEN Model

We also compared concentrations for five selected HAPs (acetaldehyde, acrolein, benzene, 1,3-butadiene, and formaldehyde) simulated by our CMAQ4.4 model with the modified SAPRC99-ARO chemical mechanism, by cell, with the ASPEN-generated concentrations, by census track, prepared for the NATA analysis of air toxics risk (U.S. Environmental Protection Agency, 2007) (Figure 8 [A-E]). The CMAQ4.4 simulations are of the 90-day mean of the 24-hour daily means for July 1 through September 28, 2000; the ASPEN simulated values are annual means for 1999. The underlying emissions inventories are similar (CMAQ4.4 = TEI base5b imputed + NEI99 final version 3 + MOBILE6 vehicle emissions vs. NATA = NEI99 final version 3), although the inventories for the CMAQ4.4 simulation offer slightly more detailed information and resolution than does the NEI99 alone for the HGA. The spatial distribution of concentrations in the two models for the five HAP species are shown in deciles, relative to the highest value in each dataset, to compare the relative spatial distribution. The two models display underlying similarity, although the effect of secondary photochemical concentrations in the CMAQ4.4

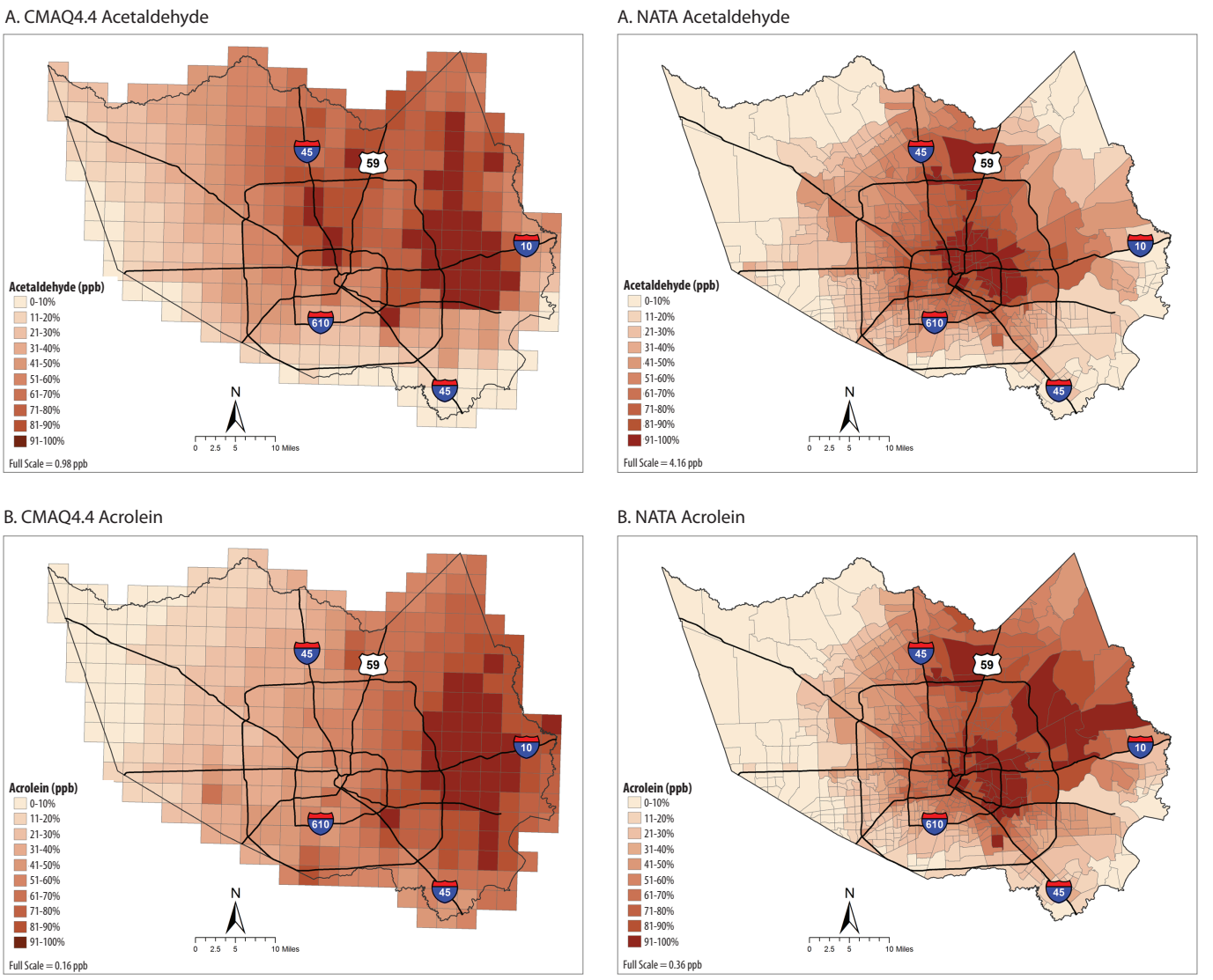


Figure 8: Comparison of CMAQ4.4- and ASPEN-simulated HAP concentrations. Shown are the mean 90-day mean of the 24-hour daily means concentrations of selected air toxics simulated by the CMAQ4.4 model with the modified SAPRC99-ARO mechanism (left) for July 1-September 28, 2000, by 4 x 4-km cell (N = 337), and the mean annual concentrations computed by the ASPEN dispersion model for year 1999 (right), by census tract (N = 649), for the EPA's National Air Toxics Assessment. A. Acetaldehyde; B. Acrolein; C. Benzene; D. 1,3-Butadiene; and E. Formaldehyde.

model output is apparent, especially for acetaldehyde, acrolein, and formaldehyde. In general (except for 1,3-butadiene), the ASPEN model simulated 1.6-4 times higher peak concentrations (e.g., full scale NATA benzene = 2.83 ppb vs. 1.74 ppb for CMAQ4.4). This would be expected since ASPEN, a dispersion model, will provide outputs more indicative of local-scale conditions as was discussed earlier. A comparison with annual maximum observed means from EPA's AirData database (U.S. Environmental Protection Agency, 2008) suggests that the values simulated by the CMAQ4.4 model were somewhat closer to the measured values. The peak annual mean benzene concentration, for example, at any Harris County monitor in 1999 was 1.50 ppb; in 2000, it was 1.23 ppb. The monitor locations, however, may or may not capture peak local values.

Hospital Admission Data

After obtaining Texas Department of State Health Services (TDSHS) and BCM Institutional Review Board (IRB) approvals, we purchased 61 fields from the Researcher Use

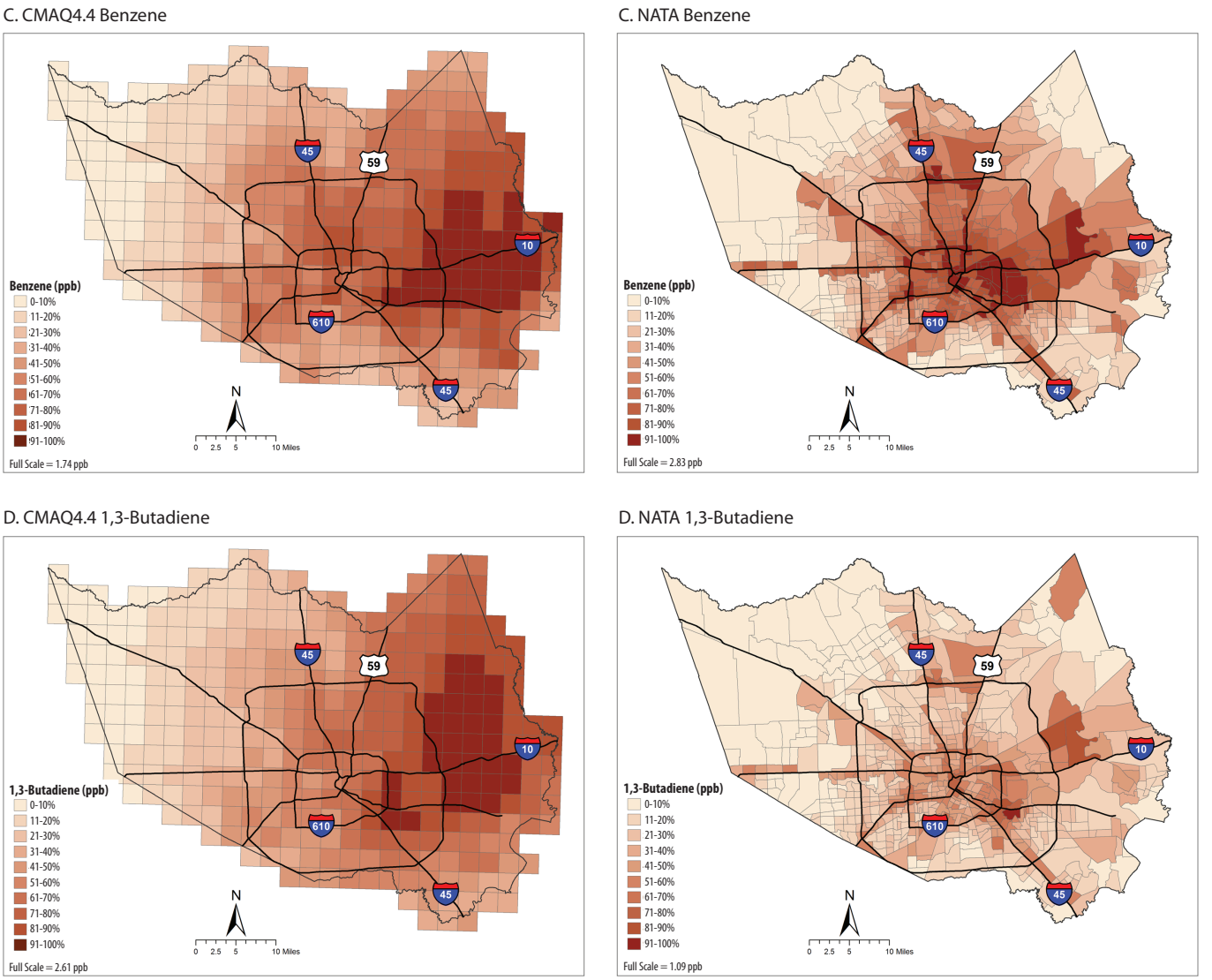


Figure 8 (cont.): Comparison of CMAQ4.4- and ASPEN-simulated HAP concentrations. Shown are the mean 90-day mean of the 24-hour daily means concentrations of selected air toxics simulated by the CMAQ4.4 model with the modified SAPRC99-ARO mechanism (left) for July 1-September 28, 2000, by 4 x 4-km cell (N = 337), and the mean annual concentrations computed by the ASPEN dispersion model for year 1999 (right), by census tract (N = 649), for the EPA's National Air Toxics Assessment. A. Acetaldehyde; B. Acrolein; C. Benzene; D. 1,3-Butadiene; and E. Formaldehyde.

Data File of the TDSHS Texas Health Care Information Collection (THCIC) Hospital Discharge Data for year 2000. The data fields included hospital name and address; patient gender, birth date, age, race, ethnicity, and employment status; residential address and marital status; employer name and address; facility type; date, time, and type of admission; length of stay; discharge status; payment source and total charges; principal ICD-9 diagnosis code plus eight additional ICD-9 codes for additional conditions; admitting diagnosis; and illness severity.

From the Texas hospital discharge data we extracted the records of all patients who were hospitalized between July 1 and October 6, 2000, and who listed a residence in Harris County (code 201) or a contiguous zip code–the latter to allow us to include persons who lived in the edge cells, partially in Harris County and partially in another county. There were initially 132,411 admissions that fit these criteria, although a substantial number of these patients lived in parts of the contiguous counties outside of our study area. To refine the cohort, we first determined our *a priori* exclusions. These included newborns (age 0, admitted 7/1-10/6/2000, ICD-9 V30-V39), accidents (ICD-9

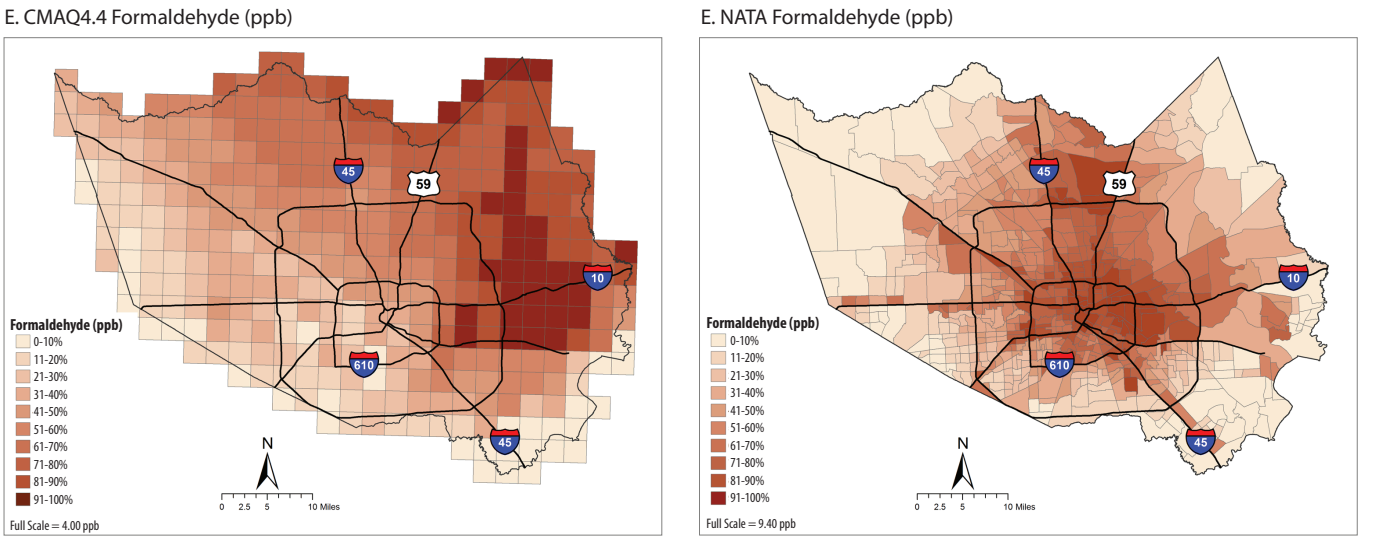


Figure 8 (cont.): Comparison of CMAQ4.4- and ASPEN-simulated HAP concentrations. Shown are the mean 90-day mean of the 24-hour daily means concentrations of selected air toxics simulated by the CMAQ4.4 model with the modified SAPRC99-ARO mechanism (left) for July 1-September 28, 2000, by 4 x 4-km cell (N = 337), and the mean annual concentrations computed by the ASPEN dispersion model for year 1999 (right), by census tract (N = 649), for the EPA's National Air Toxics Assessment. A. Acetaldehyde; B. Acrolein; C. Benzene; D. 1,3-Butadiene; and E. Formaldehyde.

E800-E999; E810-E819 = MVA), accidental poisonings (E860-E869), and various miscodings (e.g., age 100 + V30 [newborn]). The total number of *a priori* exclusions for the initial group of patients was 24,154, which included 22,742 newborns, 1,383 accidents, 17 newborns also involved in an accident, 10 accidental poisonings, and two miscodings. In addition, 126 records had no ICD-9 diagnosis in the primary discharge diagnosis field or in any of the eight "other" diagnosis fields. Thus the total for geocoding before trimming the edge cells was 108,257. The characteristics of the study cohort are shown in Table 4.

We then geocoded the patients with residences in the contiguous counties, focusing as much as possible on ZIP codes and streets that intersected the edge cells. Because an intensive effort to geocode all of the adjacent counties to access the relatively few admissions in the portion of the edge cells that abutted Harris County was beyond our resources, we chose to count as 100% successful our attempt for geocoding the partial edge cells outside Harris County. In the other-than-Harris-County edge cells, we geocoded 691 in Brazoria County (county code 039), 680 in Fort Bend County (code 157), 1,709 in Galveston County (code 167), 9 in Liberty County (code 291), 1,090 in Montgomery County (code 339), and 36 in Waller County (code 473)–for a total of 4,215. Thus our total adjusted number of hospital admissions was 94,012 in Harris County (code 201) and 4,215 from the non-Harris County portions of the edge cells, for an adjusted total number of 98,227. We included six additional days (October 1-6) of hospitalization data in our study in order to have this data for subsequent time-series or case-crossover study designs,

Table 4: Descriptive statistics of study cohort. Patients were included who listed a residential address in Harris County or an edge cell and who were admitted into a hospital in Harris County or a contiguous county between July 1 and October 6, 2000 (N = 98,227), following a priori exclusions.

Age	Mean 48.0 ± SD 23.0 year (range 0–109 year)			
	Age Group (year)	N	% of Total	% Female
	0-10	3,359	3.4%	43.2%
	11-20	7,691	7.8%	74.5%
	21-30	16,386	16.7%	84.8%
	31-40	14,473	14.7%	72.7%
	41-50	12,624	12.9%	56.0%
	51-60	11,066	11.3%	51.2%
	61-70	10,857	11.1%	53.1%
	71-80	12,721	13.0%	59.0%
	81-90	7,510	7.7%	66.9%
	91-100	1,505	1.5%	76.6%
	101-110	33	0.03%	90.9%
Gender	65.0% Female (N = 63,820); mean age 46.2 ± SD 23.2 year			
	35.0% Male (N = 34,379); mean age 51.4 ± SD 22.3 year			
	0.03% Not available (N = 28); mean age 57.0 ± SD 20.1 year			
Ethnicity	Hispanic origin		23.8%	22,370
	Not of Hispanic origin		77.1%	75,683
	Not available		0.2%	174
Race	American Indian		0.2%	206
	Asian or Pacific Islander		2.0%	1,929
	Black		22.0%	21,622
	White		56.4%	55,410
	Other		19.4%	19,017
	Not available		0.0%	43

Abbreviations: N = number; SD = standard deviation

in which we might wish to examine lags of 0 to 6 days depending on the health outcome being studied, using the same data set. For the current study design, which examines the potential effect of relatively long-term spatial differences in simulated pollutant exposure on the likelihood of adverse health effects, lag times are not relevant. Indeed, if the spatial distribution of chronic exposure to pollution as estimated by the 90-day CMAQ-simulated mean approximates the distribution across all of 2000 (i.e., the ranking of the 337 cells by pollutant is the same based on a 3-month or 12-month average), then we could potentially use admissions for all of 2000 in the current model to gain additional power. Limited resources for geoaddressing and the need to focus on key methodological questions in this pilot study restricted our initial hospital admissions cohort to 98 days.

The included hospitals were compared with lists from the THCIC and the Joint Council on Accreditation of Healthcare Organizations (JCAHO) for completeness in terms of reporting hospitals and percentage of beds. The dataset included data from 95 hospitals in the region (77.6% of hospitals and 96.0% of hospital beds). Two significant omissions included Lyndon B. Johnson General Hospital (LBJ; 324 beds) and Quentin Meese Hospital (75 beds). These hospitals are part of the Harris County Hospital District (HCHD), which supplies much of the medical care for the uninsured and underinsured in the region. We were not successful in obtaining 2000 discharge data directly from the HCHD for these two hospitals. Although both are general hospitals, LBJ has a large obstetrics program, with approximately 5,000 births each year, and Quentin Meese has fewer than 250 admissions per year. Thus we are hopeful that the omission of these two hospitals will have a relatively small effect on the study. Ben Taub General Hospital, the largest of the HCHD hospitals, was included in the data we received.

All patients were categorized by discharge diagnosis (Principal Discharge Diagnosis Code [ICD-9-CM] or Other Diagnosis Code 1, 2 or 3 [of eight]). The outcomes of primary interest for this study were, by discharge diagnosis, admission for cardiovascular, respiratory, or cardiovascular or respiratory (either) disease or symptoms. These diagnoses have been associated with ambient air pollution in other studies (Braga et al., 2001; Brunekreef and Holgate, 2002; Burnett et al., 1997; Burnett et al., 1997; Burnett et al., 1999; Dominici et al., 2006; Lee et al., 2007; Lee et al., 2007; Lee et al., 2006; Magas et al., 2007; Maheswaran et al., 2005; Maheswaran et al., 2005; Ricci and Straja, 2006; Schwartz, 1999; Thurston, 2006; Wellenius et al., 2005; Yang and Chen, 2007; Yang et al., 2004; Zanobetti et al., 2000). The "Other" category is useful as a possible control and for assessing possible bias. Because this category contains 17,335 new mothers, it is younger (mean age 37.2 vs. 64.5 years) and more female (72.6% vs. 57.9% female) than, for example, the cardiovascular group. Among the new mothers, however, it should be noted that 377 had either a cardiovascular or a respiratory diagnosis as well. A description of the four diagnostic groups follows.

- Cardiovascular (ICD-9 390-459, 745-747, 772-773, 776, 785, 790, 972, 986)
  N = 35,436 (36.1%, 57.9% F)
  Mean age 64.5 ± SD 17.1 years (range 1-109 years)

- Respiratory (ICD-9 460-519, 748, 769-770, 786, 975, 987)
  N = 17,137 (17.5%, 55.0% F)
  Mean age 58.1 ± SD 22.7 years (range 1-108 years)

- Cardiovascular or respiratory (either)
  N = 43,516 (44.3%, 55.4%F)
  Mean age 61.6 ± SD 19.8 years (range 1-109 years)

- Other
  N = 54,711 (55.7%, 72.6% F)
  Mean age 37.2 ± SD 19.3 years (range 1-104 years)

Geocoding of Hospital Admissions

We used ArcGIS, version 9.2 (ESRI, Redlands, CA), and two versions of the most detailed base map currently available for Harris County, the Southeast Texas Addressing and Referencing Map (STAR*Map; v. 2.0, 2004 and v. 4.0, 2006), for geocoding the addresses. For the July 1 through October 6, 2000 period, we were able to geocode (i.e., map by listed residential address) 84,729 (86.3%) patients out of the total of 98,227 who fit our inclusionary criteria; we were not able to geocode 13,498 patients. Within Harris County, there were 94,012 patients (95.7% of the 98,227). Of the 94,012 patients, 85.6% could be geoaddressed. As noted earlier, 4,215 (4.3% of total) patients were located in the non-Harris County portions of the edge cells. Because of the small number and the time involved in geoaddressing additional counties outside of our study area, we assumed these 4,215 to be 100% of the geoaddressable records in these small areas; thus our percentage of geoaddressed records, i.e., 86.3%, is probably very slightly overinflated.

The GIS Geocoding Service Address Locator gives a zero or low score if the ZIP code or the street name or the street address does not match. Relatively simple problems found with nongeocodable records, which were often resolvable

using other databases, included:

- street address did not match ZIP code;
- street number out of range (e.g., patient address is listed as 13500 Fannin when STAR*Map shows that Fannin street numbers range from 100 to 12000);
- street name incorrectly spelled (usually a typographical error);
- street name not found; and
- ZIP code not found.

Addresses that did not geocode using STAR*Map were reviewed for spelling or other errors, as well as searched for in other databases. We used online address locators, postal databases, the 1999 and 2006 editions of the *Houston Harris County Atlas Key Map* (Key Maps, Inc., Houston, TX), and street indexes to help with the geoaddressing. Most of the time, records were not picked by the geocoding engine due to small differences in street name spelling or a ZIP code that did not match the street address. ZIP codes were extracted from various online map sources (Google Maps, Yahoo Maps, and MapQuest) or compared with neighboring ZIP codes and edited when a reasonable match could be found. For approximately 2,500 records, we used Google Maps and Yahoo Maps to extract location coordinates directly from the Universal Resource Locator (URL) window, when a good match was found. These coordinates, generated in the Geographic Coordinate System (GCS) in decimal degrees, were then exported to ArcGIS for additional processing (creation of point layer and projection to the coordinate system used in the analysis). Records were also reviewed for other potential errors or concerns, such as the provider name accidentally being listed as the residence or a legitimate residence in a long-term hospital or nursing home. For out-of-range street numbers we geocoded to the nearest street number as long as the geocoded point fell within the same 4 x 4-km cell.

The 84,729 geocoded addresses are shown in Figure 9. Meticulous records were maintained, including the degree of geocoding accuracy (by percentage) and any address adjustments. Some of the problems encountered for the 13,548 records that we could not geocode included

- no address information;
- an address that was determined to be "not good" for various reasons and could not be resolved;
- patients listed as "unknown" address;
- patients listed as homeless;
- only a PO Box given; and
- incomplete information.

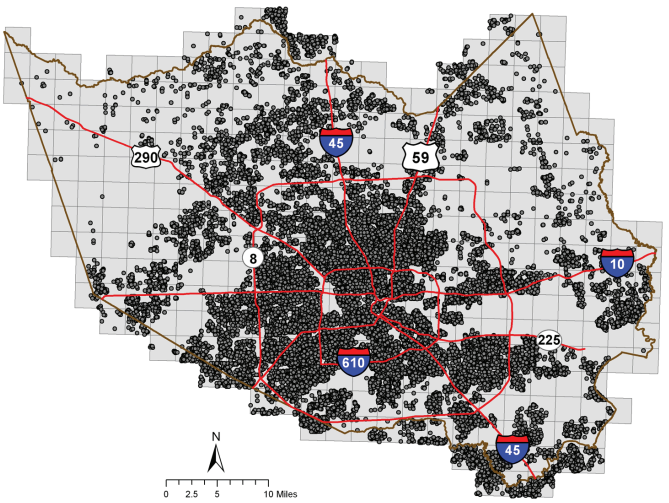


Figure 9: Geoaddressed hospital admissions. 84,729 (86.3%) were able to be geoaddressed by residential address.

Once geoaddressed, each record was assigned X (longitude) and Y (latitude) coordinates, initially using a freeware script. Since version 9.2, ArcGIS has the capability to assign XY coordinates as well. Use of STAR*Map and improvements in digital base maps for geoaddressing has facilitated extracting accurate XY coordinates, which were previously hampered in part by addressing algorithms that often positioned addresses based on beginning and end numbers of street segments, and side-of-street numbering conventions. Addition of XY coordinates reduces the potential for error, reduces computing time considerably, and produces a cleaner database for storage and/or review. Included in the database is information on the degree of accuracy of the match, and thus the accuracy of the XY coordinates. We aimed for exact matches when possible, followed by sufficient accuracy for future 1 x 1-km analyses.

Geocoded Records vs. Nongeocoded Records.

We compared the characteristics of those patients that we were able to geocode with those that we were not able to geocode to examine if any bias was introduced by the geocoding process. Based on individual characteristics from the THCIC database, the 84,729 geocoded records were, in general, statistically different from those records that did not geocode (13,498) (Table 5):

- Age: T-Test $P < 0.0001$
  Older individuals were more likely to geocode than younger individuals (48.3 vs. 46.6 years)

Table 5: Characteristics of the patients by geoaddressable status. Of the 98,277 patients, 84,729 (86.3%) could be geoaddressed to their residential address.

CHARACTERISTIC	GEOCODED N = 84,729 (%GC = 86.3%)	NOT GEOCODED N = 13,498 (%NGC = 13.7%)	P-VALUE	TOTAL (% of 98,227)
Age (years)	48.3 ± SD 23.1	46.6 ± SD 22.4	< 0.0001[T]	48.0 ± SD 23.0
Gender	55,430 (86.8%)	8,390 (13.2%)	< 0.0001[C]	63,820 (65.0%)
	29,274 (85.2%)	5,105 (14.9%)		34,379 (35.0%)
Race	18,386 (85.0%)	3,236 (15.0%)	< 0.0001[C]	21,622 (22.0%)
	47,928 (86.5%)	7,482 (13.5%)		55,410 (56.4%)
	1,664 (86.3%)	265 (13.7%)		1,929 (02.0%)
	177 (85.9%)	29 (14.1%)		206 (00.2%)
	16,537 (87.0%)	2,480 (13.0%)		19,017 (19.4%)
Cardiovascular	30,960 (87.4%)	4,476 (12.6%)	< 0.0001[C]	35,436 (36.1%)
Respiratory	14,916 (87.0%)	2,221 (13.0%)	0.0011[C]	17,137 (17.5%)
Either	37,973 (87.3%)	5,543 (12.7%)	< 0.0001[C]	43,516 (44.3%)

Abbreviations: [C] = Chi-Square; Either = cardiovascular or respiratory; GC = geocoded; N = number; NGC = not geocoded; [T] = Student T-Test

- Gender: Chi-square $P < 0.0001$
  Females geocoded more often than did males (86.8% female vs. 85.2% male)

- Race: Chi-square $P < 0.0001$
  Whites (86.5%) and Asians (86.3%) were more likely to geocode than Blacks (85.0%).

- Discharge diagnosis of cardiovascular: Chi-square $P < 0.0001$
  Individuals with a cardiovascular diagnosis (N = 30,960) were more likely to geocode than individuals with other diagnoses (87.4% vs. 85.6%).

- Discharge diagnosis of respiratory: Chi-square $P = 0.0011$

Individuals with a respiratory diagnosis (N = 14,916) were more likely to geocode than individuals with other diagnoses (87.0% vs. 86.1%).

These findings suggest that, in general, the patients in our statistical model likely slightly underrepresented persons of color, who tend to have a lower socioeconomic status. Since our "hot spot" analyses seek to delineate areas of where residents are likely to be more vulnerable to the effects of elevated pollution, which often correlates with lower socioeconomic status, underrepresentation of persons of color may dilute our findings (regression towards the null). Thus, although this bias is of concern, in context of our hypothesis it is unlikely to lead to a type 1 error, i.e., a determination of "hot spots" of elevated "exposure" and health effects when there are none.

Demographic Data

Using Census 2000 data obtained from the U.S. Census Bureau (http://factfinder.census.gov), we obtained block and block-group data for Harris County. In general, three types of data are available:

- Summary File 1 (100% data)
  Actual counts and information, such as age, sex, race, ethnicity, and whether residence is owned or rented, collected from individual surveys.

- Summary File 2 (100% data)
  Population and housing characteristics iterated for different races and ethnicities.

- Summary File 3 (sample data)
  Detailed population and housing data, such as place of birth, education, employment status, commuting, income, and year structure built, collected from a 1-in-6 sample and weighted to represent the total population.

SF1 data are available at the block level, the highest level of resolution available (approximately 14 households). Income, education, and commuting status are available at the block-group level.

Because we were interested in calculating the demographic characteristics for each of the 337 4 x 4-km cells in Harris County, we needed to calculate this information from the appropriate files at the block or block-group level of resolution. Once all the data were downloaded, they were joined to block and block-group spatial data files (shapefiles or personal geodatabase files) using a unique identifier, and then imported into our ArcGIS map of Harris County using the appropriate projection. We then proceeded with apportioning the data to the 337 4 x 4-km cells.

Apportioning Data to Cells

Apportioning data of a different shape (e.g., blocks) to another shape (e.g., a 4 x 4-km cell) on a different mapping layer requires adding appropriate percentages of data from data blocks or block groups that are transected to the data wholly encompassed by the "cookie cutter" using area determinations. A Visual Basic for Applications (VBA) script was used to calculate the area for each block, block group, and water polygon.

The basic spatial processing steps were performed using ArcGIS 9.2 and apply to both blocks and block groups as

well as to other Census 2000 variables. They are briefly described below.

1. Calculate *block* polygon areas and store values in a new field (*block_sqft*).
2. Intersect *block* and *water* polygons and delete intersected water polygons, resulting in "*land blocks*."
3. Recalculate *land-block* polygon areas and store values in a new field (l*and_block_sqft*).
4. Intersect the *land_block_sqft* and 4 x 4-km *cell_polygon* layers to create *land_cell_block*.
5. Recalculate *land-cell blocks* polygons and store values in a new attribute field (*land-cell_block_sqft*).
6. Add a new field (*percent_land_area*) and calculate *percent_land_area*:

$$\text{percent\_land\_area} = \frac{\text{land\_cell\_block\_sqft}}{\text{land\_block\_sqft}}$$

where

land_cell_block_sqft = land block area in square feet inside each 4 x 4-km cell after intersection with cell polygons

land_block_sqft = original land block area in square feet before intersection with cell polygons

percent_land_area = percent land area inside each cell polygon

7. Calculate population for each land-cell block polygon:

$$cell\_pop2000 = pop2000 * percent\_land\_area$$

where

pop2000 = total population from Census 2000

cell_pop2000 = population for the land-cell block polygon

8. Use the summarize function to calculate total block population and land block area per cell (*cell_pop2000* field).
9. Perform a table join between the 4 x 4-km *cell_polygon* layer and summarized table.
10. Create two new attribute fields to store population per cell (*scell_pop2000*) and block area per cell (*scell_block_sqft*) from joined table.
11. Add new attribute field to store population density per cell (*pop_dens*). Calculate population density per cell polygon as follows:

$$pop\_dens = \frac{scell\_pop2000}{scell\_block\_sqft}$$

12. Remove table join.

To calculate the average median income of each 4 x 4-km cell, an algorithm developed by the University of Arizona was used. First, the weighted average of all the cropped areas of block groups in each 4 x 4-km cell was calculated. For instance, if there were median incomes of $65,000, $48,000, and $75,000 for block groups having 2,500, 3,200, and 450 households respectively, the weighted average median income was calculated as:

$$\text{Average median income} = \frac{(65{,}000 x 2{,}500 + 48{,}000 x 3{,}200 + 75{,}000 x 450)}{(2500 + 3200 + 450)} = 56{,}886.17$$

In other words, average median income equals

$$= \frac{\sum (mi * hh)}{\sum (hh)}$$

where

mi = median income of the cropped block group, and

hh = number of household of the cropped block group.

Calculation of census demographic characteristics by 4 x 4-km cells was very time-consuming process. There are approximately 15-20 steps for each demographic field, and a characteristic such as age has 46 category fields (23 age categories, by gender). Done manually, it is easy to commit errors, and the process for extracting a demographic characteristic such as age can take up to one week to complete. The Baylor College of Medicine (BCM) team developed two automated programs (for block and block-group data) using python scripts that have been developed for ESRI's GIS Model Builder interface. These scripts automated data extraction and apportionment for our pilot study. The custom automated program took approximately 10 minutes per field for preparation, and then approximately 9 hours of unsupervised computer time to extract the data and do the calculations.

Demographic Characteristics

Demographic data apportioned to each of the 337 4 x 4-km cells for subsequent inclusion in the statistical model included the following variables, briefly described below. The level of resolution is shown in parentheses.

- Nighttime population count (block)
  Mean 14,116 ± SD 14,327; median 8,672; range 77-74,895
- Nighttime population density (per square mile) (block)
  Mean 2,316 ± SD 2,318; range 12-12,172
- Percentage non-White (block)
  Mean 42.3% ± SD 26.5%; median 35.0%; range 5.1%-99.5%

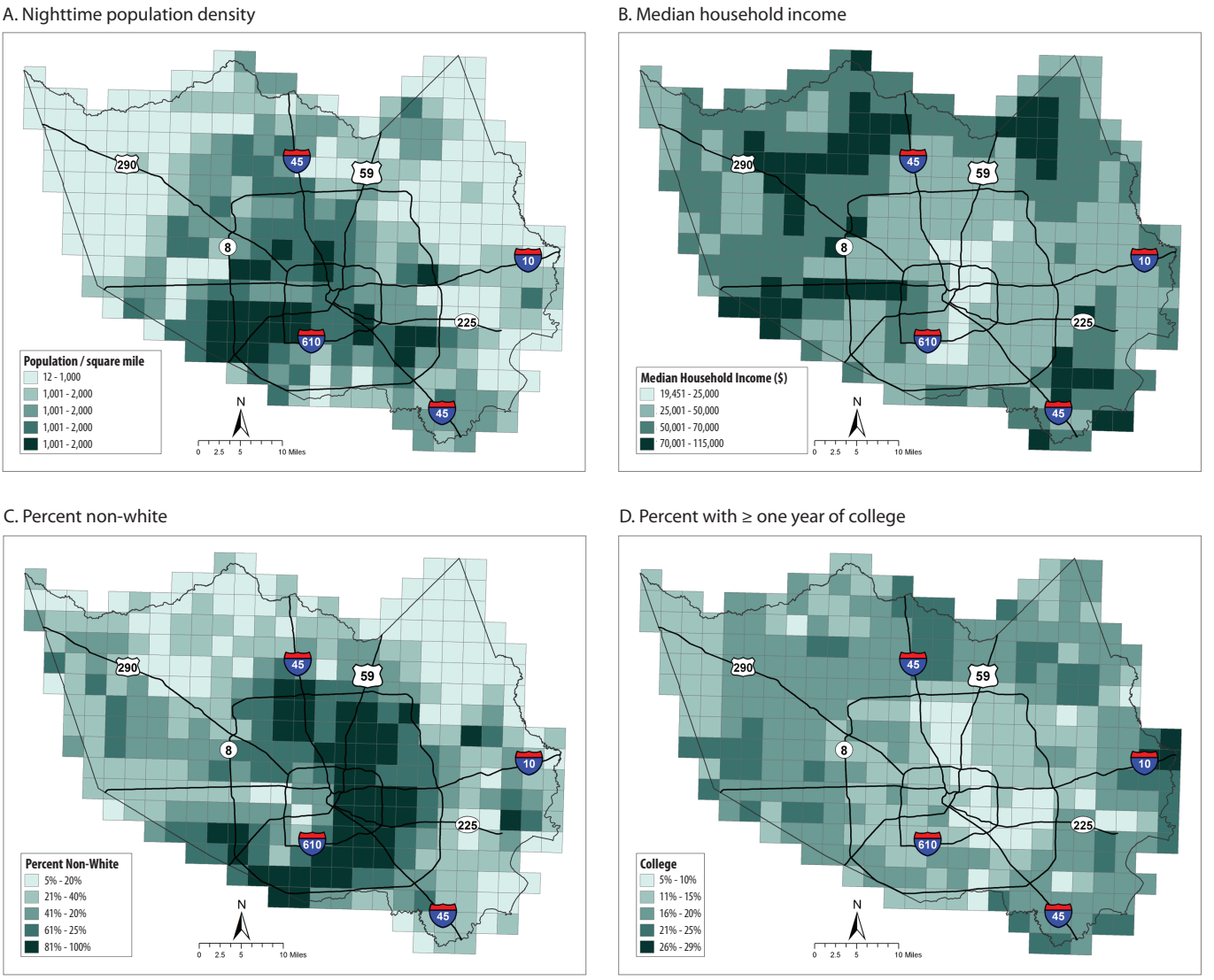


Figure 10: Demographic variables by 4 x 4-km cell. Shown are A. Nighttime population density; B. Median household income; C. Percent non-white; D. Percent with ≥ one year of college; E. Percent who commute 30 minutes or more one way to work daily; and F. Percent owner-inhabited housing.

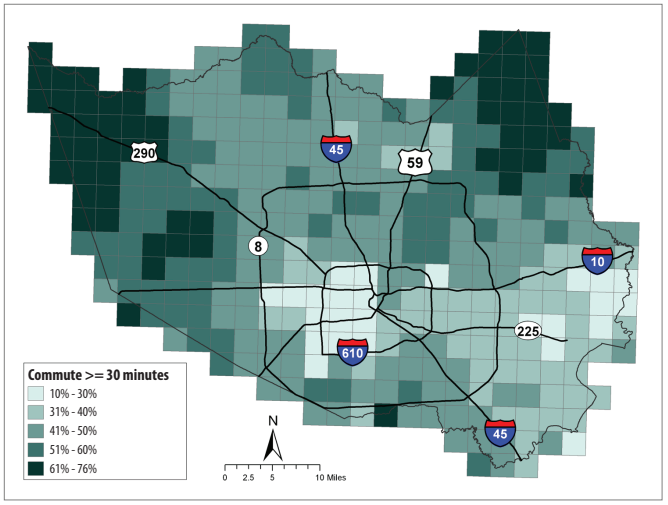
- Percent owner-inhabited housing (block)
  Mean 70.7% ± SD 19.6%; median 75.1%; range 4.7%-99.2%
- Median household income (block group)
  Mean $52,293 ± SD $16,985; median $52,119; range $19,450-$114,238
- Percent with ≥ one year of college (block group)
  Mean 20.0% ± SD 15.9%; median 17.7%; range 0%-80.5%
- Percent ≥ 16 yr who commute 30 or more minutes to work (block group)
  Mean 36.0% ± SD 23.2%; median 37.8%; range 0%-96.4%

Maps of the demographic variables by 4 x 4-km cell are shown in Figure 10 (A-F).

DATA ANALYSIS

As noted earlier, the hypothesis being tested is that the age-adjusted rates, by discharge diagnosis, of Harris County residents hospitalized during the study period differ geographically among the 337 4 x 4-km cells that overlay the county and correlate with MM5-, CMAQ4.4- and CMAQ-HAP-simulated meteorological values and/or concentrations of ambient air pollutants, controlling for

E. Commute ≥ 30 minutes to work



F. Owner-inhabited housing

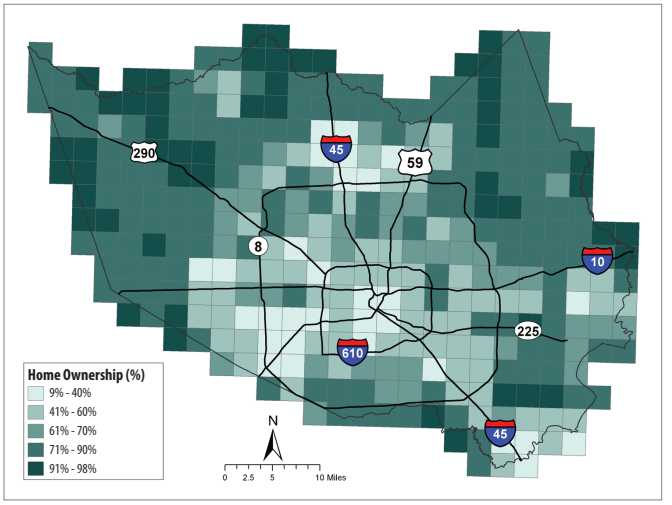


Figure 10 (cont.): Demographic variables by 4 x 4-km cell. Shown are A. Nighttime population density; B. Median household income; C. Percent non-white; D. Percent with ≥ one year of college; E. Percent who commute 30 minutes or more one way to work daily; and F. Percent owner-inhabited housing.

other variables.

The study design is a mixed-effect ecological correlation analysis in which the unit of analysis is the 337 4 x 4-km cells. The cohort (population at risk) was residents of any of the 337 4 x 4-km cells overlaying Harris County as determined by Census 2000 data (U.S. Census Bureau, 2000); cases equaled those who listed a residence in any of these 337 cells and who were admitted to a hospital in Harris County or any of the contiguous counties between July 1 and October 6, 2000. The outcome (dependent variable) of interest, as determined by discharge diagnosis, was cardiovascular disease, respiratory disease, or cardiovascular or respiratory disease (either). Independent (predictor) variables (N = 32) examined included seven demographics factors, two meteorological variables, five CAPs, 16 HAPs, and two PCA factors (see description of PCA later in this section) to represent the 16 HAPs.

Statistical Methods

SAS 9.1 was used for the majority of the statistical work, with some ancillary work done using Microsoft Excel and Access, SPSS, ESRI's ArcGIS and Geostatistical Analyst, and R. Our initial statistical work, prior to building the models, included extensive examination and validation of the raw data, creation of numerous secondary databases that would be needed to build the models, and validation of the secondary databases and their associated scripts. In most instances, the databases and SAS scripts were reviewed by at least two others on the team, in addition to statisticians Drs. Chan and Han. Refinements and/or corrections were made as appropriate. Part of the initial statistical work included extensive visual and statistical exploration of the admissions discharge records, Census 2000 data and collection methods, simulated meteorological values and pollutant concentrations, and measured pollutant concentrations. We also produced and examined a wide array of descriptive statistics and graphs; developed scripts for inputting the immense amount of CMAQ-simulated data (for each species we had 727,920 hourly values), converting UTC to CST, identifying newborns, creating secondary databases with various averaging times, testing the discharge diagnosis codes and extraction scripts for accuracy and completeness, and assessing other aspects of the statistical scripts and data; examined exclusions and geocoded vs. unable-to-geocode patients for possible bias; and explored different ways to handle the 16 HAPs in the model, including the use of ranking, PCA to reduce the dimensionality of the HAPs, and inclusion of all or of representative HAPs based on chemical properties, correlation, or other factors.

After extensive assessment of the data and several unsuccessful evaluations of potential models, we chose the statistical model for the analyses. A number of factors in this spatial analysis, not least of which were factors introduced by the gridded infrastructure, made choosing the model more challenging than we had anticipated. We first attempted using the Generalized Estimating Equation (GEE) method, in which we assumed that the number of hospital admissions in each cell was distributed as a

Poisson random variable. For this analysis, the link was log-transformed and the offset variable was the population in each cell. This approach, however, was computationally unstable, most likely because of the marked differences in population size across the 337 cells, as well as because of extreme rare events in a few cells. The result was that the models did not converge computationally, and this approach was abandoned.

Consequently, we decided to use the log-transformed age-adjusted rate for each gender group as the outcome variable and a LMM to examine the relationship between hospital admissions and simulated air pollution ("exposure"), adjusting for significant gender and cell-specific demographic factors. The 337 4 x 4-km cells were the unit of analysis. For this spatial analysis, we used a two-part approach. The mixed-effects regression model takes into consideration correlated outcome (i.e., it is a correlated-outcome model). By taking into account the correlated nature of the outcome, this method accommodates "missing" observations due to extreme age-adjusted rates. We therefore first created a specified correlation structure for outcome for each of the 337 4 x 4-km cells. This was followed by examining the spatial distributions of the residuals obtained from the fitted models.

Age-Adjusted Rates by Gender

Next we calculated the age-adjusted rates for each of the 337 cells. This was necessary to remove any spatial differences in hospital admissions that might be attributable to residents in some cells being older (or younger). Based on recent available Census 2000 age categories and reasonableness, we formed from the 23 Census 2000 age categories nine groups that corresponded to 8 of the 11 categories we had used to describe the admissions data (the top three categories were combined, as census data do not categorize after age 80) by gender: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80 years of age and older. The age-adjusted rates were then calculated as follows.

- For each of the 337 cells, we calculated the crude admission rate (using Census 2000 counts extracted for each cell as the denominator) for each age group and gender for
  - admissions with a cardiovascular discharge diagnosis,
  - admissions with a respiratory discharge diagnosis, and
  - admissions with either a cardiovascular or respiratory discharge diagnosis.
- We then multiplied the crude admission rate for each age and gender group by the total Census 2000 population for the 337 cells by age category and gender. This resulted in 18 "adjusted counts" for each cell (nine age groups by gender).
- For each cell, by gender, we added all of the adjusted counts, and divided this sum by the total Census 2000 male or female population for all 337 cells. The resulting ratio is the age-specific admission rate for each cell by gender.

Small-Cell Adjustment

As noted above, to develop the age-adjusted rate, we used nine age categories, by gender, resulting in 18 categories for each of the 337 cells. However, some of the 337 cells had small populations, which could lead to bias. We therefore did the following after extensive examination of the data:

_ Ranked the 337 cells by total population. In the lowest 5th percentile, 17 cells had a population less than 127.
_ Combined each of the 17 small-population cells with the most populous contiguous cell. By column-row number, this resulted in the following double-cell combinations (and one triple): 2931-2932; 3130-3129; 1336-1337; 1732-1731; 1134-1133; 1235-1135-1236; 3429-3430; 3042-2942; 1436-1437; 1137-1237; 1136-1036; 3031-3032; 1234-1233; 940-939; and 1335-1435.
_ Within each cell or cell combination, we dropped the admission rate of any of the 18 age-gender groups if the number in the group was °‹ 5.
_ Using our averaging schema of interest, i.e., the 90-day mean of the 24-hour daily means, we (1) ranked and inspected each cell for age-adjusted outliers for rate of admission for cardiovascular, respiratory, or cardiovascular or respiratory (either) disease, and (2) dropped age-gender categories defined as outliers (by rate or multiple admissions). This led to our dropping nine specific age-groups by gender subgroups from a total of five 4 x 4-km cells. All of these subgroups had a population < 27 and the majority tended to be older age groups in which several individuals had multiple admissions. The 4 x 4-km cells, except for these small adjustments, remained intact.

Correlation Between Variables

As part of the data exploration, as well as a component of the development of PCA factors, we examined the correlation between each of the 16 HAPs, as well as the correlation between selected CAPs, HAPs, and demographic variables. Pearson's correlation coefficients for

Table 6: Pearson's correlation coefficient matrix. Shown here are 16 HAPs simulated by the CMAQ4.4 and CMAQ-HAP models for July 1 to September 28, 2000, in Harris County, Texas. The averaging schema used is the 90-day mean of the 24-hour daily means for the 337 cells, the averaging schema used for this pilot study. The strength of correlation between two variables is reflected in the size of the coefficient and in the P-value.

	CCHO	ACET	ACROL	BENZ	BUTA	CHCL	CRES	BRC	CLC	ETOX	HCHO	CLME	CL4ET	PHEN	CL3ET	CLET
CCHO	1.000	0.316‡	0.595‡	0.607‡	0.628‡	0.317‡	0.739‡	0.430‡	0.436‡	0.080	0.822‡	0.360‡	0.407‡	0.551‡	0.192†	0.443‡
ACET	0.319‡	1.000	0.084	-0.079	0.158*	0.209†	-0.227‡	0.033	0.056	-0.247‡	0.481‡	-0.353‡	-0.173*	0.029	-0.163*	0.048
ACROL	0.595‡	0.084	1.000	0.814‡	0.916‡	0.309‡	0.528‡	0.541‡	0.529‡	0.290‡	0.635‡	0.117*	-0.103	0.933‡	0.313‡	0.589‡
BENZ	0.607‡	-0.079	0.814‡	1.000	0.699‡	0.230‡	0.756‡	0.480‡	0.550‡	0.228‡	0.697‡	0.314‡	0.073	0.563‡	0.342‡	0.545‡
BUTA	0.628‡	0.158*	0.916‡	0.699‡	1.000	0.330‡	0.556‡	0.471‡	0.459‡	0.094	0.594‡	0.113*	-0.010	0.856‡	0.183†	0.520‡
CHCL	0.317‡	0.209†	0.309‡	0.230‡	0.330‡	1.000	0.157*	0.209†	0.152*	-0.015	0.328‡	0.005	-0.013	0.220‡	0.051	0.193†
CRES	0.739‡	-0.227‡	0.528‡	0.756‡	0.556‡	0.157*	1.000	0.437‡	0.449‡	0.154†	0.578‡	0.568‡	0.481‡	0.562‡	0.336‡	0.429‡
BRC	0.430‡	0.033	0.541‡	0.480‡	0.471‡	0.209†	0.437‡	1.000	0.678‡	0.075	0.485‡	0.044	-0.078	0.545‡	0.165†	0.789‡
CLC	0.436‡	0.056	0.529‡	0.550‡	0.459‡	0.152*	0.449‡	0.678‡	1.000	0.127†	0.579‡	0.039	-0.119†	0.505‡	0.258‡	0.905‡
ETOX	0.080	-0.247‡	0.290‡	0.228‡	0.094	-0.015	0.154†	0.075	0.127†	1.000	0.083	0.260‡	-0.005	0.237‡	0.406‡	0.086
HCHO	0.822‡	0.481‡	0.635‡	0.697‡	0.594‡	0.328‡	0.578‡	0.485‡	0.579‡	0.083	1.000	0.106	0.024	0.585‡	0.274‡	0.540‡
CLME	0.360‡	-0.353‡	0.117*	0.314‡	0.113*	0.005	0.568‡	0.044	0.039	0.260‡	0.106	1.000	0.686‡	0.102	0.246‡	0.023
CL4ET	0.407‡	-0.173*	-0.103	0.073	-0.010	-0.013	0.481‡	-0.078	-0.119†	-0.005	0.024	0.686‡	1.000	-0.085	0.023	-0.117†
PHEN	0.551‡	0.029	0.933‡	0.786‡	0.856‡	0.220‡	0.562‡	0.545‡	0.505‡	0.237‡	0.585‡	0.102	-0.085	1.000	0.245‡	0.594‡
CL3ET	0.192†	-0.163*	0.313‡	0.503‡	0.183†	0.051	0.336‡	0.165†	0.258‡	0.406‡	0.274‡	0.246‡	0.023	0.245‡	1.000	0.200†
CLET	0.443‡	0.048	0.589‡	0.545‡	0.520‡	0.193†	0.429‡	0.789‡	0.905‡	0.086	0.540‡	0.023	-0.117†	0.594‡	0.200†	1.000

Abbreviations (in order by table listing): CCHO = acetaldehyde; ACET = acetone; ACROL = acrolein; BENZ = benzene; BUTA = 1,3-butadiene; CHCL = chloroform; CRES = cresols; BRC = ethylene dibromide; CLC = ethylene dichloride; ETOX = ethylene oxide; HCHO = formaldehyde; CLME = methylene chloride; CL4ET = perchloroethylene; PHEN = phenols; CL3ET = trichloroethylene; CLET = vinyl chloride; CMAQ = Community Multiscale Air Quality; CMAQ-HAP = CMAQ adapted for selected gas-phase HAPs; HAP = hazardous air pollutant; * = P-value < 0.05; † = P-value < 0.001; ‡ = P-value < 0.0001

these two sets of variables are shown in Tables 6 and 7. The degree of correlation between variables tended to be high.

Principal Components Analysis

We explored the use of PCA to identify a small number of factors that would explain much of the variance, i.e., the pattern of correlation, in the 16 HAP variables in our study. By definition, each component (PCA factor) defines a projection that encapsulates the maximum amount of variation in a dataset and is orthogonal (and therefore uncorrelated) to the previous principal component of the same dataset. After viewing the shape of the scree plot of the eigenvalues, we chose two factors, each with eigenvalues > 2, to represent the HAP set. The choice is somewhat subjective, in our case based largely on the gap between the second and third factors, a decision to include only components that represented > 10% of the total variance, and a desire to reduce the number of variables in the full statistical model. The two factors selected accounted for 42.9% (eigenvalue 6.9) and 15.1% (eigenvalue 2.4), respectively, of the variability of the 16 components examined, for a cumulative total of 57.9%. Inclusion of factors with eigenvalues above > 1 would have retained two additional factors, with the four factors accounting for 76.0% of the variability. Future work with this dataset might wish to include four factors. Varimax rotation of the matrix was selected to enhance the interpretability of the factors. Although the unrotated and rotated factors explain the same total amount of variation, varimax rotation rotates the orthogonal principal component axes so that the variability within the data set explained by each axis is maximized (Dunteman, 1989). The unrotated and rotated factors are shown in Table 8.

Univariate Linear Mixed-Effects Model

We used the LMM to examine the individual effect of each of the potential predictors variables on our outcomes of interest, age-adjusted hospital admission rates by discharge diagnosis for each of the 337 cells for Due to the skewed nature of the outcome variable, the age-adjusted

Table 7: Pearson's correlation coefficient matrix. Shown here are selected CMAQ4.4- and CMAQ-HAP-simulated HAPs, CAPs and demographic variables in Harris County, Texas. The averaging schema used is the 90-day mean of the 24-hour daily means for the 337 cells, the averaging schema used for this pilot study. The strength of correlation between two variables is reflected in the size of the coefficient and in the P-value.

	CCHO	ACET	BENZ	BRC	HCHO	PHEN	CO	0_3	NO_x	$PM_{2.5}$	SO_2	M-INC	%MIN	EDUC	OWN	POP
CCHO	1.000	0.316‡	0.607‡	0.430‡	0.822‡	0.551‡	0.420‡	-0.396‡	0.652‡	0.551‡	0.502‡	-0.248‡	0.083	0.031	-0.102	0.051
ACET	0.316‡	1.000	-0.079	0.033	0.481‡	0.029	-0.367‡	0.464‡	-0.359‡	0.204†	-0.150*	0.205†	-0.435‡	-0.257‡	0.449‡	-0.446‡
BENZ	0.607‡	-0.079	1.000	0.480‡	0.697‡	0.786‡	0.159*	-0.590‡	0.626‡	0.130*	0.511‡	-0.266‡	0.166*	0.026	-0.170*	0.012
BRC	0.430‡	0.033	0.480‡	1.000	0.480‡	0.545‡	-0.035	-0.247‡	0.296‡	0.007	0.365‡	-0.192†	0.029	-0.095	0.016	-0.115†
HCHO	0.822‡	0.481‡	0.697‡	0.485‡	1.000	0.585‡	-0.020	-0.166*	0.359‡	0.315‡	0.401‡	-0.112*	-0.160*	-0.099	0.141*	-0.270‡
PHEN	0.551‡	0.029	0.786‡	0.545‡	0.585‡	1.000	-0.021	-0.458‡	0.462‡	0.008	0.569‡	-0.179†	0.066	-0.039	0.012	-0.137†
CO	0.420‡	-0.367‡	0.159*	-0.035	-0.020*	-0.021	1.000	-0.728‡	0.820‡	0.573‡	0.263‡	-0.323‡	0.520‡	0.323‡	-0.602‡	0.740‡
0_3	-0.396‡	0.464‡	-0.590‡	-0.247‡	-0.166†	-0.458‡	-0.728‡	1.000	-0.926‡	-0.315‡	-0.483‡	0.394‡	-0.507‡	-0.208†	0.525‡	-0.531‡
NO_x	0.652‡	-0.359‡	0.626‡	0.296‡	0.359‡	0.462‡	0.820‡	-0.926‡	1.000	0.475‡	0.545‡	-0.402‡	0.468‡	0.220‡	-0.516‡	0.514‡
$PM_{2.5}$	0.551‡	0.204†	0.130*	0.007	0.315‡	0.008	0.573‡	-0.315‡	0.475‡	1.000	0.460‡	-0.161†	0.202†	0.097	-0.210†	0.424‡
$S0_2$	0.502‡	-0.150*	0.511‡	0.365‡	0.401‡	0.569‡	0.263‡	-0.483‡	0.545‡	0.460‡	1.000	-0.349‡	0.321‡	-0.050	-0.161†	0.082
M-INC	-0.248‡	0.205†	-0.266‡	-0.192†	-0.112*	-0.179†	-0.323‡	0.394‡	-0.402‡	-0.161†	-0.349‡	1.000	-0.677‡	0.219‡	0.519‡	-0.203†
%MIN	0.083	-0.435‡	0.166*	0.029	-0.160*	0.066	0.520‡	-0.507‡	0.468‡	0.202†	0.321‡	-0.677‡	1.000	-0.079	-0.524‡	0.467‡
EDUC	0.031	-0.257‡	0.026	-0.095	-0.099	-0.039	0.323‡	-0.208†	0.220‡	0.097	-0.050	0.219‡	-0.079	1.000	-0.229‡	0.376‡
OWN	-0.102	0.449‡	-0.170*	0.016	0.141*	0.012	-0.602‡	0.525‡	-0.516‡	-0.210†	-0.161†	0.519‡	-0.524‡	-0.229‡	1.000	-0.585‡
POP	0.051	-0.446‡	0.012	-0.115†	-0.270‡	-0.137†	0.740‡	-0.531‡	0.514‡	0.424‡	0.082	-0.203†	0.467‡	0.376‡	-0.585‡	1.000

Abbreviations (in order by table listing): CCHO = acetaldehyde; ACET = acetone; BENZ = benzene; BRC = ethylene dibromide; HCHO = formaldehyde; PHEN = phenols; O_3 = ozone; $PM_{2.5}$ = particulate matter < 2.5 microns in diameter; NO_x = nitrogen oxides; SO_2 = sulfur dioxide; CO = carbon monoxide; M-INC = median household income; %MIN = percent non-white; EDUC = percent with one or more years of college; OWN = percent of owner-occupied housing; POP = population density; CAP = criteria air pollutant; CMAQ = Community Multiscale Air Quality; CMAQ-HAP = CMAQ adapted for selected gas-phase HAPs; HAP = hazardous air pollutant; * = P-value < 0.05; † = P-value < 0.001; ‡ = P-value < 0.0001

Table 8: Principal components analysis. Factors 1 and 2 were chosen for the analyses, with varimax rotation, for the 90-day mean of the 24-hour daily means for the 337 cells, the averaging schema used for this pilot study. Species with one value > 0.40 and the other < 0.40 are particularly well represented by the PCA factor with the higher value in the statistical models.

HAZARDOUS AIR POLLUTANT	ABBR	PCA Factor 1 No Rotation	PCA Factor 2 No Rotation	PCA Factor 1 Varimax	PCA Factor 2 Varimax
Acetaldehyde	CCHO	0.787	0.155	0.710	0.372
Acetone	ACET	0.095	-0.582	0.257	-0.530
Acrolein	ACROL	0.887	-0.132	0.888	0.126
Benzene	BENZ	0.881	0.146	0.803	0.391
1,3-Butadiene	BUTA	0.831	-0.135	0.835	0.107
Chloroform	CHCL	0.343	-0.185	0.382	-0.080
Cresols	CRES	0.766	0.491	0.594	0.689
Ethylene dibromide	BRC	0.696	-0.225	0.732	-0.018
Ethylene dichloride	CLC	0.728	-0.235	0.765	-0.018
Ethylene oxide	ETOX	0.239	0.310	0.141	0.366
Formaldehyde	HCHO	0.810	-0.171	0.825	0.066
Methylene chloride	CLME	0.279	0.840	0.028	0.885
Perchloroethylene	CL4ET	0.093	0.776	-0.131	0.770
Phenols	PHEN	0.856	-0.115	0.853	0.133
Trichloroethylene	CL3ET	0.398	0.290	0.299	0.392
Vinyl chloride	CLET	0.756	-0.277	0.804	-0.050

Abbreviations: ABBR = abbreviation; PCA = principal components analysis

rates were log-transformed before the LMM analyses. The correlation of the outcome variable between male and female within a cell is assumed to be the same for all cells and two outcomes from different cells are assumed to be independent. In statistical terms, this means that the correlation was a bloc-diagonal structure with each block being compound symmetry. The independent variables examined included gender, percent with one or more years of college, median household income, percent non-White, percent owner-inhabited housing, percent 16 years of age or older who commute to work more than 30 minutes one-way daily, nighttime population density, two meteorological variables, five CMAQ4.4-simulated CAPs, 16 CMAQ4.4- or CMAQ-HAP-simulated HAPs, and two PCA factors, rotated using the varimax method. To reduce potential ecologic bias in the pollutant averages across the 337 cells from confounding by demographic variables, each of the meteorological and pollutant variables was preadjusted for the demographic variables (Lipfert, 1994). This multistage approach helps identify potential associations between the outcome variable, i.e., log(age-adjusted hospital admission

rate), and the exposure of particular interest, i.e., the pollutant variables.

Thus, for each of the three univariate LMM analyses (cardiovascular, respiratory, and either) using the 90-day mean averaging schema, we examined 32 potential predictor (independent) variables. For subsequent building of the multivariate models, two models for each of the three outcome variables were built, one with the HAPs and the other with the PCA factors.

Multivariate Linear Mixed-Effects Model

We used a LMM to examine the relationship between age-adjusted hospital admission rates by discharge diagnosis (the dependent variable) and a total of 32 independent variables, with the pollutant variables pre-adjusted by the demographic variables as noted earlier. The unit of analysis was the 337 4 x 4-km cells. In our preliminary analysis of four averaging times and three months of data we developed 24 multivariate models. In our final model using a single averaging schema, the 90-day mean of the 24-hour daily means, six multivariate final models were developed: two for each of the three outcome variables: for each, one that used the two PCA factors and one that used the 16 HAPs.

We used a backward elimination process, including in the development of the final model all variables with a P-value < 0.2 in the univariate analyses. We then excluded the variables that were not significant predictors of the dependent variable, one at a time. The type I error for exclusion was set at 0.05. Various measures of goodness of fit, including the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), were regularly employed as part of the model-building exercise. When developing the multivariate model, we realized that many of the air pollutant variables were highly correlated (see Tables 6 and 7), and that the resultant effect estimates would be unstable and therefore misleading with regard to the contributions of individual pollutants to the outcome of interest, i.e., hospital admission rates by discharge diagnosis. The issue of multipollutant collinearity is discussed in the next section, as well as in the Discussion.

Univariate and Multivariate LMM Results

As noted earlier, during the course of this pilot study our explorations of the data and preliminary analyses highlighted a number of concerns and questions, which need to be addressed in subsequent work. These include in particular concern about the accuracy of the CMAQ model output and averaging paradigms used in this pilot study to reasonably estimate population exposure differences across the 337 cells; and potential distortion in admission rates introduced by significant variations in population across the 337 cells, for which we only partially adjusted in the current model. These and other concerns and challenges are examined in more detail in the Discussion section. Particularly because of our concern regarding CMAQ performance in this multipollutant model and a need to further examine and possibly refine the exposure component with additional techniques, we hesitate to include the output of the univariate and multivariate LMMs in this report. This concern is accentuated by the fact that the collinearity in our multipollutant model makes the effect estimates for the pollutants largely unusable for typical interpretation of the role of specific variables. Our specific concern is readers may inadvertently see the LMM output as "results" rather than an iteration of a methodology being explored. This is further complicated by the fact that the purpose of this study was to define potential "hot spots" of elevated exposure, delineation of which has political consequences. Because the results do not yet meet the rigor needed for such delineation, we feel that it is better not to make this data publically available at this time. For these reasons we are not including any of the three univariate or six multivariate tables of the LMM results of the final analysis, or any of the preliminary analyses, in this report although the data and results may be made available for researchers wishing to examine and build upon our methods.

Conditional Predicted "Hot-Spot" Rates

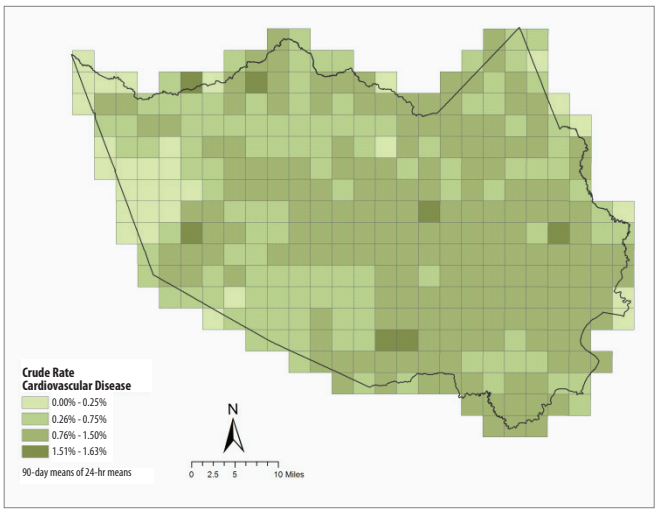
As noted above, our desire to include multiple pollutants in the model that are likely important in assessing differences in exposure in different areas of Harris County resulted in a multivariate model with roughly seven to eight pollutant and meteorological variables remaining in the final model. These variables are in some instances highly correlated, distorting the effect estimates. Although considerable work is being done to develop better statistical approaches for multipollutant models (Kim et al., 2007; Stieb et al., 2008), our multipollutant effort resulted in significant collinearity among the pollutants. Therefore the effect of an individual independent variable in the final model, controlling for the other variables in the model, cannot be accurately deduced from the effect coefficients. Nor can the relative strength of an independent variable be necessarily inferred from the size of its coefficient, P-value, or confidence intervals because of this collinearity, although

the output does contain considerable information that needs to be examined in more detail using data mining and other tools. For these reasons we have chosen not to make the LMM findings from this pilot study available at this time. Collinearity, however, does not affect the ability of the regression equations to predict the response.

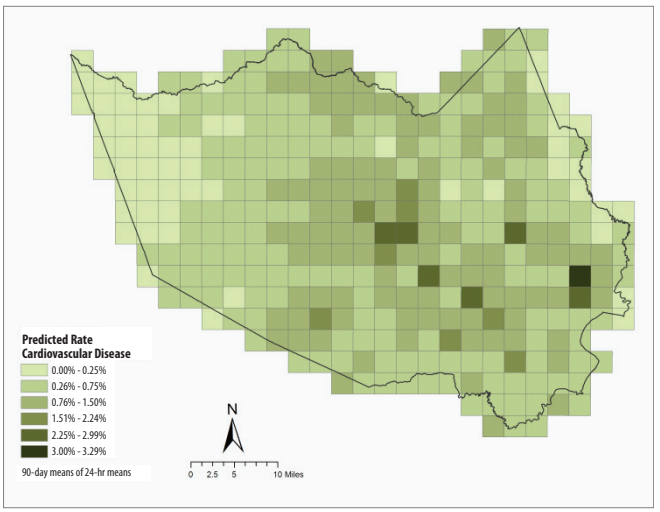
The underlying objective of this pilot study was to see if a spatial analysis utilizing modeled pollutant concentrations, including air toxics, demographic variables, and hospital admission rates by 337 4 x 4-km cells could predict potential "hot spots" of disproportionately elevated health effects, which could subsequently be targeted for more in-depth evaluation and, as appropriate, intervention. Although the quality of the MM5, CMAQ4.4, and CMAQ-HAP simulated output and averaging times for the variables chosen at this time appear to be inadequate for reliable comparisons of exposure among the 337 cells, the general approach is very promising.

The methodology used in this pilot study utilizes the coefficients from the multivariate models to predict rates of hospitalization, by gender, for cardiovascular, respiratory, and cardiovascular or respiratory (either) disease. Figure 11 graphically depicts the crude age- and gender-adjusted rates (Figure 11A) and conditional predicted rates (Figure 11B), using the multivariate output for cardiovascular disease. The crude hospitalization rates are the actual rates of hospitalization by discharge diagnosis during the study period, based on the geoaddressed THCIC data, corrected for the age distribution in each cell. The predicted rates for each outcome were calculated by summing the products, for each statistically significant variable that remained in the final multivariate LMM, of the log effect estimate times the value for that variable, for each cell. Figure 11C is the underlying equation for the output shown in Figure 11B.

A. Cardiovascular disease (crude)



B. Predicted cardiovascular disease (adjusted)



C. Equation for predicted rates of cardiovascular disease, controlling for the variables in the multivariate LMM (gender = 0 if female) for each of the 337 cells. The log rates were exponentiated for the map shown in B.

$$\log[\text{age-adjusted cardiovascular admission rate}]_{\text{Predicted}} = -52.8020 + (-0.04667 \times \text{Gender}) + (-0.0710 \times \%\text{College}) + (0.0300 \times \text{Median Income}) + (-0.00539 \times \%\text{Non-White}) + (0.0065 \times \text{Housing Owner Rate}) + (0.0318 \times \%\text{Commute}) + (0.0001 \times \text{Nighttime Population Density}) + (0.9848 \times \text{Temperature}) + (44.1927 \times \text{CO}) + (0.4787 \times \text{O3}) + (0.2065 \times \text{PM2.5}) + (-33.9496 \times \text{CCHO}) + (86.6513 \times \text{ACROL}) + (3.5481 \times \text{HCHO}) + (20.3236 \times \text{CLME})$$

Figure 11: Sample map of crude and predicted "hot spots" for rates of cardiovascular admissions. The predicted rates were calculated from the regression coefficients from the final multivariate linear mixed-levels model for cardiovascular disease. Rates for males and females were proportionally combined, and the predicted log(age-adjusted cardiovascular rate) exponentiated for mapping. A. Age-adjusted crude rate. B. Adjusted rate from multivariate LLM. C. Equation for generation of the "hot spot" map shown in B.

Although the models were stratified by gender, gender was not a significant variable in any of the LMMs. Thus for the predicted rates, the gender-specific log rates by outcome were proportionally combined based on the gender structure of each cell, and then the combined log rates were exponentiated, resulting in the percentage of predicted hospitalization by cardiovascular, respiratory, or cardiovascular or respiratory (either) discharge diagnosis for each of the 337 cells. Figure 11C is the underlying equation for the output shown in Figure 11B. By using the multivariate output, the rates are adjusted by the factors that are significant predictors of hospitalization for each outcome. Thus much of the "noise" that is reflected in the crude age-adjusted rates is removed in the predicted rates calculated from the final regression equations, resulting in a more coherent "picture" of potential pollution-related "hot spots" associated with elevated rates of hospitalization. The distortion of effect estimates for individual correlated variables is not a problem when the estimates are used together to predict hospitalization rates across the cells.

Spatial Autocorrelation

After developing the final multivariate LMMs, we explored the conditional predicted residuals for any remaining spatial autocorrelation not accounted for in the regression models. For this exploration, we used ESRI's Geostatistical Analyst to plot empirical semivariograms of the conditional predicted residuals for each of the three final models with the best fit. For each semivariogram cloud of residuals, each of which represents a pair of locations, the point where the plot flattens out indicates that the relationships between the pairs of locations beyond this distance are no longer correlated. The semivariograms (not shown) of the conditional predicted residuals from each of the final LMMs suggested little remaining autocorrelation. This suggests that most of the important variables differentiating the hospitalization rates between near or adjacent cells were included in the model, leaving little unaccounted spatial correlation between cells that could bias the findings. More sophisticated data mining techniques for potential autocorrelation of variables with varying spatial resolution and confidence in a grid-based health effects model are, however, needed. Such techniques may delineate important spatial patterns for additional exploration (Waller and Gotway, 2004).

DISCUSSION

As discussed in the previous section, concerns about the emission inventories, the accuracy of the CMAQ simulations for estimating human exposure, the appropriateness of the averaging time used, and population-based variations in statistical uncertainty-among other issues yet poorly articulated--led us to refrain from presenting the output from the multivariate linear mixed-effects regression models in this report for fear that so doing might be misconstrued. Nevertheless, we feel that using CMAQ-possibly in combination with other air quality models and/or observed values-with actual health endpoints to help to predict potential geospatial "hot spots" of concern shows considerable promise, and that the underlying methodology developed in this pilot study provides a template for testing and refining various components of this initial effort. In addition, although there are justifiable concerns about, in particular, the estimate of exposure, our study was successful methodologically in demonstrating a general procedure that is not dependent on the known shortcomings of fixed-site monitors to create predictive spatial maps of multipollutant exposure and associated health effects that can identify particular areas of concern for additional investigation. Preliminary output from the multivariate models was also encouraging in that the variables that remained in the cardiovascular and respiratory spatial models were surprisingly consistent using different averaging times and with and without PCA. In addition, the pollution and demographic variables that remained in the models were generally consistent with other studies of cardiorespiratory endpoints and ambient air pollution, and the spatial delineation of "hot spots" in Harris County based on conditional predicted hospitalization rates calculated from the full multivariate pollutant-based models suggested several areas that have already been defined as areas of concern by the TCEQ's APWL (Texas Commission on Environmental Quality, 2008) and NATA. Thus we believe that the methodology in general shows promise.

In this pilot study we sought to test, evaluate, and improve the methodology for conducting multipollutant "hot spot" analyses, using simulated pollutant concentrations and actual health effects. More precisely, we explored the usefulness of a Eulerian photochemical transport model to estimate exposure at the 4-km level, the estimated exposure then utilized in a health-based multivariate model that attempted to identify areas of disproportionate exposure and adverse health effects in Harris County, Texas. The pollutant simulation model we

used was the U.S. EPA's CMAQ model, with improved emissions input for Texas and several modifications of the model itself to explicitly represent 16 HAPs of particular concern in the Houston area. Also included were simulated vales for five criteria pollutants and two meteorological variables as potential predictors of effect, measured in this study by hospitalization for cardiovascular or respiratory disease during a 98-day period in 2000. The study attempted to differentiate levels of chronic multipollutant exposure and adverse health effects across the 337 4 x 4-km cells that overlay the county, controlling for a number of other variables that can affect the likelihood of hospitalization, such as income, ethnicity, age, education, and several other individual- and group-level demographic factors. It is fundamentally a spatial "hot spot" study with the 90-day exposure window representing chronic exposure, although within-cell exposure variability is an important variable that needs to be addressed in subsequent work. The underlying question of this pilot effort, then, is whether living in a 4 x 4-km cell with consistently higher concentrations of multiple pollutants increases the overall likelihood of residents in that cell experiencing adverse health events, such as heart attack, stroke, arrhythmias, asthma attacks, and pneumonia, that would likely result in hospitalization.

The pilot study is noteworthy for several things it attempted: (1) use of a one-atmosphere photochemical model to provide geographical coverage for all of Harris County, as well as to include multiple pollutant species (e.g., CAPs and HAPs, including secondarily formed photochemical components of these pollutants) in the same model; (2) development of a computationally efficient software program for generating a subset of HAPs for future larger studies; (3) utilization of a grid-based spatial model that lends itself to subsequent refinement; (4) a focus on defining "hot spots" rather than capturing the role of specific pollutants in the disease endpoints studied; and (5) use of actual health data to define risk in a "hot spot" model.

The pilot study is also noteworthy for the number of challenges it encountered, many of which are beyond the scope and resources of this initial effort and will need to be addressed in future work. As noted throughout this report, these include concerns about the emission inventories, the ability of the CMAQ simulations to reproduce measured concentrations and provide reasonably accurate approximation of exposure to multiple pollutants at various spatial resolutions, the appropriateness of the averaging time used, population-based variations in statistical uncertainty, and other issues yet poorly articulated. This discussion focuses on some of our more general findings and challenges encountered.

MULTIPOLLUTANT RESEARCH

As demonstrated by a recent special issue of the *Journal of Exposure Science & Environmental Epidemiology* dedicated to the "Interpretation of Epidemiologic Studies of Multipollutant Ambient Air Exposure and Health Effects," the appreciation of the need to address multipollutant exposure and its attendant challenges is increasingly a major topic (Ito et al., 2007; Kim et al., 2007). Key issues that limit multipollutant models include covariation of pollutants, potential confounding by mismeasured or unmeasured pollutants, complex health-associated interactions, the rationale for selection of pollutants or pollutant groups, and extensive exposure uncertainty.

In this pilot study, we included a large number of meteorological and pollutant variables. Although we are not including the regression output in this report, in part because of the distortion in effect estimates created by such a multipollutant model, it is useful to note that approximately five to seven air quality variables remained in most of our multivariate LMM models, with temperature, CO, O_3, $PM_{2.5}$, and formaldehyde represented in most. There were small differences based on whether cardiovascular or respiratory disease was the outcome measure, as well as with different inventories (regular or imputed TEI) and averaging times. For example, in the earlier 92-day analysis using the regular TEI that explored multiple averaging times, among the CAPs, CO, O_3, NO_2, and $PM_{2.5}$ tended to remain in the models that used the mean of the 24-hour daily means, but only CO and O_3 remained in models using the maximum daily moving six-hour mean. Although this may be the result of simulation error, it may also reflect the fact that CO and especially O_3 tend to be higher during the day, whereas NO_2/NO_x and PM tend to be somewhat higher at night. The maximum six-hour moving average tends to attenuate these daily temporal differences that, combined with different day/night activity patterns, may be reflected in the final health models. In the more recent 90-day model that used the imputed TEI, NO_x did not remain in the models and may relate to reduced NO_x levels associated with the addition of HRVOCs and thus higher O_3 generation (Kim et al., 2006). In addition to the complexity of multipollutant exposure itself, the abilities of "one-atmosphere" air quality models such as CMAQ and statistical analysis tools to adequately respond to shifting input and interactions continues to be a major impediment to such research.

We chose to effectively ignore the collinearity problem by focusing on predicting "hot spots" from the multivariate equations. This, we feel, has considerable validity for delineation of areas of elevated stress, despite the fact that interpretation of much of the model output itself is complicated by the correlation between pollutants in the final LMM. Nevertheless, multiple pollutant exposure is critical in realistically defining "hot spots," as well as in better understanding resulting health effects. There have been a number of multipollutant studies. for example, that have found CO, $PM_{2.5}$, O_3, and often NO_x/NO_2 to predict hospitalization or other adverse health outcomes (Ballester et al., 2001; Barnett et al., 2006; Burnett et al., 1997; Burnett et al., 1997; Burnett et al., 1999; Fusco et al., 2001; Hinwood et al., 2006; Koken et al., 2003; Lanki et al., 2006; Lin et al., 2003; Linn et al., 2000; Morris et al., 1995; Schwartz, 1999; Sheppard et al., 1999; Yang et al., 2007; Yang et al., 1998). Of particular import for this discussion and the challenges of this pilot effort is that these multiple pollutant models have stimulated considerable discussion about interpretation, including the possibility that some previously published work included effect estimates that may be distorted due to correlation among the pollutants. Other issues that are being addressed include the likelihood that a number of pollutants may actually be surrogates for other pollutants in these models, that key pollutants may not be included in some models at all, emerging methods for better handling collinearity with more advanced statistical techniques, and improvements in speciation monitoring and receptor modeling to expand our knowledge of pollutant characteristics (Bateson et al., 2007; Brown et al., 2007; Kim et al., 2007; Sarnat et al., 2007; Tolbert et al., 2007).

Several multipollutant studies of particular relevance are summarized here, focusing particularly on cardiovascular and respiratory effects, as these were the outcomes studied in our pilot.

Linn and associates, for example, studied the associations between ambient CO, NO_2, $PM_{2.5}$, and O_3 and cardiopulmonary hospital admissions during 1991-1995 in metropolitan Los Angeles. Carbon monoxide showed the most consistently significant relationships with hospitalizations, with an increase of approximately 1.1 ppm associated with a 4% increase in hospitalizations (Linn et al., 2000). Burnett and associates found a stronger relationship between CO and hospitalization for congestive heart failure (CHF) than between NO_2, O_3, SO_2, or haze and CHF in the elderly in a study of 134 hospitals in Canada's largest cities (Burnett et al., 1997). In a large study of 8,582 cerebrovascular admissions and air pollution in Taipei, Taiwan, Chan and associates found that CO and O_3 were more consistently associated with cerebrovascular admissions than were the other pollutants studied, including NO_2, SO_2, $PM_{2.5}$, and PM_{10} (Chan et al., 2006). In their study, the effects were most pronounced with a 0-day lag for O_3, a 2-day lag for CO, and a 3-day lag for PM.

Of increasing interest in many of the PM studies is particle mass, particle number, speciation, and the degree to which certain pollutants may be effective surrogates for other pollutants (Ito et al., 2007; Kim et al., 2007). A recent study by Janhall and Hallquist, for example, found that NO correlates strongly with ultrafine particles (UFPs) along roadsides, whereas NO_2 correlates strongly with background levels of UFP (Janhall and Hallquist, 2005). There is increasing interest in the role of UFPs, especially in cardiovascular disease.

The role of O_3 in manifestations of cardiovascular and respiratory disease in multipollutant studies is somewhat conflicting. Ozone has been associated with hospital admissions in a number of studies, although there seems to be considerable intercity differences (Medina-Ramon et al., 2006), with some suggestion that elevations in O_3 may have a greater effect on vulnerable populations in areas with generally lower background O_3 concentrations (Medina-Ramon and Schwartz, 2008). In a multipollutant study in Atlanta, GA, that used emergency department visits as its health endpoint, researchers at Emory University observed, for cardiovascular visits, associations with CO, NO_2, and $PM_{2.5}$ elemental carbon and organic carbon, with CO as the strongest predictor (Tolbert et al., 2007). For respiratory visits, associations were observed with O_3, PM_{10}, CO, and NO_2 in single-pollutant models, whereas in multipollutant models, PM_{10} and O_3 persisted as predictors, with O_3 the stronger predictor. The authors stress the difficulty of interpreting multipollutant models, noting that they can be as biased and as misleading as single-pollutant models.

CMAQ AS AN ESTIMATE OF EXPOSURE

CMAQ is intended as a multipollutant model, but do the simulated concentrations estimate exposure sufficiently well at this point for researchers and/or policy makers to have reasonable confidence in associations with health effects observed?

A rigorous analysis of the sensitivity of MM5, CMAQ4.4 (with the SAPRC99-ARO mechanism), and CMAQ-HAP to adequately simulate observed values measured at monitors (one of the more common comparisons for modeled output) is beyond the scope of this report. However, such analyses are being done elsewhere (Appel et al., 2007; Byun et al.,

2007; Ching et al., 2006; Gego et al., 2006; Hakami et al., 2007; Irwin et al., 2008; Jimenez-Guerrero et al., 2008; Marshall et al., 2008; Napelenok et al., 2008; Seigneur, 2005; Sokhi et al., 2006). The CMAQ model, the elements of which are continually being updated (see www.cmaq-model.org for the recent new updates, including CMAQ4.6), is one of three air quality models that can be used to model O_3 and PM for SIP development (U.S. Environmental Protection Agency, 2008). EPA's guidance document for the use of models, *Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, $PM_{2.5}$, and Regional Haze*, offers a number of approaches for assessing model accuracy, such as comparisons between modeled and observed ratios of indicator species, the decoupled direct method (DDM) for assessing a model's sensitivity to input perturbations, integrated process rate (IPR) analyses to assess the relative contributions of model components to output, and sensitivity tests of alternative model input in predicting known concentrations (U.S. Environmental Protection Agency, 2008).

A number of investigators have assessed the sensitivity of CMAQ using different methods. Byun and associates recently evaluated the sensitivity of CAMx and CMAQ, using similar model configurations, to simulate a high-ozone episode in Houston in August 2000 (Byun et al., 2007). They found that the two models performed similarly using the base emission inventory but that, with the addition of the imputed HRVOC emissions-which we use in our final model, CMAQ predicted lower ozone peaks in HRVOC-rich areas than did CAMx, suggesting that the CMAQ system may be radical poor relative to ozone formation. Hakami and associates used DDM and adjoint methods of CMAQ to collect sensitivity information about specific receptors, which can be used to improve model performance but is also a powerful tool for identifying emission sources that are associated with "hot spots" (Hakami et al., 2007). Napelenok and associates, including researchers at Rice University, are using a three-dimensional DDM approach to assess the sensitivity of CMAQ during a summer ozone episode at different scales across the U.S., finding good correlation for both primary and secondary pollutants (Napelenok et al., 2008).

Although the researchers noted above primarily focused on ozone, a number have addressed $PM_{2.5}$ and NO_x as well in comprehensive performance evaluations that generally also evaluated the performance measure, such as those performed on the CMAQ simulations from the Southern Oxidant Study (Zhang et al., 2006; Zhang et al., 2006; Zhang et al., 2006). Wyat Appel and co-workers have recently focused on CMAQ version 4.5's performance in predicting $PM_{2.5}$ (Wyat Appel et al., 2008), a pollutant of considerable concern in our study because of the numerous studies linking it to cardiovascular disease.

Luecken, Hutzell, and Gibson have reported on the predictive capability of CMAQ, compared with observed concentrations, for five HAPs (formaldehyde, acetaldehyde, benzene, 1,3-butadiene, and acrolein) and found generally good agreement, with a tendency for underpredicting (Luecken et al., 2006). Ching and co-workers used a version of CMAQ4.4 with a modified CB4 mechanism to study the usefulness of simulated HAP concentrations in the Philadelphia area. They found that "modeled mean values compared reasonably well against the observed concentrations" (Ching et al., 2004). These and other studies are useful not only in broadly assessing the possible utility of a study such as ours, but also with helping us to interpret our output in the context of known CMAQ strengths and weaknesses.

As part of our study we compared simulated concentrations with observed concentrations at Harris County monitors, examining hourly pairs, the averaging schema used in this pilot study (90-day mean of the 24-hour daily means), outliers, and temporal agreement (Figure 5). In general, the time series and scatter plots of hourly pairs performed better than the scatter plots of the pairs reflecting the averaging period used, i.e., the 90-day mean of the 24-hour daily means, but there were differences among pollutants. For the averaging period utilized, the R^2 values ranged from > 0.8 for temperature to 0.03 and 0.04 for $PM_{2.5}$. For O_3, the R^2 at the four monitors plotted ranged between 0.46 and 0.53. For all 337 cells, the simulated-to-observed ratio of was 0.997 for temperature, 1.385 for O_3, 0.433 for CO, 0.680 for NO_x, 0.750 for $PM_{2.5}$ and 0.252 for formaldehyde. The overprediction by CMAQ of O_3 concentrations is largely due to a nighttime bias, which is why the apparent performance greatly improves if maximum 6- or 8-hour moving averages are used.

The above comparisons with observed data may be somewhat misleading in that in several instances the monitor data were either extremely sparse or acknowledged to be of poor quality. The CO and $PM_{2.5}$ monitors, in particular, performed poorly during our study period, with the CO monitors having a problem with variable baseline shift. The $PM_{2.5}$ monitoring had just begun and was still being tested. Current $PM_{2.5}$ monitor coverage is better and the correlations have improved considerably. Current UH-IMAQS hourly 24-hour R^2 values for $PM_{2.5}$ at Channelview and Deer Park (the only two monitors measuring $PM_{2.5}$ during our study period) in April 2008 were approximately

0.3 and 0.5 (an improvement by a factor of 10), respectively, with the R^2 values for the "best-fit" comparison, i.e., the closest CMAQ:CAMS comparison within 9 adjacent 4 x 4-km cells, considerably higher.

For our pilot study the accuracy of the spatial ranking of concentrations across the 337 cells is of most interest, and in general there are too few monitors measuring the pollutants of interest, and the poor geographical distribution of the monitors (largely in East Harris County) complicates any spatial appraisal. Even the large but consistent overprediction of mean O_3 concentrations would not necessarily be a major impediment if the monitored and observed values were in the same rank order, from high to low. However, as noted in Figure 7, this is not generally the case using our averaging schema. Using a maximum 6- or 8-hour moving average would greatly improve the simulated-to-observed ratio but would it improve the estimate of exposure? Does having multiple pollutants in the conditional predicted rates generated by the multivariate LMM improve the exposure metric and delineation of "hot spots" by allowing collinearity? Or degrade it? How does one best assess spatial performance with little or no observed data? As discussed earlier, simulations even at 4-km resolution will not represent peaks that may represent a subgrid feature. We cannot address this aspect of variability in this pilot study but recognize that future studies are needed to examine the contributions of these peaks and associated subgrid variability to explaining adverse health events.

AVERAGING SCHEMA

The choice of an averaging period (or different averaging periods for each pollutant) is a key component in the study design. Inappropriate averaging can diminish or even eliminate important exposure characteristics, and temporal misalignment can lead to measurement error of a different kind (Bateson et al., 2007). The foremost consideration is an appropriate averaging schema for the health endpoint being studied. At the same time, however, especially when using simulated meteorological and pollutant data, one wants an averaging time that optimizes the accuracy of the exposure estimation. In the case in which the health outcome seems best served by a measure of chronic exposure, such as the mean of the 24-hour daily means, and the accuracy of the simulated exposure optimized by using shorter averaging times, several hybrid approaches may be useful. For example, the use of the longer-term average such as we used (90 day mean of the 24-hour daily means) to characterize baseline exposure might be weighted or augmented by a shorter averaging period, such as the maximum daily concentration.

Another approach for ambient exposure would be to ignore penetrance ratios but to limit the portion of the day averaged, such as from 9 am through 9 pm. This would better reflect times when people are likely to be outdoors, as well as reduce one common deficiency of CMAQ and other photochemical models, that is, night-time bias. This however underestimates exposure to PM, which has a high penetrance ratio. A concern, in general, with variable windows of exposure is that we may not adequately understand the pollutant characteristics or the biological mechanisms. For example, recent studies suggest that indoor exposure to O_3 is higher than previously recognized, and that formaldehyde and other oxidation products that are quickly formed indoors or within vehicles by O_3 may contribute to adverse health effects that should be attributed to O_3 (Weschler, 2006).

The choice of an averaging schema is also related to within-cell variability, which is addressed briefly in the section on temporal variability, and more comprehensively in the section on subgrid variability in Future Efforts. Additional work is needed to explore of short-term excursions in cells with high or low chronic exposure. This may be especially important in the HGA as, in addition to baseline geographic differences in exposure, some areas are regularly exposed to short-term but significant elevations in pollutant concentrations as noted earlier (Webster et al., 2007).

A better understanding of the appropriate averaging time-which is complexly linked with available exposure measurements and/or simulations, activity patterns, daily concentration variability, and biological response to exposure-is needed. As CMAQ's ability to simulate more pollutants improves, some of the pressure to choose a high performance averaging schema, despite the outcome being studied, will lessen. Similarly, a hybrid overlay with Gaussian or Lagrangian simulations may improve the exposure metric and therefore allow more appropriate averaging schema and exposure modulation, without undermining the model's findings with an averaging schema poorly matched to the exposure estimation ability of the model.

COLLINEARITY AND EFFECT ESTIMATES

Collinearity is the statistical difficulty of separating the independent effects of correlated variables in multipollutant models as more than one variable may explain the same variability in the outcome measure

(Bateson et al., 2007; Ito et al., 2007). Standard regression models routinely have difficulty assigning the "correct" variability to each variable, often resulting in highly unstable effect estimates. In this situation, the effect of an individual independent variable in the final model, controlling for the other variables in the model, cannot be calculated from its effect coefficient, nor can the relative strength of the variable be necessarily inferred from the size of its coefficient, P-value, or confidence intervals. Collinearity, however, does not affect the ability of the regression equation to predict the response, as reflected in our study's "hot spot" maps. It is only a problem if the intent is to estimate the contributions of individual predictors, which is usually the case. To reduce collinearity in multiple regression models, several techniques, including hierarchical regression methods, PCA for the entire set of pollutant variables, and elements of Bayesian model averaging can be useful in reducing collinearity in some models.

OTHER CHALLENGES

The grid-based structure effectively introduced a new variable in that the population density, and therefore the confidence in the hospitalization rates, varied significantly across the 337 cells in Harris County. In the Methods section we describe an approach that reduces the effect of low-population cells on the analysis. Future efforts are needed to qualify all of the cells based on the confidence in each cell's data, without undermining the model's power to predict "hot spots" in rural or other low population areas.

As noted in the statistical section of the Methods, we evaluated the usefulness of PCA as a means of extracting subsets of linear combinations that account for most of the variability and then using these as surrogate predictors of the 16 HAPs. The first two factors, with varimax rotation, were chosen for inclusion in the multivariate LMMs as together they represented approximately 60% of the variability. For each of the three diagnostic outcomes, we ran the regression with the PCA factors and without, i.e., with the original 16 air toxics. The AIC was improved in each instance in the model that used all 16 air toxics. Although multi-collinearity among the HAPs was eliminated by using the PCA factors, the regression coefficients in the full model were still unstable due to correlation among the CAPs and, because of difficulty in interpretation of PCA factors, the PCA factors were not used in the final predictive maps. Additional work in this area, however, is warranted.

Although in this pilot study we examined hospitalizations categorized by discharge diagnosis of cardiovascular, respiratory, and cardiovascular or respiratory (either), we might consider focusing on cardiovascular in future efforts. The primary reason would be that the underlying biological mechanisms of pollution-induced cardiovascular and respiratory disease tends to have two relatively distinct components, chronic and acute, with vulnerability to acute effects highly dependent on underlying disease. For a chronic exposure study examining different areas of Harris County or elsewhere, the primary question is whether living in an area with higher levels of pollution make one more vulnerable to acute episodes. Pollution-induced respiratory effects tend to be more acute and often reversible. Exacerbation of asthma is a typical example. In our hospital admission database, 36.1% of all admissions (N = 35,436) were for cardiovascular disease; 17.5% were for respiratory disease. Cardiovascular disease also inflicts greater morbidity and mortality on the population, and therefore may be of greater public health concern. Carbon monoxide, $PM_{2.5}$, and O_3 have all been implicated in chronic and acute effects in cardiovascular disease. Among the CO- and PM-induced effects are reduced oxygen-carrying capacity, inflammation, arterial endothelial damage, effects on atherosclerotic plaque stability, altered endothelial function, effects on autonomic function, and increased reactive oxygen species (European Commission Directorate-General XI, 1999). In addition, several investigators have suggested that CO may be a marker or surrogate for traffic-related pollution, as well as for UFPs and/or particulate number concentration (PNC) (Lanki et al., 2006), both which appear to play a strong role in cardiovascular disease and secondary events. In addition, O_3 has been linked with arrhythmias, decreased oxygen-carrying capacity secondary to inflammation and increased reactive oxygen species, although O_3 has more typically been associated with shorter-term respiratory effects.

Using a pollutant-associated outcome that is affected by both chronic multipollutant exposure and short-term excursions and using the more focused biological mechanisms to help determine variables and averaging time may result in a in a model that is better able to discriminate effect and resolve spatially with more accuracy.

LIMITATIONS

Concerns about the ability of the CMAQ model to adequately simulate the pollutant concentrations temporally and spatially for the purposes of this pilot health effects study continue to be the most significant concern, ultimately making any conclusions from the health-based model premature or qualitative. Questions about the best averaging schema, and poor geographical coverage and some measurement problems at the monitors increase the difficulty of adequately assessing the model. In addition, there are known problems with the emissions inventories, and poor spatial resolution for area sources is a particular concern for "hot spot" analyses. Other concerns and limitations not discussed earlier include misclassification errors in the hospital admissions databases, probably largely nondifferential but also including some differential interhospital error (for example, public inner-city hospitals may be less likely to send in complete records); and questionable generalizations from domain data (ecologic bias). We are aware that some bias was introduced in geocoding, i.e., in general the disenfranchised who are also more likely to be affected by elevated exposure to air pollution are also most likely to not get health care, to not have their healthcare records sent to the state, and to have an address (or not have an address) that doesn't exist, doesn't geocode because of mistakes, or doesn't represent where this person lives. However, loss of high-risk individuals would more likely reduce any significant findings rather than bias the study toward spurious unsupported findings. Other potential problems include differential error of migration (for example, at-risk populations may move from a high exposure area in order to live in a lower exposure area and to be close to medical facilities, or sick individuals may move to a poorer and higher exposure area because of the expense of their illness), lack of information on smoking and other lifestyle choices, and lack of information on indoor exposure or activity (e.g., time spent driving which, especially for air toxics, has been shown to be the source of highest exposure to air toxics for many individuals).

IMPLICATIONS

Our pilot project utilized the EPA's Eulerian CMAQ4.4 and CMAQ-HAP models (Byun and Schere, 2006; Byun and Ching, 1999; Byun et al., 2003), the first of which was adapted for this study by modifying the chemical mechanism, SAPRC99, to explicitly simulate a numbers of HAPs. CMAQ-HAP was especially developed for this project to work in concert with the adapted CMAQ4.4 output to significantly reduce computational resources needed. This may aid in using CMAQ for future health studies, as health effects studies typically need sufficient time or geography to supply an adequate number of cases to have sufficient statistical power to detect a difference.

In addition, the framework and georeferenced pollutant, patient, and Census 2000 databases that we have developed lend themselves to other analyses, as well as continuing analysis of these or additional compatible datasets. As the exposure metric is improved, the grid-based GIS model is amenable to future incorporation of additional layers of spatial attribute data. This approach also has the advantage of building increasing complexity into the model in an iterative fashion, nesting additional information into previously completed work or, as appropriate, aggregating outward (e.g., patient addresses) at different resolutions and/or geographical shapes. This includes the possibility of adding subgrid information or building hybrid models, both of which are discussed in the next section. Yanosky and associates have recently taken a similar approach in building a layered high resolution PM map from disparate geospatial and temporal sources for subsequent use with health data from the Nurses' Health Study cohort (Yanosky et al., 2008).

Despite significant and continuing problems and uncertainties, we feel that the use of a multipollutant model to generate equations for calculating conditional predictive rates by 4 x 4-km or other unit of geographical analysis is a potentially promising approach for delineation of "hot spots," which are poorly captured by monitoring networks and most study designs. These areas could then be targeted for additional exploration using fixed and/or mobile monitors, aerial photography, "differential light absorption and ranging (DIAL) leak-finding cameras, speciation, receptor modeling, community questionnaires, or other methods. These targeted studies could lead to improved monitoring and emissions data, as well as novel approaches to potentially reduce model biases in the "hot-spot" areas, leading to improved multipollutant model predictions, and the subsequent increase in the confidence of the statistical outcomes.

FUTURE EFFORTS

Numerous refinements of our methods and data input are needed, including additional exploration of the existing data with more rigorous and focused techniques, as well as improved methods to better estimate exposure and deal with uncertainty in a multipollutant health-effects model. Future efforts might include the following.

DATA MINING

Other future efforts that could indirectly support the efforts to refine the estimate of exposure might include more extensive analysis of the existing spatial datasets using advanced data mining techniques, and more rigorous examination of different averaging schema that may be important. For example, additional information on daily temporal variations in simulated pollutants along with penetrance and exposure factors might suggest that certain averaging periods and/or times of the day are particularly important. This could potentially be determined at the individual pollutant level. Care would need to be taken to not inadvertently introduce new variables or misrepresent hazards captured by simpler averaging. For example, recent evidence suggests that health effects associated with O_3 may be misrepresented by only considering outdoor maximum moving eight-hour outdoor averages due to secondarily formed oxidation products indoors that may attenuate O_3-associated health effects (Weschler, 2006). Data mining can be used to help explore potential approaches and identify underlying patterns that may be important.

IMPROVED LOCAL-SCALE EXPOSURE ESTIMATION

Emission Inventories

Although improvement of the emission inventories is generally beyond our capabilities, the inventories are being continually refined and the TCEQ has invested significant resources into improving the emission information available for the HGA and into modeling the region accurately. Industrial sources are improving their emissions estimates, and state and federal governments are increasingly requiring more chemicals, metals, and particulates to be reported. In addition, increased stack and fence-line monitoring and use of new technologies such as DIAL and aircraft capable of high-resolution samplings are regularly identifying emissions or concentration levels that are being subsequently used to refine the inventories. More realistic vehicle emission factors are also improving model input. On a more local scale, area sources such as gasoline stations and restaurants could be geoaddressed for better resolution. These improvements in emissions input and understanding will improve CMAQ and the exposure metric, especially for spatial "hot spot" analyses.

Subgrid Variability

At 4-km resolution, the modeled pollutant signal is known to have filtered high pollutant concentrations levels arising from local sources, sources that are important for "hot spot" delineation. Exposures to these localized high pollutant levels would not be discriminated using the grid mesh for this study. As can be seen in Figure 2, which maps the standard deviations of the simulated hourly concentrations, there is considerable variability within and between cells that is not captured in our current model. Finer resolution modeling, which entails considerable computer resources, improved emissions data, and/or alternative approaches, is required to gain additional precision and confidence in further assessing the rigor of the hypothesis. Subgrid variability (SGV) pollutant distribution functions could be used to better characterize intracell variability in a stochastic framework, and methods for such SGV functions are currently under development by Ching and colleagues (Ching et al., 2005; Ching et al., 2006; Isakov et al., In press). Such additional data on spatial texture may indeed be needed if health triggers are exacerbated by populations exposed and compromised by elevated pollutant levels currently filtered by the coarseness of operational grid model outputs.

Hybrid Modeling

A hybrid model could also be used to increase local-scale resolution. In this approach, the CMAQ output would likely serve as the base model, which would then be overlaid with additional georeferenced information such as from measured concentrations at fixed-site monitors, dispersion models such as AERMOD that might, for example, use a link-node roadway overlay to boost concentration levels near roadways, and/or a Lagrangian model such as HYSPLIT to better account for variability in emission sources (Ching et al., 2006; Stein et al., 2007). This allows for a weighted ensemble model that benefits from several approaches to improve the exposure metric.

Exposure Factors

The use of SGV distribution functions are especially attractive in this hypothesis testing mode since it is population exposure in a grid cell that is being assessed, and it is unlikely that any person's exposure is constant but rather is variable, i.e., many degrees of freedom, depending on individual and/or group patterns of activities and time in microenvironments that might occur within any specific

model grid. Additionally, explicit activity and indoor-outdoor penetration factors have not been factored into the current study. Future efforts should take these exposure variables into consideration. Personal monitoring data and penetration factors are available from several local studies, including the Relationships of Indoor, Outdoor, and Personal Air (RIOPA) study (Weisel et al., 2005) and the Houston Exposure to Air Toxics Study (HEATS) (Morandi and Stock, 2008). Between the two studies, approximately 300 Houston-area residents participated in personal monitoring, with microenvironment and monitor measurements also obtained. Recent Harris County-specific survey data from the U.S. Center for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) (U.S. Centers for Disease Control and Prevention, 2008) may also be helpful. Utilization of such information for refining the exposure metric would likely involve utilizing existing exposure software such as HAPEM or SHEDS (Georgopoulos et al., 2005). This could be achieved fairly efficiently given that the pollutant, demographic, and hospital admissions data already exist.

NESTED EXPLORATIONS

The CMAQ nested grid-based model allows for increasing resolution in areas of particular interest and encourages models that continue to refine previous efforts. Local-scale HAP modeling using adaptations of the chemical mechanisms associated with CMAQ has been done for several urban areas, including portions of Houston (Ching et al., 2004) and Philadelphia (Ching et al., 2004; Isakov et al., 2007). The identification of areas of interest can also provide rationale not only for higher resolution local-scale modeling but also for the temporary or long-term placement of monitors in areas of concern.

Another potentially useful effort, building on the work completed to date, might use a time-series design to examine day-to-day variations in hospital admissions in several blocks of contiguous cells (possibly two 12-km blocks of 4 x 4-km cells) with significantly different levels of exposure, as determined by the 90-day mean of the 24-hour means. Such a study would help tease apart the effects of chronic and short-term elevations on health outcomes, and could focus on areas with sufficient population to reduce some of the grid-based error associated with the associated uncertainty in small-population cells.

CONCLUSIONS

Although ours is a pilot study, the geospatial methods and databases we have developed to bring together simulations from MM5 and CMAQ, hospital admission data, and demographic variables will be useful for future refinements of our "hot spot" approach, as well as to better characterize Harris County in general. With improved exposure input, we anticipate that the model will be useful in delineating potential "hot spots" of disproportionate exposure and vulnerability to adverse health effects in Harris County, areas that can then be targeted for additional research and/or intervention.

ACKNOWLEDGEMENTS

Primary funding support for the research described in this report came from the Mickey Leland National Urban Air Toxics Research Center (NUATRC), an organization jointly funded by the U.S. Environmental Protection Agency (EPA) and private industry sponsors. The contents of this report do not necessarily reflect the views of NUATRC, the U.S. EPA, or any of the private industry sponsors. Additional direct and indirect financial support was provided by the Houston Endowment Inc, Baylor College of Medicine Chronic Disease Prevention and Control Research Center, University of Houston Institute for Multi-dimensional Air Quality Studies, University of Texas School of Public Health Department of Biostatistics, and Texas Environmental Leadership Program.

The authors also wish to acknowledge and thank the following individuals and institutions, listed alphabetically, for their valuable contributions to this project: Jonathan A. Garrett, BS, for his diligent and careful assessment of the quality and completeness of the observed data; Lilian Y. Mitchell, for her help with editing and proofing the report; and Shantikumar S. Ningthoujam, ME, MS, who developed many of the initial databases and geospatial methodology for the one-month pilot study.

We'd also like to thank numerous individuals with the Texas Commission on Environmental Quality who supplied the team the observed pollutant and meteorological data from Harris County monitors, answered questions about monitor performance, and supplied us with scripts and other materials that were very helpful in developing aspects of the study methodology. In particular, we'd like to thank Amir Poursamadi who orchestrated getting us the data, along with Melanie Hotchkiss, Heather Stewart, Ken Rozacky, Raj Nadkami, Mark Estes, and Bryan Lambeth, each of whom contributed in his or her area of expertise.

The authors also wish to thank data managers with the State of Texas Department of State Health Services Texas Health Care Information Collection for their help in assessing the quality and completeness of the hospital admissions data, and the Baylor College of Medicine Institutional Review Board for their helpful suggestions and ongoing oversight of all aspects of the study.

REFERENCES

Alberdi Odriozola JC, Diaz Jimenez J, Montero Rubio JC, Miron Perez IJ, Pajares Ortiz MS, Ribera Rodrigues P, 1998. Air pollution and mortality in Madrid, Spain: A time-series analysis. Int Arch Occup Environ Health 71:543-9.

American Lung Association, 2001. Urban air pollution and health inequities: A workshop report. Environ Health Perspect 109 Suppl 3:357-74.

Andria G, Cavone G, Lanzolla AML, 2008. Modeling study for assessment and forecasting variation of urban air pollution. Measurement 41:222-229.

Appel KW, Gilliland AB, Sarwar G, Gilliam RC, 2007. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance. Part I-Ozone. Atmos Environ 41:9603-9615.

Ballester F, Tenias JM, Perez-Hoyos S, 2001. Air pollution and emergency hospital admissions for cardiovascular diseases in Valencia, Spain. J Epidemiol Community Health 55:57-65.

Barnett AG, Williams GM, Schwartz J, Best TL, Neller AH, Petroeschevsky AL, Simpson RW, 2006. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australian and New Zealand cities. Environ Health Perspect 114:1018-23.

Bateson TF, Coull BA, Hubbell B, Ito K, Jerrett M, Lumley T, Thomas D, Vedal S, Ross M, 2007. Panel discussion review: Session three--issues involved in interpretation of epidemiologic analyses--statistical modeling. J Expo Sci Environ Epidemiol 17 Suppl 2:S90-6.

Bell ML, Dominici F, 2008. Effect modification by community characteristics on the short-term effects of ozone exposure and mortality in 98 US communities. Am J Epidemiol 167:986-97.

Blumenstock J, Fagliano J, Bresnitz E, 2000. The Dover township childhood cancer investigation. N J Med 97:25-30.

Braga AL, Saldiva PH, Pereira LA, Menezes JJ, Conceicao GM, Lin CA, Zanobetti A, Schwartz J, Dockery DW, 2001. Health effects of air pollution exposure on children and adolescents in Sao Paulo, Brazil. Pediatr Pulmonol 31:106-13.

Briggs D, 2005. The role of GIS: Coping with space (and time) in air pollution exposure assessment. J Toxicol Environ Health A 68:1243-61.

Brown JS, Graham JA, Chen LC, Postlethwait EM, Ghio AJ, Foster WM, Gordon T, 2007. Panel discussion review: Session four--assessing biological plausibility of epidemiological findings in air pollution research. J Expo Sci Environ Epidemiol 17 Suppl 2:S97-105.

Brown PJ, Le ND, Zidek JV, 1994. Multivariate spatial interpolation and exposure to air pollutants. Can J Stat 22:489-509.

Brunekreef B, Holgate ST, 2002. Air pollution and health. Lancet 360:1233-42.

Buckeridge DL, Glazier R, Harvey BJ, Escobar M, Amrhein C, Frank J, 2002. Effect of motor vehicle emissions on respiratory health in an urban area. Environ Health Perspect 110:293-300.

Bullard RD, 1983. Solid waste sites and the black Houston community. Sociol Inq 53:273-288.

Burnett RT, Brook JR, Yung WT, Dales RE, Krewski D, 1997. Association between ozone and hospitalization for respiratory diseases in 16 Canadian cities. Environ Res 72:24-31.

Burnett RT, Dales RE, Brook JR, Raizenne ME, Krewski D, 1997. Association between ambient carbon monoxide levels and hospitalizations for congestive heart failure in the elderly in 10 Canadian cities. Epidemiology 8:162-7.

Burnett RT, Smith-Doiron M, Stieb D, Cakmak S, Brook JR, 1999. Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. Arch Environ Health 54:130-9.

Byun D, Schere KL, 2006. Review of the governing equations, computational algorithms, and other components

of the models-3 Community Multiscale Air Quality (CMAQ) modeling system. Appl Mech Rev 59:51-76.

Byun DW, Ching JKS. Science algorithms of the EPA Models 3 Community Multiscale Air Quality (CMAQ) Modeling System. EPA600/R99/030, 1999.

Byun DW, Kim ST, Kim SB, 2007. Evaluation of air quality models for the simulation of a high ozone episode in the Houston metropolitan area. Atmos Environ 41:837-853.

Byun DW, Lacser A, Yamartino R, Zannetti P. 2003. Eulerian dispersion models. In: Air Quality Modeling: Theories, Methodologies, Computational Techniques, and Available Databases and Software, Vol I: Fundamentals (Zannetti P, ed). EnviroComp Institute and the Air & Waste Management Association.

California Air Resources Board. AB 2588 California "Hot Spots" Program. (2006). www.arb.ca.gov/ab2588/ab2588.htm. Accessed October 6, 2006.

Chan CC, Chuang KJ, Chien LC, Chen WJ, Chang WT, 2006. Urban air pollution and emergency admissions for cerebrovascular diseases in Taipei, Taiwan. Eur Heart J 27:1238-44.

Chang S, Allen DT, 2006. Atmospheric chlorine chemistry in southeast Texas: impacts on ozone formation and control. Environ Sci Technol 40:251-62.

Chen L, Mengersen K, Tong S, 2007. Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia. Sci Total Environ 373:57-67.

Cheng FY, Byun DW, 2008. Application of high resolution land use and land cover data for atmospheric modeling in the Houston-Galveston metropolitan area, Part I: Meteorological simulation results. Atmos Environ 42:7795-7811.

Ching J, Dupont S, Gilliam R, Burian S, Tang R. Neighborhood scale air quality modeling in Houston using urban canopy parameters in MM5 and CMAQ with improved characterization of mesoscale lake-land breeze circulation. 5th Symposium on the Urban Environment, 2004; 369-377.

Ching J, Herwehe J, Swall J, 2006. On joint deterministic grid modeling and sub-grid variability conceptual framework for model evaluation. Atmos Environ 40:4935-4945.

Ching J, Isakov V, Majeed M. Incorporating sub-grid variability concentration distributions with CMAQ. 4th Community Modeling and Analyses System (CMAS) Conference, Chapel Hill, NC, September 26-28, 2005; www.cmascenter.org/conference/2005/ppt/5_4.pdf.

Ching J, Isakov V, Majeed M, Irwin S. An approach for incorporating sub-grid variability (SGV) information into air quality modeling. 14th Joint Conference on the Applications of Air Pollution Meteorology with the Air and Waste Management Association, Atlanta, GA, January 30-February 2, 2006.

Ching J, Majeed M, Isakov V, Khlystov A. Fine scale air quality modeling using a hybrid dispersion and CMAQ modeling approach: An example application in Wilmington, DE. 5th Community Modeling and Analyses System (CMAS) Conference, Chapel Hill, NC, October 16-18, 2006; www.cmascenter.org/conference/2006/ppt/session6/ching.ppt.

Ching J, Pierce T, Palma T, Hutzell W, Tang R, Cimorelli A, Herwehe J. Linking air toxics concentration from CMAQ to the HAPEM5 exposure model at neighborhood scales for the Philadelphia area. 84th American Meterological Society Annual Meeting, Vancouver, British Columbia, Canada, 2004.

Cook R, Isakov V, Touma JS, Benjey W, Thurman J, Kinnee E, Ensley D, 2008. Resolving local-scale emissions for modeling air quality near roadways. J Air Waste Manag Assoc 58:451-61.

Dockery DW, Pope CA, 3rd, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Jr, Speizer FE, 1993. An association between air pollution and mortality in six US cities. N Engl J Med 329:1753-9.

Dolinoy DC, Miranda ML, 2004. GIS modeling of air toxics releases from TRI-reporting and non-TRI-reporting facilities: Impacts for environmental justice. Environ Health Perspect 112:1717-24.

Dominici F, McDermott A, Zeger SL, Samet JM, 2003. Airborne particulate matter and mortality: Timescale effects in four US cities. Am J Epidemiol 157:1055-65.

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM, 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295:1127-34.

Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet JM, 2004. Hierarchical bivariate time series models: A combined analysis of the effects of particulate matter on morbidity and mortality. Biostatistics 5:341-60.

Dunteman GH. Principal Components Analysis. Newbury Park, CA:SAGE Publications, Inc., 1989.
Elliott P, Wakefield JC, Best NG, Briggs DC, eds. 2000. Spatial Epidemiology: Methods and Applications. New York, NY:Oxford University Press, 2000.

European Commission Directorate-General XI. Ambient Air Pollution: Carbon Monoxide (Draft Version 5.2). Apeldoorn, The Netherlands, 1999; http://ec.europa.eu/environment/air/pdf/pp_co.pdf.

Furtaw EJ, Jr., 2001. An overview of human exposure modeling activities at the USEPA's National Exposure Research Laboratory. Toxicol Ind Health 17:302-14.

Fusco D, Forastiere F, Michelozzi P, Spadea T, Ostro B, Arca M, Perucci CA, 2001. Air pollution and hospital admissions for respiratory conditions in Rome, Italy. Eur Respir J 17:1143-50.

Gego E, Gilliland A, Godowitch J, Rao ST, Porter PS, Hogrefe C, 2008. Modeling analyses of the effects of changes in nitrogen oxides emissions from the electric power sector on ozone levels in the eastern United States. J Air Waste Manag Assoc 58:580-8.

Gego E, Steven Porter P, Hogrefe C, Irwin JS, 2006. An objective comparison of CMAQ and REMSAD performances. Atmos Environ 40:4920-4934.

Georgopoulos PG, Wang SW, Vyas VM, Sun Q, Burke J, Vedantham R, McCurdy T, Ozkaynak H, 2005. A source-to-dose assessment of population exposures to fine PM and ozone in Philadelphia, PA, during a summer 1999 episode. J Expo Anal Environ Epidemiol 15:439-457.

Green Media Toolshed. Scorecard: The Pollution Information Site, 2008; www.scorecard.org; Accessed September 29, 2008.

Greenland S, 2001. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. Int J Epidemiol 30:1343-50.

Hakami A, Henze DK, Seinfeld JH, Singh K, Sandu A, Kim S, Byun D, Li Q, 2007. The adjoint of CMAQ. Environ Sci Technol 41:7807-7817.

Health Effects Institute. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institute's Particle Epidemiology Reanalysis Project. Cambridge, MA: Health Effects Institute, 2000; http://pubs.healtheffects.org/view.php?id=6.

Herbert R, Moline J, Skloot G, Metzger K, Baron S, Luft B, Markowitz S, Udasin I, Harrison D, Stein D, Todd A, Enright P, Stellman JM, Landrigan PJ, Levin SM, 2006. The World Trade Center disaster and the health of workers: Five-year assessment of a unique medical screening program. Environ Health Perspect 114:1853-8.

Hinwood AL, De Klerk N, Rodriguez C, Jacoby P, Runnion T, Rye P, Landau L, Murray F, Feldwick M, Spickett J, 2006. The relationship between changes in daily air pollution and hospitalizations in Perth, Australia 1992-1998: A case-crossover study. Int J Environ Health Res 16:27-46.

Irwin JS, Civerolo K, Hogrefe C, Appel W, Foley K, Swall J, 2008. A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-h ozone concentrations. Atmos Environ 42(21):5403-5412.

Isakov V, Irwin JS, Ching J, 2007. Using CMAQ for exposure modeling and characterizing the subgrid variability exposure estimates. J Appl Meteorol Climatol 46:1354-1371.

Ito K, Thurston GD, Silverman RA, 2007. Characterization of $PM_{2.5}$, gaseous pollutants, and meteorological interactions in the context of time-series health effects models. J Expo Sci Environ Epidemiol 17 Suppl 2:S45-60.
Jacquez GM, Greiling DA, 2003. Local clustering in breast, lung and colorectal cancer in Long Island, New York. Int J Health Geogr 2:3.

Janhall S, Hallquist M, 2005. A novel method for determination of size resolved, submicrometer particle traffic emission factors. Environ Sci Technol 39:7609-15.

Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, Brook JR, Finkelstein N, Finkelstein MM, 2007. Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada. J Toxicol Environ Health A 70:200-12.

Jerrett M, Burnett RT, Ma R, Pope CA, 3rd, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE, Thun MJ, 2005. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 16:727-36.

Jimenez-Guerrero P, Jorba O, Baldasano JM, Gasso S, 2008. The use of a modelling system as a tool for air quality management: Annual high-resolution simulations and evaluation. Sci Total Environ 390:323-340.

Kim JY, Burnett RT, Neas L, Thurston GD, Schwartz J, Tolbert PE, Brunekreef B, Goldberg MS, Romieu I, 2007. Panel discussion review: Session two--interpretation of observed associations between multiple ambient air pollutants and health effects in epidemiologic analyses. J Expo Sci Environ Epidemiol 17 Suppl 2:S83-9.

Kim JY, Grant L, Burnett RT, 2007. Special issue on interpretation of epidemiologic studies of multipollutant ambient air exposure and health effects. J Expo Sci Environ Epidemiol 17:81.

Kim S, Czader B, Jang M-D, In-Bo O, Cheng F-Y, Byun DW. Comparison of national and state emissions inventories used for Houston ozone simulations. 2006 U.S. Environmental Protection Agency Emissions Inventory Conference, New Orleans, LA, May 16-17, 2006; www.epa.gov/ttn/chief/conference/ei15/session9/kim_pres.pdf.

Ko FW, Tam W, Wong TW, Chan DP, Tung AH, Lai CK, Hui DS, 2007. Temporal relationship between air pollutants and hospital admissions for chronic obstructive pulmonary disease in Hong Kong. Thorax 62:779-84.

Koken PJ, Piver WT, Ye F, Elixhauser A, Olsen LM, Portier CJ, 2003. Temperature, air pollution, and hospitalization for cardiovascular diseases among elderly people in Denver. Environ Health Perspect 111:1312-7.

Krewski D, Burnett RT, Goldberg MS, Hoover BK, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH, 2003. Overview of the reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality. J Toxicol Environ Health A 66:1507-51.

Laden F, Schwartz J, Speizer FE, Dockery DW, 2006. Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med 173:667-72.

Lanki T, Pekkanen J, Aalto P, Elosua R, Berglind N, D'Ippoliti D, Kulmala M, Nyberg F, Peters A, Picciotto S, Salomaa V, Sunyer J, Tiittanen P, von Klot S, Forastiere F, 2006. Associations of traffic related air pollutants with hospitalisation for first acute myocardial infarction: The HEAPSS study. Occup Environ Med 63:844-51.

Lee IM, Tsai SS, Chang CC, Ho CK, Yang CY, 2007. Air pollution and hospital admissions for chronic obstructive pulmonary disease in a tropical city: Kaohsiung, Taiwan. Inhal Toxicol 19:393-8.

Lee IM, Tsai SS, Ho CK, Chiu HF, Yang CY, 2007. Air pollution and hospital admissions for congestive heart failure in a tropical city: Kaohsiung, Taiwan. Inhal Toxicol 19:899-904.

Lee JT, Son JY, Kim H, Kim SY, 2006. Effect of air pollution on asthma-related hospital admissions for children by socioeconomic status associated with area of residence. Arch Environ Occup Health 61:123-30.

Leikauf GD, 2002. Hazardous air pollutants and asthma. Environ Health Perspect 110 Suppl 4:505-26.

Liao D, Peuquet DJ, Duan Y, Whitsel EA, Dou J, Smith RL, Lin HM, Chen JC, Heiss G, 2006. GIS approaches for the estimation of residential-level ambient PM concentrations. Environ Health Perspect 114:1374-80.

Lin M, Chen Y, Burnett RT, Villeneuve PJ, Krewski D, 2003. Effect of short-term exposure to gaseous pollution on asthma hospitalisation in children: A bi-directional case-crossover analysis. J Epidemiol Community Health 57:50-5.

Linn WS, Szlachcic Y, Gong H, Jr, Kinney PL, Berhane KT, 2000. Air pollution and daily hospital admissions in metropolitan Los Angeles. Environ Health Perspect 108:427-34.

Lipfert FW. Air Pollution and Community Health: A Critical Review and Data Sourcebook. New York, NY:Van Nostrand Reinhold Publishers, 1994.

Luecken DJ, Hutzell WT, Gipson GL, 2006. Development and analysis of air quality modeling simulations for hazardous air pollutants. Atmos Environ 40:5087-5096.

Maantay J, 2007. Asthma and air pollution in the Bronx: Methodological and data considerations in using GIS for

environmental justice and health research. Health Place 13:32-56.

Maantay J, 2001. Zoning, equity, and public health. Am J Public Health 91:1033-41.

Magas OK, Gunter JT, Regens JL, 2007. Ambient air pollution and daily pediatric hospitalizations for asthma. Environ Sci Pollut Res Int 14:19-23.

Maheswaran R, Elliott P, 2003. Stroke mortality associated with living near main roads in England and Wales: A geographical study. Stroke 34:2776-80.

Maheswaran R, Haining RP, Brindley P, Law J, Pearson T, Fryers PR, Wise S, Campbell MJ, 2005. Outdoor air pollution and stroke in Sheffield, United Kingdom: A small-area level geographical study. Stroke 36:239-43.

Maheswaran R, Haining RP, Brindley P, Law J, Pearson T, Fryers PR, Wise S, Campbell MJ, 2005. Outdoor air pollution, mortality, and hospital admissions from coronary heart disease in Sheffield, UK: A small-area level ecological study. Eur Heart J 26:2543-9.

Maheswaran R, Haining RP, Pearson T, Law J, Brindley P, Best NG, 2006. Outdoor NO_x and stroke mortality: Adjusting for small area level smoking prevalence using a Bayesian approach. Stat Methods Med Res 15:499-516.

Marshall JD, Nethery E, Brauer M, 2008. Within-urban variability in ambient air pollution: Comparison of estimation methods. Atmos Environ 42:1359-1369.

Martello DV, Pekney NJ, Anderson RR, Davidson CI, Hopke PK, Kim E, Christensen WF, Mangelson NF, Eatough DJ, 2008. Apportionment of ambient primary and secondary fine particulate matter at the Pittsburgh National Energy Laboratory particulate matter characterization site using positive matrix factorization and a potential source contributions function analysis. J Air Waste Manag Assoc 58:357-68.

Martins LC, Pereira LA, Lin CA, Santos UP, Prioli G, Luiz Odo C, Saldiva PH, Braga AL, 2006. The effects of air pollution on cardiovascular diseases: lag structures. Rev Saude Publica 40:677-83.

Mayor's Task Force on the Health Effects of Air Pollution. A Closer Look at Air Pollution in Houston: Identifying Priority Health Risks. Houston: Institute for Health Policy, 2006; www.sph.uth.tmc.edu/uploadedFiles/Centers/IHP/UTReportrev.pdf.

McMillan NJ, Holland DM, Morara M, Feng J, in press. Combining numerical model output and particulate data using Bayesian space-time modeling. Environmetrics.

Medina-Ramon M, Schwartz J, 2008. Who is more vulnerable to die from ozone air pollution? Epidemiology 19:672-679.

Medina-Ramon M, Zanobetti A, Schwartz J, 2006. The effect of ozone and PM_{10} on hospital admissions for pneumonia and chronic obstructive pulmonary disease: A national multicity study. Am J Epidemiol 163:579-88.

Moolgavkar SH, 2000. Air pollution and hospital admissions for diseases of the circulatory system in three US metropolitan areas. J Air Waste Manag Assoc 50:1199-206.

Morandi MT, Stock TH. Houston Exposure to Air Toxics Study (HEATS), 2008; www.tceq.state.tx.us/implementation/tox/research/heats.html.

Morello-Frosch R, Pastor M, Jr., Porras C, Sadd J, 2002. Environmental justice and regional inequality in southern California: Implications for future research. Environ Health Perspect 110 Suppl 2:149-54.

Morello-Frosch RA, Woodruff TJ, Axelrad DA, Caldwell JC, 2000. Air toxics and health risks in California: the public health implications of outdoor concentrations. Risk Anal 20:273-91.

Morgan G, Corbett S, Wlodarczyk J, 1998. Air pollution and hospital admissions in Sydney, Australia, 1990 to 1994. Am J Public Health 88:1761-6.

Morris RD, 2001. Airborne particulates and hospital admissions for cardiovascular disease: A quantitative review of the evidence. Environ Health Perspect 109 Suppl 4:495-500.

Morris RD, Naumova EN, Munasinghe RL, 1995. Ambient air pollution and hospitalization for congestive heart failure among elderly people in seven large US cities. Am J Public Health 85:1361-5.

Napelenok SL, Cohan DS, Odman MT, Tonse S, 2008. Extension and evaluation of sensitivity analysis capabilities in a photochemical model. Environmental Modelling and Software 23:994-999.

Neuberger M, Rabczenko D, Moshammer H, 2007. Extended effects of air pollution on cardiopulmonary mortality in Vienna. Atmos Environ 41:8549-8556.

Nuckols JR, Ward MH, Jarup L, 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect 112:1007-15.

Ozkaynak H, Glenn B, Qualters JR, Strosnider H, McGeehin MA, Zenick H, 2009. Summary and findings of the EPA and CDC symposium on air pollution exposure and health. J Expo Sci Environ Epidemiol 19:19-29.

Payne-Sturges DC, Burke TA, Breysse P, Diener-West M, Buckley TJ, 2004. Personal exposure meets risk assessment: A comparison of measured and modeled exposures and risks in an urban community. Environ Health Perspect 112:589-98.

Perlin SA, Wong D, Sexton K, 2001. Residential proximity to industrial sources of air pollution: interrelationships among race, poverty, and age. J Air Waste Manag Assoc 51:406-21.

Poloniecki JD, Atkinson RW, de Leon AP, Anderson HR, 1997. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London, UK. Occup Environ Med 54:535-40.

Pope CA, 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD, 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287:1132-41.

Pope CA, 3rd, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ, 2004. Cardiovascular mortality and long-term exposure to particulate air pollution: Epidemiological evidence of general pathophysiological pathways of disease. Circulation 109:71-7.

Radian Corporation. Emissions inventory and dispersion modeling reports: A report to the Texas Natural Resources Conservation Commission, Office of Air Quality/Toxicology & Risk Assessment Section. Austin, TX, 1995.

Ricci PF, Straja SR, 2006. Hospital admissions and fine particulate air pollution. JAMA 296:1966-7.

Riccio A, Barone G, Chianese E, Giunta G, 2006. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. Atmos Environ 40:554-566.

Saez M, Tobias A, Munoz P, Campbell MJ, 1999. A GEE moving average analysis of the relationship between air pollution and mortality for asthma in Barcelona, Spain. Stat Med 18:2077-86.

Sanhueza PA, Reed GD, Davis WT, Miller TL, 2003. An environmental decision-making tool for evaluating ground-level ozone-related health effects. J Air Waste Manag Assoc 53:1448-59.

Sarnat JA, Wilson WE, Strand M, Brook J, Wyzga R, Lumley T, 2007. Panel discussion review: Session one--exposure assessment and related errors in air pollution epidemiologic studies. J Expo Sci Environ Epidemiol 17 Suppl 2:S75-82.

Schwartz J, 1999. Air pollution and hospital admissions for heart disease in eight US counties. Epidemiology 10:17-22.

Scoggins A, Kjellstrom T, Fisher G, Connor J, Gimson N, 2004. Spatial analysis of annual air pollution exposure and mortality. Sci Total Environ 321:71-85.

Seigneur C. Air Toxics Modeling. Current Status, Challenges and Prospects. CRC Project Number A-49: Coordinating Research Council, Inc., 2005; www.crcao.com/reports/recent_reports_and_study_results.htm.

Seigneur C, Pun B, Loman K, Wu S. Air Toxics Modeling. CP079-02-3. San Ramon: Atmospheric and Environmental Research, 2002.

Sheppard E, Leitner H, McMaster RB, Tian H, 1999. GIS-based measures of environmental equity: Exploring their sensitivity and significance. J Expo Anal Environ Epidemiol 9:18-28.

Sheppard L, Levy D, Norris G, Larson TV, Koenig JQ, 1999. Effects of ambient air pollution on nonelderly asthma hospital admissions in Seattle, Washington, 1987-1994. Epidemiology 10:23-30.

Sokhi RS, San Jose R, Kitwiroon N, Fragkou E, Perez JL, Middleton DR, 2006. Prediction of ozone levels in London using the MM5-CMAQ modelling system. Environ Model Software 21:566-576.

South Coast Air Quality Management District. Estimation of Health Benefits of South Coast Air Basin 2007 AQMP/SIP Oceangoing Marine Vessel Control Measures, 2007; www.epa.gov/oms/regs/nonroad/marine/ci/mvbenefits20071018-b.pdf.

South Coast Air Quality Management District. Multiple Air Toxics Exposure Study in the South Coast Air Basin (MATES-II). Diamond Bar, CA: South Coast Air Quality Management District, 1999.

Stein AF, Isakov V, Godowitch J, Draxler RR, 2007. A hybrid modeling approach to resolve pollutant concentrations in an urban area. Atmos Environ 41:9410-9426.

Stieb DM, Burnett RT, Smith-Doiron M, Brion O, Shin HH, Economou V, 2008. A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. J Air Waste Manag Assoc 58:435-50.

Sunderland EM, Cohen MD, Selin NE, Chmura GL, 2008. Reconciling models and measurements to assess trends in atmospheric mercury deposition. Environ Pollut 156:526-535.

Texas Commission on Environmental Quality. Air Pollutant Watch List: Regulating Chemicals in Specific Areas, 2008; www.tceq.state.tx.us/nav/eq/eq_aq_apwl.html; Accessed September 26, 2008.

Texas Commission on Environmental Quality. Effects Screening Level Lists, 2008; www.tceq.state.tx.us/implementation/tox/esl/list_main.html; Accessed September 29, 2008.

Texas Commission on Environmental Quality. TexAQS II Air Quality Field Study. (2007). www.tceq.state.tx.us/implementation/air/airmod/committee/scc.html. Accessed June 10 2007.

The University of Texas at Austin. The Texas Air Quality Study 2000. (2000). www.utexas.edu/research/ceer/texaqs/participants/about.html. Accessed June 10 2007.

Thurston GD, 2006. Hospital admissions and fine particulate air pollution. JAMA 296:1966-7.

Tolbert PE, Klein M, Peel JL, Sarnat SE, Sarnat JA, 2007. Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta. J Expo Sci Environ Epidemiol 17 Suppl 2:S29-35.

U.S. Census Bureau. Census. (2000). www.census.gov. Accessed June 10, 2007.

U.S. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System (BFRSS), 2008; www.cdc.gov/brfss; Accessed September 29, 2008.

U.S. Environmental Protection Agency. 1999 National-Scale Air Toxics Assessment. (2007). www.epa.gov/ttn/atw/nata1999/index.html. Accessed April 28, 2008.

U.S. Environmental Protection Agency. 1999 National-Scale Air Toxics Assessment. (2006). www.epa.gov/ttn/atw/nata1999/index.html. Accessed July 20, 2006.

U.S. Environmental Protection Agency. 1999 National Emission Inventory Documentation and Data - Final Version 3.0, 2001; www.epa.gov/ttn/chief/net/1999inventory.html.

U.S. Environmental Protection Agency. Air Quality Models. (2008). www.epa.gov/scram001/aqmindex.htm. Accessed April 22, 2008.

U.S. Environmental Protection Agency. AirData: Access to Air Pollution Data, 2008; www.epa.gov/air/data/; Accessed September 29, 2008.

U.S. Environmental Protection Agency. Example Application of Modeling Toxic Air Pollutants in Urban Areas. EPA-454/R-02-003. Research Triangle Park: Office of Air and Radiation, Office of Air Quality Planning and Standards, 2002.

U.S. Environmental Protection Agency. Industrial Source Complex (ISC) Dispersion, Model User's Guide, second edition (revised). EPA-450/4-88-002a. Research Triangle Park, NC: U.S. Environmental Protection Agency (USEPA), 1987.

U.S. Environmental Protection Agency. Technology Transfer Network Air Toxics Web Site. (2008). www.epa.gov/ttn/atw/allabout.html. Accessed September 26, 2007.

U.S. Environmental Protection Agency. User's Guide for the Industrial Source Complex (ISC3) Dispersion Models, Vol I - User Instructions and Vol II - Description of Model Algorithms. EPA-454/B-95-003a,b. Research Triangle Park, NC, 1995.

U.S. Environmental Protection Agency. User's Guide to RAM, second edition. EPA/600/8-87/046. Research Triangle Park, NC: Atmospheric Sciences Research Laboratory, Office of Research and Development, 2002.

U.S. Office of the President. Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations. 1994. In: Executive Order 12898.

University of Houston Institute for Multi-dimensional Air Quality Studies. University of Houston East Texas Air Quality Forecasting System, 2008; www.imaqs.uh.edu/aqfmain.htm; Accessed September 30, 2008.

Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Hoboken, NJ:John Wiley & Sons, Inc., 2004.

Webster M, Nam J, Kimura Y, Jeffries H, Vizuete W, Allen DT, 2007. The effect of variability in industrial emissions on ozone formation in Houston, Texas. Atmos Environ 41:9580-9593.

Weisel C, Zhang J, Turpin B, Morandi M, Colome S, Stock T, Spektor D. Relationships of Indoor, Outdoor, and Personal Air (RIOPA): Part I. Data Collection and Descriptive Analyses. HEI Research Report 130; NUATRC Research Report 7. Boston, MA, and Houston, TX: Health Effects Institute and Mickey Leland National Urban Air Toxics Research Center, 2005; www.healtheffects.org/Pubs/RIOPA-I.pdf.

Weisel CP, 2002. Assessing exposure to air toxics relative to asthma. Environ Health Perspect 110 Suppl 4:527-37.

Wellenius GA, Schwartz J, Mittleman MA, 2005. Air pollution and hospital admissions for ischemic and hemorrhagic stroke among Medicare beneficiaries. Stroke 36:2549-53.

Wellenius GA, Schwartz J, Mittleman MA, 2006. Particulate air pollution and hospital admissions for congestive heart failure in seven United States cities. Am J Cardiol 97:404-8.

Weschler CJ, 2006. Ozone's impact on public health: Contributions from indoor exposures to ozone and products of ozone-initiated chemistry. Environ Health Perspect 114:1489-96.

Whitworth KW, Symanski E, Coker AL, 2008. Childhood lymphohematopoietic cancer incidence and hazardous air pollutants in southeast Texas, 1995-2004. Environ Health Perspect 116:1576-80.

Woodruff TJ, Axelrad DA, Caldwell J, Morello-Frosch R, Rosenbaum A, 1998. Public health implications of 1990 air toxics concentrations across the United States. Environ Health Perspect 106:245-51.

Woodruff TJ, Parker JD, Kyle AD, Schoendorf KC, 2003. Disparities in exposure to air pollution during pregnancy. Environ Health Perspect 111:942-6.

Wyat Appel K, Bhave PV, Gilliland AB, Sarwar G, Roselle SJ, 2008. Evaluation of the community multiscale air quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II-particulate matter. Atmos Environ 42:6057-6066.

Xie Y, Berkowitz CM, 2007. The use of conditional probability functions and potential source contribution functions to identify source regions and advection pathways of hydrocarbon emissions in Houston, Texas. Atmos Environ 41:5831-5847.

Yang CY, Chen CC, Chen CY, Kuo HW, 2007. Air pollution and hospital admissions for asthma in a subtropical city: Taipei, Taiwan. J Toxicol Environ Health A 70:111-7.

Yang CY, Chen CJ, 2007. Air pollution and hospital admissions for chronic obstructive pulmonary disease in a subtropical city: Taipei, Taiwan. J Toxicol Environ Health A 70:1214-9.

Yang CY, Chen YS, Yang CH, Ho SC, 2004. Relationship between ambient air pollution and hospital admissions for cardiovascular diseases in Kaohsiung, Taiwan. J Toxicol Environ Health 67:483-493.

Yang W, Jennison BL, Omaye ST, 1998. Cardiovascular disease hospitalization and ambient levels of carbon monoxide. J Toxicol Environ Health 55:185-96.

Yanosky JD, Paciorek CJ, Schwartz J, Laden F, Puett R, Suh HH, 2008. Spatio-temporal modeling of chronic PM_{10} exposure for the Nurses' Health Study. Atmos Environ 42:4047-4062.

Zandbergen PA, Chakraborty J, 2006. Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. Int J Health Geogr 5:23.

Zanobetti A, Schwartz J, 2008. Mortality displacement in the association of ozone with mortality: An analysis of 48 cities in the United States. Am J Respir Crit Care Med 177:184-189.

Zanobetti A, Schwartz J, 2000. Race, gender, and social status as modifiers of the effects of PM_{10} on mortality. J Occup Environ Med 42:469-74.

Zanobetti A, Schwartz J, Dockery DW, 2000. Airborne particles are a risk factor for hospital admissions for heart and lung disease. Environ Health Perspect 108:1071-7.

Zanobetti A, Schwartz J, Gold D, 2000. Are there sensitive subgroups for the effects of airborne particles? Environ Health Perspect 108:841-5.

Zanobetti A, Schwartz J, Samoli E, Gryparis A, Touloumi G, Peacock J, Anderson RH, Le Tertre A, Bobros J, Celko M, Goren A, Forsberg B, Michelozzi P, Rabczenko D, Hoyos SP, Wichmann HE, Katsouyanni K, 2003. The temporal pattern of respiratory and heart disease mortality in response to air pollution. Environ Health Perspect 111:1188-93.

Zhang C, 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ Pollut 142:501-11.

Zhang F, Bei N, Nielsen-Gammon JW, Li G, Zhang R, Stuart A, Aksoy A, 2007. Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts. J Geophys Res D: Atmos 112.

Zhang Y, Liu P, Pun B, Seigneur C, 2006. A comprehensive performance evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study episode-Part I: Evaluation protocols, databases, and meteorological predictions. Atmos Environ 40:4825-4838.

Zhang Y, Liu P, Pun B, Seigneur C, 2006. A comprehensive performance evaluation of MM5-CMAQ for the summer 1999 southern oxidants study episode, Part III: Diagnostic and mechanistic evaluations. Atmos Environ 40:4856-4873.

Zhang Y, Liu P, Queen A, Misenis C, Pun B, Seigneur C, Wu SY, 2006. A comprehensive performance evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study episode-Part II: Gas and aerosol predictions. Atmos Environ 40:4839-4855.

Zhou Y, Levy JI, 2007. Factors influencing the spatial extent of mobile source air pollution impacts: A meta-analysis. BMC Public Health 7:89.

Zhu X, Fan Z, Wu X, Meng Q, Wang Sw, Tang X, Ohman-Strickland P, Georgopoulos P, Zhang J, Bonanno L, Held J, Lioy P, 2008. Spatial variation of volatile organic compounds in a "Hot Spot" for air pollution. Atmos Environ 42:7329-7338.

ABOUT THE AUTHORS

Alphabetical by last name

Daewon Byun, PhD, assumed the position of Group Leader for the Air Quality Modeling Team at the Air Resources Laboratory, Office of Atmospheric Research, at the National Oceanic Atmospheric Administration (NOAA) in Washington, DC, in early 2009. He continues his appointment as professor in the Department of Geosciences, with a joint appointment in the Department of Chemistry, University of Houston (UH), where he was since 2001 until his 2009 departure to DC. Dr. Byun was also director for the Institute for Multidimensional Air Quality Studies (IMAQS) at UH from 2001 through 2008, and oversaw all of the modeling for our project. The UH-IMAQS team performs atmospheric modeling and monitoring studies specializing in urban and regional air quality. Prior to joining UH, Dr. Byun was the science team leader of the U.S. Environmental Protection Agency (EPA) Models-3 Community Multiscale Air Quality (CMAQ) model development project. He was a senior science specialist at the Computer Sciences Corporation, a staff meteorologist at the ERT environmental consulting company, and a physical

scientist for NOAA, on assignment to U.S. EPA. Dr. Byun was awarded the EPA Silver Medal for Superior Service (development of Models-3 CMAQ) in 1999. He received the NOAA Air Resources Laboratory "Paper of the Year" Award in 2000. He served as the Science Board member of the NCAR and NOAA Weather Research and Forecasting (WRF) project and the External Advisory Committee for the Community Modeling and Analysis System (CMAS) until 2004. Also, he was the member of the American Meteorological Society (AMS) Committee for the Meteorological Aspects of Air Pollution during 2003-2006, and AMS Committee for the Atmospheric Chemistry during 2004-2007. Until his appointment at NOAA, he also served as an Executive Board member of the Texas Commission for Environmental Quality (TCEQ) Science Steering Committee, Advisory Board for Mothers for Clean Air, and the Community Advisory Panel for Environmental Health Houston.

Wenyaw Chan, PhD, is currently professor of biostatistics at the School of Public Health, University of Texas-Health Science Center at Houston, where he started his faculty appointment in 1989. Committed to interdisciplinary research, he has participated in more than two dozen public health research projects. His methodological research includes design and analysis of longitudinal studies, stochastic models and the development of biostatistical methods. In addition to biostatistical methodology, Dr. Chan's areas of research involvement include Alzheimer's disease, behavioral medicine, cancer, cardiovascular disease, dementia, environmental and occupational health, health outcome research, health promotion, human biology, nursing research, ophthalmology and stroke. He has authored or co-authored more than 80 peer-reviewed articles in statistics, medical sciences or public health, and one textbook in mathematical statistics.

Jason K. S. Ching, PhD, is currently a physical scientist with the U.S. Environmental Protection Agency's (U.S. EPA) Atmospheric Modeling Division (AMD), National Exposure Research Laboratory (NERL) in Research Triangle Park, NC. From 1970 until July 2008, Dr. Ching was a meteorologist with the National Oceanic and Atmospheric Agency's (NOAA) Air Resources Laboratory (ARL), and between 1974 and 2008 he's been on interagency assignment to the U.S. EPA's AMD/NERL in NC. He received degrees in meteorology from the University of Hawaii (BS/1962), Penn State University (MS/1964) and University of Washington (PhD/1974). Dr. Ching has led and participated in major U.S. EPA field studies from 1974 to 1989 including RAPS, TPS and PEPE/NEROS. He also led EPA's AcidMODES RADM model evaluation field study for NAPAP, for which he received a Certificate of Recognition. Around 1989 Dr. Ching's research interests shifted to model development. He initiated and developed the conceptual design of the Community Multiscale Air Quality Modeling (CMAQ) system, led the development team and delivered the first implementation in 1998. He has also been involved in the development of modeling approaches for dry deposition, fugitive dust, and smoke emissions modeling from wild land and prescribed burns. He is now primarily engaged in the development of activities and methods directed towards promoting, developing and applying CMAQ and meteorological models (MM5 and WRF) at urban-to-neighborhood scales, with a focus on improving linkages between fine-grid CMAQ simulations and exposure models in urban areas. He recently developed the prototype National Urban Database and Assess Port Tools (NUDAPT) system to provide the infrastructure and resources for advanced meteorological and air quality modeling for urban applications. For his modeling efforts, he has received three bronze medals from the U.S. EPA. He has authored 150 journal articles, project reports and conference proceedings. Dr. Ching is engaged in several research collaborations on CMAQ, MM5, and WRF models, and participates actively in the AMS and European COST programs.

Violeta F. Coarfa, PhD, recently completed a postdoctoral fellowship at the Institute for Multi-dimensional Air Quality Studies (IMAQS) at the University of Houston. Her most recent research focused on measurements of atmospheric formaldehyde, working with Dr. Bernhard Rappenglueck. She has also been working with Green Environmental Consulting, Inc., where she assists with air emissions inventory calculations by compiling physical data for numerous compounds, as well as working on air permitting and dispersion modeling. She earned her PhD in atmospheric sciences from the University of Houston Geosciences Department, working under Daewon Byun, PhD. Her thesis focused on modeling air toxics in the Houston-Galveston area, with special emphasis on aromatic hydrocarbons, such as benzene, toluene, xylene isomers, and ethylbenzene. Dr. Coarfa also developed a computationally efficient model, termed CMAQ-HAP, which uses the oxidant fields created by the Community Multiscale Air Quality (CMAQ) model in conjunction with a modification in the chemical SAPRC mechanism to explicitly generate selected hazardous air pollutant (HAP) species. This model greatly reduces the time required to simulate many HAPs, without affecting the quality of the output.

Winifred J. Hamilton, PhD, is an assistant professor in medicine and neurosurgery at Baylor College of Medicine, and director of the Environmental Health Section of the Chronic Disease Prevention and Control Research Center. She is also a faculty member at Rice University, where she teaches a course on environmental health. She earned her graduate degrees from the University of Michigan, Rice University, and the Harvard School of Public Health, the latter in environmental health epidemiology. Her interests include geospatial analysis of environmental health indicators, the use of research to drive policy to improve public health, community-based participatory research, and quality-of-life initiatives at the neighborhood level. She has been program director of three collaborative pediatric environmental health symposia in the Houston region; is principal investigator of the 2007 geospatial analysis and investigative report, "Childhood Lead Poisoning in Galveston, Texas;" is co-author of the 2006 report, "The Control of Air Toxics: Toxicology, Motivation and Houston Implications," and is principal investigator of three lead-exposure studies in Houston, TX. For her work to improve pediatric environmental health, she received the U.S. EPA's 2008 Children's Environmental Health Champion Award. She currently serves on numerous Boards and task forces in the Houston-Galveston region addressing exposure, quality-of-life, and environmental justice issues. She is co-chair of the Medical Center Recycling Collaborative in the Texas Medical Center, and is principal investigator of an initiative to integrate environmental health into Baylor's medical school curriculum, and teaches environmental health in the APEX program at Baylor College of Medicine. She also chairs a collaborative team to establish a children's environmental health clinic in the Texas Medical Center. She is principal or co-investigator of numerous epidemiologic and public outreach initiatives, has published more than 50 peer-reviewed articles and chapters, and has co-authored or co-edited several books.

Younghun Han, PhD, received his MS in statistics from Yeungnam University, Kyoungsan, South Korea in August 1993, and an MS in statistics from Iowa State University in August 2001. He earned his PhD in biostatistics in May 2008 from The University of Texas School of Public Health, Health Science Center at Houston. He is proficient in use of various computer systems, programming languages, and statistical software packages. He has extensive experience in analyzing large datasets, and developed or refined a number of statistical procedures for working with grid-based geospatial air pollution and health effects data. He is currently a statistical analyst with the Department of Epidemiology at the University of Texas M.D. Anderson Cancer Center. Among his current projects and research interests include analysis of longitudinal data, meta-analysis of genome-wide association data for lung cancer, genetic analyses to identify susceptibility factors for complex diseases, and pharmacogenetic analyses to identify genotypic risk factors for adverse treatment outcomes.

DaeGyun Lee received his MS in environmental science from Kangwon National University, South Korea in February 2001, and subsequently worked at the National Institute of Environmental Research for five years (2001-2005). He is a currently PhD candidate and research assistant with the University of Houston Institute of Multi-dimensional Air Quality Studies (UH-IMAQS), which is part of the Department of Geosciences. He has expertise in sophisticated air quality models and is currently overseeing much of the daily air quality forecasting of criteria air pollutants using the version 4.4 of the U.S. Environmental Protection Agency's Community Multiscale Air Quality (CMAQ4.4) at the UH-IMAQS. In addition to the daily CMAQ4.4 runs, Mr. Lee has developed much of the automated statistical infrastructure that generates near-real time spatial and temporal comparisons between CMAQ-simulated output and hourly measurements from regional fixed-site monitors, which are available online.

Ricardo A. López holds a BS in civil engineering (Universidad de los Andes - Colombia) and a MS in environmental science (University of Houston - Clear Lake). He presently serves as a Senior Risk and Safety Consultant for IRC Risk and Safety LLC, Houston, TX. Since joining IRC in 2008, Mr. López has led efforts on two RAM (Reliability, Availability and Maintainability) analyses, a computational model for Medical Emergency Response, and an economic model for selection of drilling rig blow-out insurance. Until accepting the position at IRC, Mr. Lopez held a faculty position at Baylor College of Medicine in the Environmental Health Section of the Chronic Disease Prevention and Control Research Center, where he specialized in geospatial modeling and bioinformatics. Before joining Baylor, he was with Texas A&M University (Texas Sea Grant and the Texas Cooperative Extension) where he worked in geo-information technologies and website development. His experience with geographical information systems (GIS) includes design and implementation of geospatial databases and processing models in projects such as environmental impact assessment studies, air toxics mapping, hospital admissions geoaddressing, wetland loss monitoring, urban

sprawl, sea level rise response, and use/water quality modeling. Mr. López also has extensive experience with database development, spatial analysis, and interactive website design. In addition, he holds an adjunct faculty position (since 2001) with the University of Houston - Clear Lake, where he teaches basic and advanced GIS courses.

OTHER PUBLICATIONS RESULTING FROM THIS RESEARCH

No publications in peer-reviewed journals have yet resulted from this research. One abstract and two oral presentations have resulted from the pilot work.

Hamilton WJ, Byun D, Lopez RA, Coarfa VF, Han Y, Chan W, Ching JKS: Geospatial Analysis of Air Toxics and Hospital Admissions in Harris County, Texas. 17th Annual Conference of the International Society of Exposure Analysis (ISEA), Durham/Research Triangle Park, NC, October 14-18, 2007.

Hamilton WJ, Byun D, Chan W, Ching JKS, Han Y, Lopez RA, Coarfa VF, Lee DG: Preliminary Explorations into Using EPA's CMAQ Model and Hospital Admission Data to Identify Multipollutant "Hot Spots" of Concern in Harris County, Texas. U.S. Environmental Protection Agency Atmospheric Modeling Division, National Exposure Research Laboratory and Office of Research and Development (USEPA AMD/NERL/ORD) Seminar. Research Triangle Park, NC, October 17, 2008.

ABBREVIATIONS

AERMOD	American Meteorological Society/ Environmental Protection Agency Regulatory Model Improvement Committee Model
AIC	Akaike Information Criterion
AIRS	U.S. EPA's Aerometric Information Retrieval Service
APEX	Air Pollutant Exposure Model
APWL	Air Pollution Watch List
ASPEN	Assessment System for Population Exposure Nationwide
ATSDR	Agency for Toxic Substances and Disease Registry
BCM	Baylor College of Medicine
BEIS	EPA's Biogenic Emissions Inventory System
BIC	Bayesian Information Criterion
CAA	Clean Air Act
CAMS	Continuous Air Monitoring Stations
CAMx	Comprehensive Air Model with Extensions
CARB	California Air Resources Board
CMAS	Community Modeling & Analysis System
CAP	Criteria air pollutant
CB4	Chemical Bond 4
CMAQ	Community Multiscale Air Quality model
CMAQ-HAP	Community Multiscale Air Quality modified for selected HAPs
CDT	Central daylight time
CHG	Congestive Heart Failure
CO	Carbon monoxide
COPD	Chronic obstructive pulmonary disease
CST	Central standard time
CTM	Chemical transport model
DDM	Decoupled direct method
EBI	Euler Backward Iterative
EHS	Environmental Health Section
EMS-HAP	Emissions Modeling System for Hazardous Pollutants
EPA	U.S. Environmental Protection Agency
ESL	Effects Screening Level
FIPS	Federal Information Processing Standard
GCS	Geographic Coordinate System
GEE	Generalized Estimating Equation
GloBEIS	Global Biosphere Emissions and Interactions System
HAP	Hazardous air pollutant
HAPEM	Hazardous Air Pollutant Exposure Model
HARC	Houston Advanced Research Center
HCHD	Harris County Hospital District
HGA	Houston-Galveston Area
HRVOC	Highly reactive volatile organic compound
HYSPLIT	Hybrid Single-Particle Lagrangian Integrated Trajectory
ICD-9	International Classification of Diseases, 9th revision
IPR	Integrated process rate
IRB	Institutional Review Board
ISCST	Industrial Source Complex Short Term
JCAHO	Joint Council on Accreditation of Healthcare Organizations
LBJ	Lyndon B. Johnson hospital
LMM	Linear mixed-effects model
LSM	Modified Land-Surface Model
LULC	Land Use/Land Cover
MCIP	Meteorology-Chemistry Interface Processor
MM5	Fifth-Generation National Center for Atmospheric Research / Penn Stat Mesoscale Model
MOBILE6	EPA's Vehicle Emission Modeling Software, version 6

MRF	Medium Range Forecast
NAD	North American Datum
NATA	National-Scale Air Toxic Assessment
NAAQS	National Ambient Air Quality Standard
NEI	National Emissions Inventory
NMMAPS	National Morbidity, Mortality, and Air Pollution Study
NO	Nitrogen oxide
NO_2	Nitrogen dioxide
NO_x	Nitrogen oxides (NO + NO2)
NO_y	Total reactive nitrogen
NOAA	National Oceanic and Atmospheric Administration
O_3	Ozone
PAN	Nitric anhydride
PBL	Planetary Boundary Layer
PCA	Principal Components Analysis
$PM_{2.5}$	Particulate matter with an aerodynamic mass median diameter ° 2.5 µm
PM_{10}	Particulate matter with an aerodynamic mass median diameter ° 10 µm
ppb	Parts per billion
ppm	Parts per million (volume)
PPM	Piecewise Parabolic Method
PPN	Peroxypropionic nitric anhydride
PNC	Particle number concentration
RAM	Gaussian-Plume Multiple Source Air Quality Algorithm
REL	Reference Exposure Level
RRTM	Rapid Radiative Transfer Model
SHEDS	Stochastic Human Exposure and Dose Simulation
SMOKE	Sparse Matrix Operator Kernal Emissions Modeling
SAPRC	Statewide Air Pollution Research Center chemical mechanism
SGV	Subgrid variability
SIP	State Implementation Plan
SO_2	Sulfur dioxide
SPCS	State Plane Coordinate System
STAR*Map	Southeast Texas Addressing and Referencing Map
TCEQ	Texas Commission on Environmental Quality
TERC	Texas Environmental Research Consortium
TexAQS-I	Texas Air Quality Study I, 2000
TexAQS-II	Texas Air Quality Study II, 2006
THCIC	Texas Health Care Information Collection
TDSHS	Texas Department of State Health Services
TEI	Texas Emissions Inventory
TFS	Texas Forest Service
TRI	Toxic Release Inventory
UFP	Ultrafine particles
UH-IMAQS	University of Houston Institute for Multi-dimensional Air Quality Studies
URL	Universal Resource Locator
UTC	Coordinated Universal Time
VBA	Visual Basic for Applications

BOARD OF DIRECTORS

Wilma Delaney
Dow Chemical Company (Retired)

Jane L. Delgado
National Alliance for Hispanic Health

Shawn L. Gerstenberger
University of Nevada Las Vegas

John E. Hiatt (Treasurer)
Quest Diagnostics Inc.

R. Bruce LaBoon (Chair)
Locke Lord Bissell & Liddell LLP

Herminia Palacio
Harris County Public Health and Environmental Services

Monica Samuels
Attorney

John Walke
Natural Resources Defense Council

SCIENTIFIC ADVISORY PANEL

Ed Avol
University of Southern California

James J. Collins (Chair)
Dow Chemical Company

Michael L. Cunningham
National Institute of Environmental Health Sciences

George Delclos
University of Texas Houston School of Public Health

David H. Garabrant
University of Michigan School of Public Health

Pertti J. (Bert) Hakkinen (Vice Chair)
National Institutes of Health

Harvey Jeffries
University of North Carolina

Bertram Price
Price Associates, Inc.

Nathan Rabinovitch
National Jewish Medical Research Center

Anne Rea
U.S. Environmental Protection Agency

Linda Sheldon
U.S. Environmental Protection Agency

NUATRC STAFF

Craig Beskid
President

Rebecca Jensen Bruhl
Assistant Staff Scientist

Debra A. Kaden
Consulting Staff Scientist

Lata Shirnamé-Moré
Consulting Staff Scientist

Sherry Stevenson
Executive Assistant

Carolyn Wade
Financial Manager

P.O. Box 20286
Houston, Texas 77225-0286
Tel: 713.500.3450
Fax: 713.500.0345
http://www.sph.uth.tmc.edu/mleland/

Printed on 100% Recycled Paper
Authorized by the Clean Air Act Amendments of 1990 (Title III, Section 301/p)