

Data Organization and Operating Procedures

Revision 11, July 2018

- Overview/Purpose and Included Data 4
- 1. Sources, Contents, and Quality Control of Data 8
 - 1.1. Overview of Location Data..... 8
 - 1.1.1. Map Projections and Coordinates 8
 - 1.1.2. Political Identifiers – County, State 8
 - 1.2. Air Quality System (AQS) Data 9
 - 1.2.1 Description of Data Source 9
 - 1.2.2 AQS Locations..... 13
 - 1.2.3 AQS Monitor Deployment Height..... 13
 - 1.2.5 Data completeness rules for averaging 13
 - 1.2.6 Quality Control 14
 - 1.2.7 Data Archiving..... 15
 - 1.3. Monitoring Data from DEOHS Air Monitoring Studies 15
 - 1.3.1 Fixed Site Monitoring 15
 - 1.3.2 Home Outdoor Monitoring 15
 - 1.3.3 Home Indoor Monitoring 16
 - 1.3.4 Personal Monitoring..... 16
 - 1.3.5 In-Vehicle Monitoring 16
 - 1.3.6 Roadway Gradient Snapshot Monitoring..... 16
 - 1.3.7 Other Kinds of Snapshot Monitoring..... 16
 - 1.3.8 Data Description 18
 - 1.3.9 Quality Control of Data 18
 - 1.3.10 Raw Monitoring Data 19
 - 1.4 Monitoring Data from Other Sources 19
 - 1.4.1 Inhalable Particulate Network (IPN) 19
 - 1.4.2 New York City Community Air Survey (NYCCAS) 20
 - 1.5 Participant Address History 20
 - 1.5.1 MESA Air and MESA Neighborhood Participants 20
 - 1.5.2 Technical Note: Address geocoding and geocode flags 21
 - 1.5.3 Cardiovascular Health Study (CHS) Participants 23
 - 1.5.4 Women’s Health Initiative – Observational Study (WHI - OS) Participants 23
 - 1.5.5 SPIROMICS Air Participants 23
 - 1.5.6 ACT – Air Pollution Study 24
 - 1.5.7 The Sister Study and the Two Sister Study Participants..... 24
 - 1.5.6 PSID Study Participants..... 24
 - 1.5.7 BRFSS Study Participants 24
 - 1.5.6 Small-Fee Projects 24
 - 1.6. Participant Exam Dates 24
 - 1.7. Participant Time Location..... 25
 - 1.8. AQS Monitors for ‘Acute’ Analysis (Averaging Period < 1 Year)..... 25
 - 1.8.1 Pre-Adjusted PM_{2.5} Exposure 26
 - 1.9. AQS Monitors Eligible for ‘Nearest Monitor’ Analysis (Averaging Period >= 1 Year) .. 26

1.10. Grid Locations	27
1.11. Geographic Covariates.....	28
1.11.1 Sources of GIS data	29
1.11.2 Creation or Projection of Shapefiles from Raw Data Sources.....	30
1.11.3 Land use data	30
1.11.4 TeleAtlas Road Data.....	33
1.11.5 Distance to Road and Near Road Determination.....	33
1.11.6 Airports and Major Airports	33
1.11.7 Coastlines, Railroads, and Rail Yards.....	34
1.11.8 Ports	34
1.11.9 Distance to Nearest Truck Route and Length of Truck Routes in Buffers.....	34
1.11.10 Population	34
1.11.11 Emissions Data.....	34
1.11.12 Normalized Difference Vegetation Index.....	35
1.11.13 Impervious Surface	35
1.11.14 CALINE Long-Term Average.....	35
1.11.15 Motor Vehicle Emissions Trends.....	36
1.11.16 Residual Oil in New York City.....	37
1.11.17 Distance to City Hall.....	38
1.11.18 Elevation	39
1.11.19 Urban Topography	39
1.11.20 Census Data	40
1.11.21 Distance to Bus Route.....	45
1.11.22 Satellite Data: PM _{2.5} , NO ₂ , SO ₂ , HCHO, CO	46
1.11.23 Method of Covariate (Variable) Calculation	46
1.11.24 Data Quality	46
1.12. Exposure Model Outputs	47
1.12.1 NO _x , NO ₂ , and PM _{2.5} Likelihood Model Predictions.....	50
1.12.2 O ₃ Likelihood Model Predictions	50
1.12.3 SPIROMICS-Specific Likelihood Model Predictions	51
1.12.4 Light Absorption Coefficient (LAC) Predictions	51
1.12.5 Pragmatic PM _{2.5} Predictions	51
1.12.6 National Model PM _{2.5} , PM ₁₀ , and NO ₂ Predictions	52
1.12.7 National Model Historical PM _{2.5} Predictions	52
1.12.8 National Model for As, Cr, Cu, EC, Ni, NO ₃ ⁻ , OC, S, Si, SO ₂ , SO ₄ ²⁻ and V, Predictions.....	52
1.12.9 Coarse PM Land Use Regression Predictions	52
1.12.10 Alternative National NO ₂	53
1.12.11 Individual-level Exposures to Ambient PM _{2.5}	53
1.13. Meteorological Data.....	53
1.13.1 B-Spline Variables for Temperature, Humidity, and Calendar Time.....	54
2. Exposure Assessment Core Data Request System.....	55
2.1 Placement of a Request.....	55
2.2 Maps of Participant-Identifying Information.....	55
2.3 Multiple Users of Requested Data, or Accessing an Existing Request	55
2.4 Fulfillment of a Request.....	56

2.4.1 Data Request Fulfillment: Work Flow Details for Internal Users	56
Appendices.....	58
A. Statistical Analysis Plan (SAP).....	58
B. List of Acronyms and Abbreviations	59
C. Referenced Documents and Code Locations	60
D. Suggested Citations.....	63
E. Known Data Quality Issues	64
F. Quick Reference for Averaged Exposure Variable Names	64
H. MESA Air Modeling Regions.....	67
I. Ozone Modeling Regions	68
J. Performance Statistics for Exposure Models	69

Overview/Purpose and Included Data

The purpose of this document is to describe the collection, processing, and management of exposure data for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) and associated environmental epidemiology studies. The primary aims are: 1) to record the treatment of all of the Kaufman/Vedal Lab's environmental monitoring data prior to its upload to the Exposure Assessment Core's (EAC) database; 2) to describe the content and processing of third-party (e.g., Air Quality Systems, NYCCAS) data; and 3) to describe the data request process. Health data are collected and stored at study-specific coordinating centers, and documentation for these data handling procedures can be found in a separate Standard Operating Procedure (SOP) document. **The EAC does not fill data requests for health data.** We are currently distributing data from Revision 11 (Rev 11) of the EAC Database. The data available in Rev 11 include:

Location data

Location identifiers for monitoring sites and participant homes include: county according to the 2000 Census, state, metropolitan statistical area as of the year 2000, census tract, block group, and block according to years 2000 and 2010 censuses, projected coordinates [x,y] referenced to the national US Conical Lambert projection in meters, latitude and longitude in decimal degrees.

- **Residential Locations**
 - Addresses for participants in the Women's Health Institute Observational Study and Clinical Trial (WHI-OS and CT) (not participant-linked)
 - Addresses for participants in the Cardiovascular Health Study (CHS) (not time resolved)
 - Addresses for NOMAS (not time resolved)
 - "Current" and secondary (both at baseline), childhood, and longest-lived addresses for participants in the Sister and Two Sister Studies
 - Locations of intersections near BREATH participant residences
 - Time-resolved address histories for MESA Air participants
 - Time-resolved address histories for SPIROMICS Air participants
 - Time-resolved address histories for ACT-AP participants
 - Time-resolved address histories for GEMS participants
- **Air Monitoring Locations and Weather Stations**
 - New York City Community Air Survey (NYCCAS) monitoring sites
 - Inhalable Particle Network (IPN) sites
 - AQS (state and local) and IMPROVE (federal) monitoring network sites
 - MESA Air monitoring sites
 - Yesler Terrace monitoring sites
 - Diesel Exhaust Exposure in the Duwamish Study (DEEDS) monitoring sites
 - CCAR Project 1 mobile monitoring routes in Baltimore, MD and passive badge sites in Atlanta, GA; Baltimore, MD; Winston-Salem, NC; and Los Angeles, CA
 - CCAR Project 5 monitoring sites
 - SPIROMICS Air monitoring sites
 - LAX flightpath monitoring locations
 - NOAA meteorological monitoring stations throughout the continental US
 - PM Center Panel Study monitoring locations
 - ACT-AP monitoring locations
- **Other Locations**

- Locations for grids of points in each of the MESA Air study areas to facilitate the production of maps illustrating area-wide surfaces (regional grids) and small-scale variations (fine-scale grids) of model predictions
- Fine-scale grid points in Seattle covering the area relevant to DEEDS
- Fine-scale grid and on-road points covering Oakland
- Locations for a large-scale grid of points nationwide, also to facilitate map production
- Grid point locations used in the CCAR Collaboration Project
- Year 2010 Census Tract centroids

Address History

- All available time-resolved address histories for all MESA and MESA Air participants (with the exception of 11 MESA participants who explicitly declined to participate in MESA Air) from 1980 through the end of Exam 5 in the MESA Air extended region. “At exam” addresses are identified for the residences of MESA Air participants on the dates of clinical exams, Computed Axial Tomography (CT) scans, magnetic resonance imaging (MRI) scans, ultrasound scans, and spirometry
- All available time-resolved address histories for SPIROMICS Air participants
- All available time-resolved address histories for ACT participants

Federal, State, and Local Agency Monitoring data

AQS, Speciation Trends Network (STN), and IMPROVE monitoring data are available throughout the lower 48 United States and Hawai'i. The years for which data are available for specific pollutants are variable, as regulatory needs have changed over time. The last date for which data are available varies by location and depends on the diligence with which agencies report updates. In general, AQS and STN data are considered up-to-date through January 2018 for most locations. IMPROVE data are available from March 1988 through June 2017 for PM_{2.5} and its species. Specific availability by pollutant is as follows:

- AQS monitoring data for NO₂, NO_x, SO₂, and O₃ beginning in 1979, and CO beginning in 1980
- AQS monitoring data for PM₁₀ beginning in 1983 and IMPROVE monitoring beginning in 1999
- AQS monitoring data for PM_{2.5} beginning in 1999, IMPROVE monitoring data beginning in 1988
- AQS monitoring data for sulfate (SO₄²⁻) and nitrate (NO₃⁻) beginning in 2000
- STN monitoring data for elemental carbon and organic carbon (EC/OC) beginning February 2000, IMPROVE monitoring data for EC/OC beginning March 1988
- STN and IMPROVE monitoring data for selected elemental species (aluminum, arsenic, bromine, cadmium, calcium, chromium, cobalt, copper, iron, potassium, manganese, sodium, silicon, selenium, sulfur, nickel, vanadium, and zinc) beginning in 1988
- IPN monitoring data for PM_{2.5}, PM₁₅, PM_{15-2.5}, and PM₃₀ between 1979 and 1982

Air quality data collected through MESA Air and related studies

- PM_{2.5}, PM_{2.5} species, NO_x, NO₂ and O₃ data collected as part of MESA Air and the Health Effects Institute (HEI) / National Particle Components Toxicity (NPACT) study
- PM_{2.5}, NO_x, NO₂ and O₃ monitoring data collected as part of the SPIROMICS Air Study
- PM_{2.5} and NO_x monitoring data collected as part of DEEDS
- NO_x monitoring data collected in Yesler Terrace
- NO_x monitoring data collected under the LAX approach flightpath
- PM_{2.5}, PM₁₀, PM_{10-2.5}, associated elemental species, and endotoxin monitoring data collected as a part of the MESA Coarse PM Study

- NO_x and O₃ Ogawa badge data collected at ‘fuzzy points’ in Atlanta, Baltimore, LA, Winston-Salem, and St. Paul by CCAR Project 1
- NO_x, O₃, and VOC badge data collected at residential locations and in vehicles in Winston-Salem and LA as part of CCAR Project 5

Exposure model outputs – MESA and SPIROMICS regions

- PM_{2.5}, NO_x, NO₂, and O₃ predictions for MESA Air participants’ addresses throughout the MESA Air areas for the years 1999 – 2013 at 2-week resolution
- PM_{2.5}, NO_x, NO₂, and O₃ predictions for SPIROMICS Air participants’ addresses throughout the SPIROMICS Air areas for the years 1999 – 2016 at 2-week resolution
- LAC predictions for MESA Air participants’ addresses in the MESA Air areas representing long-term exposure (predictions averaged over 2006 – 2008) and at two-week resolution (2005 – 2009)
- Land use regression predictions for PM_{10-2.5} and silicon, copper, zinc, and phosphorus in the coarse fraction for MESA Air participants’ addresses in the MESA Coarse Study monitoring areas (Chicago, Winston-Salem, and St. Paul)

Exposure model outputs – National Models (Continental US, Annual Averages)

- PM_{2.5} predictions for each year from 1999 – 2015 at MESA Air, SPIROMICS Air, ACT-AP, WHI-OS, Sisters, Two Sisters and CHS participants’ residential locations
- PM_{2.5} predictions for each year from 1980 – 2010 at MESA Air, WHI-OS, Sisters, and Two Sisters participants’ residential locations from an annual spatio-temporal model (“historical PM2.5 model”)
- EC, OC, arsenic, silicon, sulfur, chromium, copper, SO₄²⁻, SO₂, nickel, nitrate, and vanadium predictions for 2009 at MESA Air and WHI-OS participants’ residential locations
- PM₁₀ predictions for each year from 1990 – 2014 for MESA Air, SPIROMICS Air, ACT-AP, WHI-OS, Sisters, Two Sister, and CHS participants’ residential locations
- NO₂ predictions for each year from 1990 – 2014 for MESA Air, SPIROMICS Air, ACT-AP, WHI-OS, Sisters, and Two Sister participants
- NO₂ predictions for the year 2006 at census block centroids from the model developed by Novotny et. al.¹ These predictions are available for all locations in the continental US.

Exposure estimates based on AQS data

- Time series data of PM_{2.5} for each MESA Air and SPIROMICS Air area designated for acute analyses
- Specified “long-running” monitors for long-term ‘nearest monitor’ analyses for PM_{2.5}
- Cascading nearest monitor assignments (described in more detail later), for more unusual analyses

Meteorological data

- Visibility, temperature, wind speed, wind direction, humidity, sea level pressure, station-level pressure and ceiling height data at the daily time scale at meteorological stations throughout the nation from the years 1979 – 2017

Geographically-linked data

¹ Novotny EV, MJ Bechle, DB Millet, and JD Marshall. 2011. "National satellite-based land-use regression: NO₂ in the United States". *Environmental Science & Technology*. 45 (10): 4407-14.

- Geographic covariate values
 - USGS aerial photography-based land use in buffers, year 2006 satellite-based land use in buffers, impervious surface, normalized difference vegetative index (NDVI), distance to features, emissions in buffers, sum of road lengths in a buffer, sum of truck route lengths in a buffer, counts of intersections in buffers, population in buffers, and elevation nationally
 - Distance to city hall for some locations in the MESA Air regions
 - Distance to nearest residual oil source, sums of oil emissions in buffers, nearest bus route, and sum of bus routes in buffers for the New York City area
 - Urban topography for locations in New York City and Chicago
- Selected census variables for all block groups, blocks, and tracts in the lower 48 United States for both Census 2000 and Census 2010
- Average column NO₂ for the year 2006 at census block centroid as measured by satellite image processing
- Average column NO₂ for a Moore neighborhood of pixels near each location in the continental US for the years 2005 – 2007
- CALINE3QHCR dispersion model averages of the years 2004 – 2006 for monitoring locations and MESA Air participant addresses
- Vehicle emissions model outputs (MOVES data) by month for the years 1990 and 1999 – 2012.

Additional data will be added in future releases of the database. Planned updates are slated to include:

- Additional updates of existing models (e.g., national PM_{2.5}, national NO₂, likelihood-based PM_{2.5}), are expected, both to incorporate additional AQS data not currently available and to reflect improvements to modeling methodology.
- National scale spatio-temporal models
- MESA Air address updates
- K-means cluster predictions
- Updated bus and truck routes

Database revisions are planned to occur approximately once per year. If the available data does not satisfy the needs of a particular analysis, data users are encouraged to contact the EAC as soon as possible to negotiate new data's inclusion in an upcoming revision.

1. Sources, Contents, and Quality Control of Data

1.1. Overview of Location Data

Location data are central to the design of the EAC MySQL database, as all exposure data compiled for the relevant studies are associated with a location and all exposure predictions will be made at locations. Locations of interest include air quality monitoring sites, meteorological monitoring sites, and participants' homes. Other important locations include the intersection points of grids developed at various scales to facilitate preparation of visual aids like concentration maps. "Location data" refers to the data necessary to specify a position and its relevant characteristics. For example, a participant's home is located at a certain latitude and longitude. The latitude and longitude can then be used to determine geographic variables, such as population density at that location according to a given Census year. The coordinates of each location were determined as specified in each relevant subsection.

1.1.1. Map Projections and Coordinates

Geographic coordinates and projected coordinates are stored in the EAC database for all locations. Latitude and longitude are stored as decimal degrees as referenced to the North American Datum geographic coordinate system. Two different sets of projected coordinates will be available. The projected coordinates (x,y) represent the number of meters that the location lies from the origin of the projection. One set of coordinates will be referenced to a specific State Plane Zone. Another set of coordinates will be referenced to the US Conical Lambert projection (lambert x, lambert y). Distance calculations using State Plane coordinates are expected to be more accurate, but these coordinates cannot be used to calculate distances across State Plane Zones. Distances calculated between two points within a state plane could vary up to almost 1% depending on the coordinate system used (up to approximately 2 kilometers absolute distance).

Locations were assigned latitudes and longitudes ("geocoded") in one of three ways: created in ArcGIS from street addresses, recorded directly from digital global positioning system (GPS) units, or acquired directly from a third-party data provider. The geographic coordinates were then projected into the State Plane Coordinate System (SPCS). The latitude and longitude and the projected coordinates of the SPCS for participants' residences are available to users with appropriate permission. Coordinates that uniquely identify a participant's home are considered identifying information by the University of Washington Human Subjects Review Board. Access to participant home locations may be granted by the Principal Investigator for data users that demonstrate a need for that information, complete any relevant human subjects training, and sign the Data and Materials Distribution Agreement (DMDA).

1.1.2. Political Identifiers – County, State

Political identifiers of location that are available include state, county, and block group. As a default, counties and block groups are assigned by overlaying the locations' latitudes and longitudes onto the TeleAtlas Dynamap 2000 v. 16.1 county shapefile. In the future, if necessary, historical counties will be determined by overlaying the locations onto historical census shapefiles in later versions of the database. State boundaries of the lower 48 states have not changed during the lifetimes of any study participants, so state is assigned only by the TeleAtlas 2000 boundaries.

The term 'city' is vague in the context of our health studies and is to be avoided, although we are able to provide the year 2000 Metropolitan Statistical Area (MSA) for locations. For convenience, a standard set

of monitors were selected in each MESA Air study area for exposure modeling. These are indicated on the maps on the internal website.

1.2. Air Quality System (AQS) Data

1.2.1 Description of Data Source

Air quality data, *i.e.* pollutant concentrations, are collected nationally by networks of federally- and locally-funded agencies, called the Air Quality System (AQS). The AQS data from most agency monitoring locations were obtained from the Environmental Protection Agency's (EPA) website at https://aqs.epa.gov/aqsweb/airdata/download_files.html. Monitoring data from national parks and some rural locations is collected by a specific sub-network of the AQS, called the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. Data collected by the IMPROVE network were obtained from the Visibility Information Exchange Web System (VIEWS), at <http://views.cira.colostate.edu/iwdw/QueryWizard/Default.aspx>. Users should be aware that non-IMPROVE monitoring locations are generally selected to assess population exposures, and that IMPROVE monitoring locations are generally selected to assess pollution impacts in sparsely populated environments. Therefore, users may determine that IMPROVE monitors are not representative of pollutant levels to which "nearby" subjects are exposed if those subjects live in reasonably populous areas.

The AQS network is comprised of a heterogeneous mixture of state, county, and local agencies throughout the United States. As funding availability and monitoring aims evolve, monitors come online or are discontinued, and monitoring schedules are altered. The monitors near a given location that collected PM_{2.5} data from 2000 to 2002 may not have collected PM_{2.5} data from 2003 to 2005. Furthermore, an individual monitor that collected PM_{2.5} data every day in 2002 may have collected data every third day in 2003. Another monitor that collects PM_{2.5} data every day during the winter may only collect data every third day in the summer. These features of the data primarily affect analysts creating exposure models. In general, health analysts conducting analyses based on AQS monitoring data will receive data from monitors as described in Sections 2.6 and 2.7.

PM_{2.5} data is collected by air quality agencies via several methods, and is consistently reported in units of $\mu\text{g}/\text{m}^3$. Federal Reference Methods (FRM) are used to determine compliance with National Ambient Air Quality Standards (NAAQS). These data are considered to be of the most reliable quality, so only PM_{2.5} data collected by FRM monitors will be available in the EAC database. FRM methods are similar to one another, and involve collecting PM_{2.5} on a filter for 24 hours at a time. Most commonly, agencies measure PM_{2.5} one out of three days, with occasional data loss. The IMPROVE network collects PM data using FRM-type methods. A comparison of IMPROVE data to AQS FRM data collected concurrently at the same locations indicated that these methods were comparable enough to be considered equivalent. Occasionally, EPA audits of AQS data have found that measurement conditions did not meet the specifications required to validate those measurements for regulatory purposes. Many of these measurements were re-classified as "raw data" by the EPA. Although there is some evidence that these data are noisier than validated measurements, they are collected using reference methods and EPA staff believe that they are "potentially of quality." We have included these measurements with a flag where no validated measurement was available.

A number of users have asked if we will include data from Federal Equivalent Methods (FEMs). These include tapered element oscillating monitors (TEOMs) and beta attenuation monitors (BAMs), which are both one-hour integrated methods. When used for compliance purposes, these monitors are used

alongside and corrected to filter methods. Uncorrected methods are not as well-correlated with the filter methods, show seasonal bias, and are relatively sparse in locations where we do not already have filter-based monitors. Our judgment has been that these monitors would add noise to our models rather than improve their spatial resolution.

We note that some agencies may operate two monitors concurrently at the same location for quality control purposes. The EAC database is intended to house no more than one concentration for a single point in time and space. Therefore, AQS monitoring methodologies were ranked in terms of reliability based on best professional judgment. A detailed list of the rankings can be found in Tables 1 through 5. Data for the more reliable monitor was retained for time periods during which two monitors' data were available from exactly the same location. If two co-located monitors used the same method, the data from the monitor with a more complete time series were retained. The alternative monitor's data were inserted for time periods during which only that monitor's data were available. Thus, a given location could have a single time series of monitoring data sourced from more than one monitor.

Table 1. Ranking of PM_{2.5} collection methods

Rank	Method Code	Method Description
1	145	R & P Model 2025 PM _{2.5} Sequential Air Sampler w/VSCC
2	118	R & P Model 2025 PM _{2.5} Sequential w/WINS
3	144	R & P Model 2000 PM _{2.5} Audit Sampler w/VSCC
4	143	R & P Model 2000 PM _{2.5} Air Sampler w/VSCC
5	117	R & P Model 2000 PM _{2.5} Sampler w/WINS
6	116	BGI Model PQ200 PM _{2.5} Sampler w/WINS
7	155	Thermo Electron Model RAAS2.5-300 Sequential w/VSCC
8	142	BGI Models PQ200-VSCC or PQ200A-VSCC
9	129	R & P Model 2000 PM _{2.5} Audit w/WINS
10	153	Thermo Electron Model RAAS2.5-100 w/VSCC
11	154	Thermo Electron Model RAAS2.5-200 Audit w/VSCC
12	135	URG-MASS100 Single PM _{2.5} Sampler
13	136	URG-MASS300 Sequential PM _{2.5} Sampler
14	128	Andersen RAAS2.5-2000PM _{2.5} Aud w/WINS
15	119	Andersen RAAS2.5-100 PM _{2.5} SAM w/WINS
16	120	Andersen RAAS2.5-300 PM _{2.5} SEQ w/WINS
17	123	Thermo Env Model 605 CAPS [Computer Assisted Particle Sampler]
18	177	Thermo Scientific Partisole 2000-D Dichot.
19	179	Thermo Scientific Dichot. Partisole-Plus Model 2025-D Seq

Table 2. Ranking of PM₁₀ collection methods

Rank	Method Code	Method Description
1	127	R - P Co Partisol Model 2025
2	126	R - P Co Partisol Model 2000
3	098	R&P Model 2000 Partisol
4	079	INSTRUMENTAL-R&P SA246B-INLET
5	125	BGI Inc. Model PQ200 PM10
6	124	BGI Inc. Model PQ100 PM10
7	900	BGI Inc. frmOMNI at 5 lpm
8	063	HI-VOL SA/GMW-1200
9	064	HI-VOL-SA/GMW-321-B
10	065	HI-VOL-SA/GMW-321-C
11	162	Hi Vol SSI Ecotech Model 3000
12	130	Andersen RAAS10-100 Single channel
13	131	Andersen RAAS10-200 S-Channel
14	132	Andersen RAAS10-300 M-channel
15	141	Tisch Environ Model-6070 PM10 Hi-Vol
16	062	HI-VOL-WEDDING-INLET
17	773	LO-VOL-DICHOT-INTERIM
18	073	LO-VOL-DICHOTOMOUS-SA246B-INLT
19	071	OREGON-DEQ-MED-VOL

Table 3. Ranking of O₃ collection methods

Rank	Method Code	Method Description
1	047	Ultraviolet Absorption – Thermo Electron 49
2	103	Open Path Ozone Analyser - OPSI Model AR 500 Ozone Analyzer
3	056	Ultraviolet Absorption - DASIBI 1008-AH
4	019	Ultraviolet Absorption - DASIBI 1003-AH--PC—RS
5	053	Ultraviolet Absorption - Monitor Labs 8810
6	078	Ultraviolet Absorption - Environics Series 300
7	087	Ultraviolet Absorption - Model 400 Ozone Analyzer
8	112	Ultraviolet Absorption – Horiba APOA-360
9	105	UV Photometric - Environment SA Model Q341M
10	020	Chemiluminescence - Beckman 950A
11	--	Other

Table 4. Ranking of SO₂ collection methods

Rank	Method Code	Method Description
1	560	Pulsed Fluorescence - Thermo Electron 43C-TLE
2	060	Pulsed Fluorescence - Thermo Electron 43A, 43B, 43C
3	009	Pulsed Fluorescence - Thermo Electron 43
4	--	Other

Table 5. Ranking of EC/OC collection methods

Rank	Method Code	Method Description
<i>Chemical Speciation Network (CSN) Temperature Protocol, Total Optical Transmittance – AQS parm 88305 and 88307</i>		
1	847	R & P Model 2025 PM2.5 Sequential Quartz VSCC
2	843	R & P Model 2025 PM2.5 Sequential Quartz
3	853	R&P MDL2300 PM2.5 Seq Spec Quartz
4	873	URG MASS450 Quartz VSCC
5	833	URG MASS450 Quartz WINS
6	823	Andersen RAAS Quartz
7	813	Met One SASS Quartz
<i>IMPROVE Temperature Protocol, Total Optical Reflectance - AQS parm 88320 and 88321</i>		
1	859	R & P Model 2025 PM2.5 Sequential Quartz VSCC
2	857	R & P Model 2025 PM2.5 Sequential Quartz
3	855	R&P MDL2300 PM2.5 Seq Spec Quartz
4	877	URG MASS450 Quartz VSCC
5	829	URG 3000N w/Pall Quartz filter and Cyclone Inlet
6	838	URG 3000N w/Pall Quartz filter and Cyclone Inlet
7	837	URG MASS450 Quartz WINS
8	815	Met One SASS Quartz
9	825	Andersen RAAS Quartz
10	809	IMPROVE Module C with Cyclone Inlet-Quartz Filter
11	805	IMPROVE
<i>IMPROVE Temperature Protocol, Total Optical Transmittance</i>		
--	838	URG 3000N w/Pall Quartz filter and Cyclone Inlet, Unadjusted – AQS parm 88355 and 88357
--	840	URG 3000N w/Pall Quartz filter and Cyclone Inlet, Unadjusted – AQS parm 88355 and 88357
<i>Revised CSN Temperature Protocol, Total Optical Reflectance, Unadjusted</i>		
--	831	URG 3000N w/Pall Quartz filter and Cyclone Inlet – AQS parm 88370 and 88380

Elemental and organic carbon (EC/OC) are not listed as criteria air pollutants, and as such have no federal reference methods. Similar equipment is used for the collection of samples for EC/OC analysis as is used for the collection of PM_{2.5} for gravimetric analysis, so similar rankings were used to prioritize EC/OC methods. One issue that is important to consider when requesting EC/OC data is that these pollutants are operationally defined. Through 2007, EC/OC was primarily measured using the Chemical Speciation Network (CSN) temperature protocol. In 2007, many locations switched to using the IMPROVE temperature protocol. Comparisons between HEI/NPACT EC/OC measurements and monitoring network measurements by different protocols are presented in the XRF and EC/OC supplement to the QA report.

AQS monitors collect hourly NO_x and NO₂ data by chemiluminescence, generally reported in units of ppm (and stored in the EAC database in units of ppb). All chemiluminescent methods are considered equivalent, so data for the monitor with the more complete time series were retained for days with two monitors' data available. If a duplicate monitor existed, data from that monitor were inserted into the time series for time periods during which only that monitor's data were available.

Speciated data for the same species are rarely assessed using more than one method, so data for the monitor with the more complete time series were retained for days with two monitors' data available.

If a duplicate monitor existed, data from that monitor were inserted into the time series for time periods during which only that monitor’s data was available.

1.2.2 AQS Locations

AQS locations were downloaded from AQS Data Mart as latitude and longitude. Flat files were converted to shapefiles in ArcGIS and re-projected into the appropriate State Plane Coordinate System before geographic variables were calculated.

AQS monitoring locations are identifiable by unique identifiers assigned by the EPA. These identifiers have three parts: a two digit state code, a three digit county code, and a four digit site code. Individual monitors at monitoring sites also have a five digit pollutant code, and a two-digit ‘parameter occurrence code’ (POC). The two digit state code and three digit county code are defined under the Information Technology Management Reform Act (Public Law 104-106) and the National Institute of Standards and Technology (NIST) for Federal computer systems. The four digit site code site code defines unique AQS monitor locations within certain counties, and the five digit parameter code defines the specific parameter being collected by a designated monitor. For example, a PM_{2.5} monitor in Los Angeles County, California at site 0001 could be numbered 0603700018810101. As described above, multiple monitors for the same pollutant are occasionally operated concurrently at the same site. The POC identifies unique monitors in this case, with lower numbers indicating more established monitors. Maps of the AQS monitor locations for which we retain data are available online.

1.2.3 AQS Monitor Deployment Height

AQS monitors are generally deployed at ground level or on the roof of a low building, but this is not universally the case. This may be of particular importance in New York, so the floor at which monitors were deployed was obtained from the air quality agency. Monitors deployed above the ground floor in the New York City metropolitan area are flagged with this information.

1.2.4 Data completeness rules for averaging

Data completeness rules are based on a minimum number of data points and a maximum gap between consecutive measurements, which are derived from the length of the averaging period and the typical monitoring frequency of the pollutant in question.

Table 6. Data completeness rules for averaging

Averaging Period	Maximum Gap	Minimum Data Points		
		1-in-6 Days (PM ₁₀ , Elements, Carbon Species)	1-in-3 Days (PM _{2.5})	Daily (Gases)
Two Weeks	12	2	4	10
One Month	12	4	7	21
Six Months	15	21	41	122
Nine Months	30	31	61	127
One Year	45	41	82	244

Note that ozone is measured seasonally at approximately half of locations. For many applications, it may be more appropriate to request the April – October average rather than the annual average.

These standards were established for convenience and are not expected to cover the needs of every data request. All available AQS data are housed in the EAC database, not just those meeting these

criteria. Users are welcome to design their own inclusion criteria; space is provided in the electronic data request form to outline criteria specific to each user’s needs. The monitor identification numbers are indicated on the study area maps at <http://www.uwchsc.org/MESAAP/Data.aspx>, so that users can list individual monitor numbers for which they would like data. These maps indicate the monitors that are included in the standard set, as well as additional monitors in the region. If no inclusion criteria are described, the standards above will be used as the default.

1.2.5 Quality Control

Air pollution monitoring is primarily conducted to comply with federal regulations, though individual air quality agencies may have other secondary objectives. Siting priorities, funding, local culture, and chance all have the potential to impact the quality of data from individual monitors. Therefore, certain criteria have been set for the inclusion of AQS data. After data are downloaded, the “Basic Level” of QC consists of the following steps:

- Limits of detection (LOD) are determined from the documentation available on the AQS website; some may depend on the collection or analysis method. A range is provided for these species.

<u>Parameter</u>	<u>LOD</u>
NO _x /NO ₂	1 ppb
SO ₂	2-5 ppb
O ₃	1.5-5 ppb
AQS PM _{2.5}	2 µg/m ³
IMPROVE PM _{2.5}	0.0562-2 µg/m ³
speciated elements from the STN	0.0001-0.0347 µg/m ³
speciated elements from the IMPROVE network	0.00001-2.23 µg/m ³
EC/OC	0.002-0.245 µg/m ³
IMPROVE EC and OC	Not listed

- Measurements below the limit of detection (LOD) are replaced with a value equal to half the LOD. These values, and any daily averages associated with them, are assigned an ‘LOD’ flag.
- Very high outliers (≥ 1000 ppb for gases or ≥ 1000 µg/m³ for PM_{2.5}) are investigated. Occasionally, the wrong units appear to have been assigned to the raw data (i.e. ppm instead of ppb) by the local agency. A more reasonable unit may be assigned by the data manager, but since the air quality agency typically is not asked to confirm this error, the measurement is associated with a ‘unit’ flag.
- Daily averages are calculated for those pollutants for which AQS captures hourly data, provided that the hourly measurements (where they exist) meet the following criteria. The same applies to daily meteorology data that is computed from hourly data:
 - At least 18 hours for the day are available
 - At least 4 hours between 4:00 am and 9:00 am are available
 - At least 4 hours between 1:00 pm and 6:00 pm are available
- For ozone, the 10 am – 6 pm average will also be available. The inclusion criterion for these measurements is that 6 of the intended 8 hours must be available.
- For ozone, the maximum 8-hour average will also be available. This measurement can be calculated only if that day’s data meets the daily average criteria and if at least 6 hours’ data are available for each 8-hour averaging window.
- A small amount of additional data are excluded based on personal communication with local air monitoring agencies, which is documented in the Monitor Issue Log (see Appendix C). If extremely suspicious data were brought to the attention of a local agency, but no response was

received or the agency confirmed its validity, then those data are retained in the database but may be assigned an 'Agency' flag to alert the analyst.

1.2.6 Data Archiving

Older versions of the AQS data already used for publications and manuscripts have sometimes been archived on DVD, including a pre-database version of AQS data for the years 1980-1999. Historical versions of data within the EAC database are archived in compressed files that can be accessed to fill old data requests or retrieve older versions of covariates and predictions.

1.3. Monitoring Data from DEOHS Air Monitoring Studies

Below are overviews of the types of air monitoring that our working group has conducted. There are study-specific QA/QC documents and other materials that are available and can be accessed as needed. Some of these documents are listed in Appendix C. The EAC database contains the dates that sampling started and ended for each individual measurement, as well as the middle day of the intended two-week sampling period (typically a Wednesday). Samples with concentrations below the LOD are flagged, and the LODs associated with these samples are provided. Additional flags for minor circumstances were associated with a small number of measurements as described in the final Quality Assurance / Quality Control Report (see Appendix C).

In addition to the geographic covariates that are available for all locations, DEOHS monitoring locations are also associated with a building floor at which the data were collected. This is primarily of interest in cities such as New York where many living units are located well above street level. For the Coarse monitoring sites and where the living floor is not indicated, data were collected at the first (ground) or second floor.

1.3.1 Fixed Site Monitoring

Fixed sites are similar to AQS sites in that they are designed to collect a continuous series of measurements over the monitoring period. Fixed sites were operated continuously in all MESA Air regions (2-5 per region) from approximately July 2005 until July 2009. The exact start and end dates vary from site to site. The methods used for these monitoring sites collected two-week integrated measurements, using either Teflon filters for PM or Ogawas for gases. Each PM sampler was intended to run on a 50% duty cycle for consecutive two-week intervals. These sites were chosen to represent MESA Air participant exposures, near road exposures, or were co-located with an AQS monitor in the area. PM_{2.5} (for mass, LAC, and elements), NO₂, NO_x, SO₂, and sometimes O₃ were measured at these sites. The HEI/NPACT study collected PM_{2.5} for EC/OC analysis at these sites during some time periods.

Fixed sites were set up in each MESA city and Seattle during 2017. These sites were primarily co-located with AQS sites, and remote air data (RAD) monitors were used to report 5-minute resolution for PM_{2.5}, O₃, NO₂, NO_x, and CO. These data are not yet available.

1.3.2 Home Outdoor Monitoring

Home outdoor monitoring refers to air monitoring conducted outside residential locations, primarily of participants involved in one of the health studies. In theory, repeated monitoring periods (or rounds) at the same home were intended to occur in distinct seasons: summer, winter, or "transitional" (spring or fall). Home outdoor monitoring was conducted in MESA Air, SPIROMICS Air, ACT-AP, CCAR Project 5 (Ogawas and VOCs), and the PM Center Panel Study.

1.3.3 Home Indoor Monitoring

Home indoor monitoring occurred at a subset of outdoor home monitoring locations. Similar sampling set-ups were operated concurrently inside and outside the participant's home. PM_{2.5} (for mass, LAC, and elements), NO₂, NO_x, SO₂, and O₃ were measured at these sites in MESA Air. SPIROMICS Air additionally included nicotine. CCAR Project 5 included VOCs but not PM_{2.5}.

1.3.4 Personal Monitoring

Personal monitoring occurred at a subset of indoor home monitoring locations. Participants carried samplers for NO_x, NO₂, SO₂, and PM_{2.5} (for mass, LAC, and elements) for two weeks in MESA Air. In SPIROMICS Air, participants carried NO_x, NO₂, SO₂, and O₃ samplers for two weeks. CCAR Project 5 participants carried these Ogawas as well as VOC samplers.

1.3.5 In-Vehicle Monitoring

In-vehicle monitoring occurred in the personal vehicles of subjects that participated in CCAR Project 5. These samplers were sealed in canisters when the vehicle was not in use.

1.3.6 Roadway Gradient Snapshot Monitoring

Snapshot monitoring included ~100-150 individual locations monitored simultaneously for a two-week period using Ogawas on telephone poles. The aim of this type of campaign is to learn more about pollutant gradient near roads. Gradient-type snapshot monitoring was conducted by MESA Air, SPIROMICS Air, and ACT-AP (data not yet available).

1.3.7 Other Kinds of Snapshot Monitoring

MESA Air Coarse PM: The Coarse PM Study collected snapshot data in the winter and summer at MESA Air participant homes in Chicago, Winston-Salem, and St. Paul. This study measured PM_{2.5}, PM₁₀, and endotoxin, and was conducted primarily at participant homes. NO_x/NO₂/SO₂ Ogawa data from these home locations is available for most sampling rounds.

CCAR "Fuzzy Points": The Center for Clean Air Research Project 1 conducted mobile monitoring campaigns in Atlanta, Winston-Salem, St. Paul, LA, and Baltimore. "Fuzzy points" were intersections that the monitoring vehicle passed through several times from different directions. Ogawa and VOC badges were hung near these intersections over the 2-week monitoring periods. Each city was sampled in a heating (winter) and non-heating (summer) season.

LAX "Flightpath": Two airplanes land at LAX every minute between 6:30 am until midnight, with all airplanes following the same flightpath over Inglewood. During Fall 2014, 24 NO_x/NO₂ Ogawas were distributed in roughly a grid pattern in this community.

Yesler Terrace: The Seattle Housing Authority was planning a re-development of the Yesler Terrace neighborhood. NO_x/NO₂ Ogawas were deployed at 28 locations in February 2010 and 86 locations in March 2010.

Diesel Exhaust Exposure in the Duwamish Study (DEEDS): DEEDS measured PM_{2.5}, light absorbing carbon (indicated by filters' light absorption coefficients), NO_x, NO₂, and SO₂ at snapshot sites in Seattle, WA in Summer and Winter 2012. These data are described in Jill Schulte's DEOHS master's thesis. All monitoring results are based on a two-week integrated sampling design.

Table 7. Pollutant availability by study. Rounds per home refers to home outdoor monitoring. Rounds per pole refers to snapshot monitoring. The air monitoring methods were either two-week integrated (TWI) using HPEMs and passive badges or remote air data (RAD).

Study	Geographic Area	Time Period	Rounds per Home	Rounds per Pole	PM _{2.5}	NO _x	NO ₂	O ₃	Other
Yesler Terrace	Seattle	2010		2		TWI	TWI		
DEEDS	Seattle	2012		2	TWI	TWI	TWI		1-NP
MESA Air	MESA Air cities ²	2005 – 2009	1-3	3	TWI	TWI	TWI	TWI	XRF
MESA Air – Coarse	Winston-Salem, St. Paul, Chicago	2009	1-2		TWI	TWI	TWI	TWI	Endotoxin, PM ₁₀
CCAR Project 1	LA, Baltimore, Winston-Salem, St. Paul, Atlanta	2011-2013		1-2		TWI	TWI	TWI	VOCs
CCAR Project 5	LA, Winston-Salem	2013-2014	1-2			TWI	TWI	TWI	VOCs
CCAR Flightpath	LA	2014		1		TWI	TWI		
SPIROMICS Air	SPIROMICS cities ³	2014-2016	1-2	1-2	TWI	TWI	TWI	TWI	
ACT – Air Pollution ¹	Seattle	2017-2019	2 (Target)	3 (Target)	RAD	RAD	RAD	RAD	CO (RAD)
MESA air R56	Baltimore	2017-2018	1-2		RAD	RAD	RAD	RAD	CO (RAD)
PM Center Panel Study	Seattle	1999-2002	1-2		TWI/Neph	TWI	TWI		

¹ Monitoring data not finalized as of Rev 11 publishing date

² MESA Air Cities: Winston-Salem, NC; New York City, NY; Baltimore, MD; Chicago, IL; St. Paul, MN; Los Angeles, CA

³ SPIROMICS Air Cities: Winston-Salem, NC; New York City, NY; Baltimore, MD; Los Angeles, CA; Ann Arbor, MI; Salt Lake City, UT; San Francisco, CA

1.3.8 Data Description

Sampling designs prior to 2017 were generally intended to yield consecutive two-week integrated air pollutant measurements. The start and end dates of each sampling event are available, and the 'intended middle day' is available for MESA Air measurements. Pump failure, building closures, and participant availability often resulted in measurements that were shorter than two weeks or that were shifted from the standard Wednesday-to-Wednesday schedule by one or two days. A few Coarse PM Study snapshots were scheduled for a two week period that began and ended on a day other than Wednesday. For all MESA Air samples, the 'intended middle day' is the Wednesday that best aligns that sample with the standard Wednesday-to-Wednesday schedule in that study area. For the Coarse PM Study snapshots, the intended middle day aligns all of the Coarse PM samples that were collected concurrently. SPIROMICS Air homes were deployed over the course of 1-2 weeks and do not have an intended middle day.

Measurements are reported as ppb for gases. Gas samples collected via Ogawas are analyzed using ion chromatography (IC) analysis, which is quite sensitive and can detect very small concentrations of ions. However, a small number of measurements were made that were below the detection limit of the instrument, also known as the lab LOD. These measurements will be reported as half the noise in the instrument's baseline.

Samples typically accumulate a small amount of contamination during handling and shipping. This level will often exceed the lab limit of detection. "Field blanks" are used to determine this level of contamination, and are used to determine a correction for contamination and the "field" limit of detection. The limit of detection is provided along with the measured value (after correction), as some users may choose to replace the measurements made below the LOD with other values. More information about correction and limits of detection can be found in the QA/QC report (see Appendix C).

Measurements are reported as $\mu\text{g}/\text{m}^3$ for particulate matter. Because particulate matter mass is determined gravimetrically for the two-week integrated method, and because the microbalance can always provide a filter's weight, lab LODs are not available. Other measures of uncertainty are addressed in the QA/QC report. The field LOD is determined for mass for PM, EC/OC, and elemental species. Data users will receive this LOD as a concentration, calculated by the volume of air sampled. Because the volume of air varies by sample based on sampling duration and flow rate, the concentration LOD will not be the same for all samples. As for gases, data users will receive both the corrected measurement and the LOD. RAD monitors evaluate particulate matter using a laser (Plantower). Data processing for this method has not been finalized.

1.3.9 Quality Control of Data

A full treatment of the QA/QC process for all of our monitoring data is out of scope for this document. Please refer to a project-specific Quality Assurance Project Plan (QAPP) and the final Quality Assurance/Quality Control Report (see Appendix C) as needed.

Flags are associated with a small number of measurements and are provided to data users. Measurements taken near a source of pollution or with concentrations that might be impacted by slight deviations from the sampling protocol are indicated with a Source or Concentration flag. A full description of the flags exists in the QAPP and the final QA/QC report (see Appendix C). There are also a

number of measurements that were noted as unbelievable by analysts that have worked extensively with these data. These flags are provided with a brief description of the issue that the analyst identified.

1.3.10 Raw Monitoring Data

Raw data include field log data, questionnaire data, and measurements. The collection, measurement, and transmittance of samples and data are covered extensively in the relevant field and lab Standard Operating Procedures documents (see Appendix C). Locations of the raw monitoring data files are listed in this Appendix.

1.4 Monitoring Data from Other Sources

1.4.1 Inhalable Particulate Network (IPN)

The Inhalable Particulate Network (IPN) was an EPA monitoring campaign from 1979 to 1984. This network included fine particulate measurements ($PM_{2.5}$) and coarse particulate measurements ($PM_{2.5-PM_{15}}$) taken from dichot samplers, “inhalable particulates” (PM_{15}) taken from size-selective input HIVOL samplers, and “total suspended particulate matter” (PM_{30}) taken from HIVOL samplers, as well as a variety of other particulate components.

The EPA no longer has any data associated with this network. We digitized a printout of the data that was transferred to us from colleagues at NYU. Note that while this dataset was printed in 1984, measurements only extend through 1982 (indicating the printed version we have was not the final dataset). This printout contained site codes but not site locations. We manually linked site codes to tables with geographic locations; these tables were found in two separate EPA documents (Analysis of Inhalable and Fine Particulate Matter Measurements 1981 and Directory of Air Quality Monitoring Sites Active in 1977). Out of the 132 sites with fine PM ($PM_{2.5}$) measurements, 102 of these sites were listed in either of the documents containing geolocations, and therefore 30 sites are missing lat/long data and geocovariates. Data from these 30 sites are therefore unusable.

Table 8. Description of particle pollution data available from IPN, including particle size and collection methods.

Pollutant	Equipment	Description	Particle size range (µm)
PM30	11 ½" x 15 " hi volume sampler (HIVOL)	Total suspended particulate matter (TSP) is the portion of suspended material collected by the HIVOL filter sampler for a 24hr period; 50% cut size from 30 to 65 µm for wind speeds between 2 to 24 km/hr	0 – 30
PM15	11 ½" x 15 " hi volume sampler (HIVOL) equipped with 15 µm size-selective inlet (SSI) and flow controller	USEPA "inhalable" particulate matter (IP); that portion collected by a sampler with a 50% cut size of 15 µm	0 – 15
PM25	Sierra 244, 244e or Beckman SAMPLAIR virtual impactor (aka dichotomous sampler)	Fine particulate matter; that portion of an aerosol which penetrates a particle collector with a 50% cut size at 2.5µm	0 – 2.5
PMcf	Sierra 244, 244e or Beckman SAMPLAIR virtual impactor (aka dichotomous sampler)	Course particulate matter; presumably everything from the dichot not included in mass_f	2.5 – 15

1.4.2 New York City Community Air Survey (NYCCAS)

NYCCAS collected 2-week measurements of air pollutants (PM_{2.5}, LAC, NO_x, O₃) at utility pole locations throughout the 5 boroughs beginning in December 2008. Our database currently contains measurements collected through November 2015. More information regarding site selection can be found on NYCCAS' website:

<http://www.nyc.gov/html/doh/html/environmental/community-air-survey.shtml>

1.5 Participant Address History

1.5.1 MESA Air and MESA Neighborhood Participants

The main MESA study collected current addresses on all MESA Classic participants from the baseline exam and updated these addresses at all subsequent follow up calls and clinic visits. At the beginning of MESA, addresses were only used for mailings, etc. and out-of-date addresses were overwritten by current addresses. In Exams 2 and 3, the MESA Neighborhood study, a study ancillary to MESA and directed by Dr. Ana Diez Roux at the University of Michigan, administered a Residential History Questionnaire to all MESA Classic participants who attended these exams and consented to MESA Neighborhood. This questionnaire acquired historical addresses from 1980 until the administration of the survey. MESA Air funded identical Residential History Questionnaires on all MESA Air New Recruits and MESA Air participants recruited from the MESA Family study during Exams 3/4. In 2006, MESA Air requested that all addresses collected by the main MESA study during clinic visits and follow up phone calls be retained by the Coordinating Center. Residential histories are available participants in the

MESA Neighborhood Study for the period between 1980 and Exams 2/3 (2002-2004), and from the main MESA study from 2006 onward. Living floor and building type are available for locations for which an Air Questionnaire was administered.

The MESA Neighborhood study cleaned all addresses they collected, and contracted with Mapping Analytics to geocode the cleaned addresses. In geocoding these addresses, Mapping Analytics employed a 50 foot offset from the centerline of the road. The EAC database includes the geocodes generated by Mapping Analytics for the addresses at which participants lived between 1980 and 1999, with a few exceptions as described below. The MESA Air EAC geocoded all addresses where participants reported residing from January 1, 1999 forward with ArcGIS using a 30 foot offset perpendicular to the street. During data cleaning, MESA Air recovered some addresses that Mapping Analytics was not able to geocode. These were geocoded with a 30 foot offset.

Efforts were made to establish an address history that was sensible and complete, such that each participant had a single residence for every time point between 1980 and 2010. In the residential history, not all addresses had move-in or move-out dates. Addresses were ordered by available move dates and track date (contact date or date of questionnaire administration). A missing move-in date was assigned as the day after the previous addresses' move-out date. A missing move-out date was assigned the day before the track date of the next address, unless the last-known date at the previous address was also a track date. If only consecutive track dates were available, then the midpoint of the dates was used as the move-out date and the next day was used as the move-in date. Some track dates were misclassified as move-in dates; these were identified by comparison to available exam dates and track dates. PO Box addresses were dropped whenever possible, though included (and flagged) if no other address appeared to be valid for the concurrent time period. If only month and year (not day) were available for move dates, the move-out date was assigned to the end of the month, and the move-in date was assigned to the beginning of the month. If the year was available for the move date, but not the month or day, the middle date of the year was used. If two addresses had the same track dates, the midpoint of the difference was used for the move dates. The majority of these rules were based on those employed by the MESA Neighborhood study at the University of Michigan.

Decease dates were provided by the Coordinating Center. For these participants, an end date for the last known address was set as the decease date. Current addresses for living participants have a null end date.

The majority of addresses included in the database are the participants' primary addresses, but some participants also provide secondary addresses, which are indicated as appropriate.

1.5.2 Technical Note: Address geocoding and geocode flags

Participants' home addresses (street address, city, state, and zip code) are geocoded in ArcGIS 9.2 or 9.3 (ESRI, Redlands, CA) using data provided by the MESA Coordinating Center and the TeleAtlas Dynamap 2000 v.16.1 road network (Boston, MA).

MESA Air participant locations are geocoded at the EAC using automated geocoding procedures in ArcGIS for all addresses that match up to a selected sensitivity (80% for this study). Spelling sensitivity and minimum match score are both set to 80 for automated geocoding. The default minimum

candidate score, 10, is used. Originally, the “side offset”² used was 30 feet, with a 0 foot end offset (the default in ArcGIS 9.2 and 9.3). Addresses that were geocoded onto A1, A2, or A3 roads due to the 0 offset setting were re-geocoded with a 3 percent end offset (the default in ArcGIS 10). In the event that ArcGIS is unable to match the address with 80% accuracy, it will prompt the user to match the addresses interactively. The user must exercise good judgment for this process, and common fixes include removing apartment or unit numbers, fixing spelling errors, and checking the address with Google Earth, Google Maps, or Bing Maps.

In addition, a subset of addresses was geocoded with the parcel-based geocoding available in ArcGIS10. This was done to increase the spatial accuracy of geocodes for residences located very close to roadways, since very small differences in locations near roads can have a large influence on exposures. Ideally, all locations would be geocoded using parcel-based geocoding, but it would be time-consuming to create new geocodes (and, more importantly, new geocovariates) for all locations, and the differences in locations between street- and parcel-based geocodes are generally small and likely only important for locations near roads. In addition, the data underlying parcel-based geocoding is not available for all locations, so at best, parcel-based geocoding could only happen in a subset of homes. To select the locations to re-geocode and relocate using the parcel data, we selected all locations identified as being within 150 m of a major road based on the street geocodes and, when possible, calculated the parcel geocodes. If the parcel geocode and street geocodes were more than 100m apart or more than 50% different, and the parcel geocode was more than 10 m from any major road, the parcel geocode replaced the original geocode. A street geocode was retained if the parcel geocode fell within 10 m of a road.

All addresses geocoded to the exact location by street geocoding will be noted as “Exact” under the geocode type; those geocoded to an exact location by parcel geocoding will be noted as “Parcel”. Occasionally a road may have more than one name, so any addresses that were geocoded with different street names to exactly the same location were considered exact matches. Some addresses were identifiable through Google Earth but could not be geocoded to the exact street number in ArcGIS. These addresses were geocoded to the nearest intersection on the same street. For invalid street addresses with a valid zip code, participant locations were geocoded to the centroid of the zip code. Intersection and zip code geocodes were always created via street (not parcel) geocoding methods. Invalid addresses, such as addresses with no valid zip code, PO Boxes, and addresses outside the continental US, were noted as fatal. Each address will be associated with one of the geocode types that are listed below.

² Roads are provided as line features by TeleAtlas, and as such have no width. The position of the line is considered to be the centerline of the road. The “side offset” is the number of meters away from the centerline that the address is presumed to lie in the perpendicular direction. Road shapefiles contain information as to which side of the street has even or odd street addresses, so that points are placed on the correct side of the street.

Table 9. Geocode types

Geocode Level of Precision	Geocoding Agent	Geocode Type
Address was not found (fatal)	EAC or Mapping Analytics	FATAL
Address was a PO Box (fatal)	EAC or Mapping Analytics	POBOX
Address is outside the continental US (fatal)	EAC or Mapping Analytics	FOREIGN
Outside the larger MESA Air Area	EAC or Mapping Analytics	OUT OF AREA
Exact match, 30 foot offset, 80% match rate	EAC	EAC-EXACT
Exact match to parcel data, no offset	EAC	EAC-PARCEL
Centroid of zip code	EAC	EAC-ZIP
Nearest intersection, 30 foot offset	EAC	EAC-INTERSECTION
Exact match, 50 foot offset, 100% match rate	Mapping Analytics	MICH-EXACT
Nearest intersection, 50 foot offset	Mapping Analytics	MICH-INTERSECTION
Centroid of block group	Mapping Analytics	MICH-BLOCK
Centroid of census tract	Mapping Analytics	MICH-TRACT
Centroid of zip code	Mapping Analytics	MICH-ZIP
Centroid of county	Mapping Analytics	MICH-COUNTY
Exact match to parcel data, no offset	V. C. Van Hee	EXT-PARCEL
Exact match (interpolated along the street), no offset	V. C. Van Hee	EXT-EXACT
Approximate match (interpolated along the street), no offset	V. C. Van Hee	EXT-POOR
Midpoint of the street, no offset	V. C. Van Hee	EXT-ST_CENTER
Centroid of zip code	V. C. Van Hee	EXT-ZIP
Centroid of city	V. C. Van Hee	EXT-CITY
Centroid of state	V. C. Van Hee	EXT-STATE

1.5.3 Cardiovascular Health Study (CHS) Participants

Geocoding for CHS was conducted by Victor C. Van Hee using ArcGIS 10 (TeleAtlas/TomTom) and Google. The ArcGIS geocode was used unless the Google geocode indicated a higher level of accuracy. No offset was applied and no data cleaning was performed by the EAC.

1.5.4 Women’s Health Initiative – Observational Study (WHI - OS) Participants

Addresses for participants in the Women’s Health Initiative were collected and geocoded by the WHI Coordinating Center. These addresses were collected at the participants’ initial interview and at follow-up. Follow-up after 2009 was conducted primarily by mail. The latitudes and longitudes of these locations were transmitted to the EAC without street addresses, participant names, or participant unique numbers. The exact method of geocoding for these addresses is not known, and no data cleaning was performed by the EAC.

1.5.5 SPIROMICS Air Participants

Addresses were provided by the Coordinating Center at UNC with a dummy participant ID and were geocoded at the EAC. Participants were asked to provide the current address at baseline as well as all addresses at which they had lived for 10 years prior to the study. The current address and any recent addresses were recorded at follow-up visits and during quarterly follow-up calls. Geocoding for Rev 11 was conducted using Business Analyst for ArcGIS 10.3, with no offset.

1.5.6 ACT – Air Pollution Study

Billing addresses were provided by Kaiser Permanente Research Institute (KPRI) with a dummy participant ID and were geocoded at the EAC. Geocoding for Rev 11 was conducted using Business Analyst for ArcGIS 10.3, with no offset.

1.5.7 The Sister Study and the Two Sister Study Participants

Addresses for participants in The Sister Study and the Two Sister Study were geocoded at the EAC. The Sister Study is interested in four kinds of addresses: “current” (at enrollment), secondary (at enrollment), longest-lived, and childhood (indicated in that order by a string of 4 binary digits at the end of the native_id). Participants were enrolled between 2003 and 2009. The same types of addresses were geocoded for the Two Sister Study. Participants in The Sister Study were women whose sisters had breast cancer; the Two Sister Study was an ancillary study to The Sister Study that enrolled the women with breast cancer themselves (the sisters of The Sister Study original participants.) Participants in the Two Sister Study were enrolled between 2008 and 2010. Geocode types are similar to the EAC types that are listed in Table 6: exact, intersection, zip, or fatal.

1.5.6 PSID Study Participants

IDs of census blocks where participants in the Panel Study on Income Dynamics (PSID) reside were provided by Kyle Crowder’s group. Locations were approximated at the EAC by calculating census block centroids using ArcGIS 10.

1.5.7 BRFSS Study Participants

Zip codes were provided for survey respondents in the Center for Disease Control’s (CDC) Behavioral Risk Factor Surveillance System (BRFSS) in Washington State. Depending on the area of the ZCTA, between one and four census block centroids within each zip code were randomly selected to represent the average pollution level. Census blocks were required to have a total year 2010 population > 0. Geocovariates and predictions were generated for each point selected.

1.5.6 Small-Fee Projects

Collaborators sometimes send us a small set of locations for which they would like model predictions and sometimes geographic covariates. These cohorts include: 1) 1000 locations relevant to the OMEGA study, 2) a group of women that underwent IVF, 3) residence locations for the Washington Heights Inwood Aging Project (WHICAP), 4) a small number of locations relevant to the Religious Orders Study (ROS), 5) residence locations for Ginkgo Evaluation of Memory Study (GEMS) and 6) monitor and other locations for a study conducted in Oakland.

1.6. Participant Exam Dates

The dates on which each MESA or MESA Air participant came into the clinic for each of the five MESA exams are provided to the EAC by the Coordinating Center. In addition to the primary exam date, the Coordinating Center also provided the dates for each participant’s coronary artery CT scan, ultrasound, spirometry, and MRI as these tests sometimes occurred on a separate day.

Since a very limited number of participants have an exam at a given field center on any particular day, these dates are considered identifying information and, as such, cannot be distributed by the EAC. These dates are used by the EAC to provide “year prior to exam” averaging or time boundaries. Thus, in some cases, address histories and exam dates may be ‘masked’ by providing a move date or exam date as the number of days since an event of interest (such as a participant’s baseline exam).

In addition, baseline exam dates for SPIROMICS Air participants are available.

1.7. Participant Time Location

As part of the MESA Air Questionnaire, participants reported location patterns by season (summer or winter) and day of the week. This included questions specific to the number of hours spent in transit, at home indoors and out, at work indoors and out, and at “other activity locations” (such as volunteering) indoors and out. The primary purpose of these data is to provide time-weighted, infiltration-adjusted, aggregated estimates of PM_{2.5} exposure, or to provide the percent of time spent indoors and outdoors. For these calculations, we sum the total time spent in indoor locations (reported as home indoor, work indoor, and other indoor) and outdoor locations (reported as home outdoor, work outdoor, and other outdoor) for each season-day. We then average time indoors/outdoors across all days of the week separately for the summer and winter. When calculating individual-level exposures integrating indoor and outdoor concentrations with time-location information, the “summer” answers will be used when the two-week average temperature exceeds 18 degrees Celsius, and the “winter” answers will be used for periods with average temperatures equal to or below 18 degrees Celsius. A number of participants did not complete the Air Questionnaire, and we imputed missing responses for these participants. In instances where a participant completed some, but not all, of the time-location section of the questionnaire, we assumed that a missing day was the same as a weekday or weekend day in the same season. Otherwise, it was assumed to be the average of non-missing days. If an entire season was missing, we used the responses from the other, non-missing season. For more specific analyses involving time-location, the raw, unaggregated data are available from the Coordinating Center.

1.8. AQS Monitors for ‘Acute’ Analysis (Averaging Period < 1 Year)

Studies of the effects of short-term variations in air pollution exposure (“acute analyses”) generally leverage the variability in day-to-day pollution concentrations, rather than the variability in concentrations over space. The most common exposure period for this type of analysis is 1, 2, or 3 days prior to the exam or event. As MESA Air and SPIROMICS models are generally on the 2-week resolution, AQS data must be relied upon for these acute analyses.

The MESA Air EAC developed a daily time series of PM_{2.5} data collected by AQS monitors in each of the MESA Air study regions (Baltimore, Chicago, New York/Rockland County, LA/Coastal, Riverside (defined as those locations in Riverside or San Bernadino County), St. Paul, and Winston-Salem), as well as for Ann Arbor, San Francisco and Salt Lake regions. To ensure that analyses are not biased by the effects of inconsistent or erratic monitor schedules, we only included data from monitors within these regions that recorded daily PM_{2.5} levels using a federal reference method and were in operation from July 1999 through the end of 2011.

In areas other than LA/Coastal and NYC, these time series consists of monitoring data from a single monitor in each region. In LA, there were several qualifying monitors, and a single time series was constructed by taking each monitor's mean concentration, subtracting that from each individual measurement for that monitor, averaging the residuals across monitors for each day, and then adding the average of the monitors' means back to each of the daily averaged residuals. In NYC, the only long-running daily monitor, 360050110, is missing 73 days of data in 2001 and was offline from mid-2010 through mid-2012. To create a time series with better coverage, data from the ten nearest monitors that operated while 360050110 was offline were used in addition to this daily monitor. The ten monitors were combined according to the procedure used in LA to approximate a regional trend, which

was then used to model the output at site 360050110. Measurements that were modelled in this way are flagged in the column 'estimate_flag'.

Additionally, the time series in St. Paul is based on monitor 271230871. This monitor collected data on a daily basis during the Exam 1 period but switched to a 1-in-3 day schedule in 2003. No other federal reference method (FRM) daily monitoring data was available for this area. In order to improve coverage, the FRM measurement was modelled from the measurements reported from two continuous monitors co-located with the FRM monitor, but using non-reference methods. While this supplemented time series has been made available for analysts who require daily data, it should be noted that the agreement between non-reference and reference methods is not great (mean RPD = 19%) and analysts should consider using a three-day average in Saint Paul for acute analysis. For more information, see the acute monitoring appendix. As with NYC, these estimates are flagged in the column 'estimate_flag'.

Analysts should note that studies that include a spatial component and an averaging time greater than 14 days may find an average of modeled exposures (see Section 2.12) more appropriate than these AQS data. Requests that include an exposure time period between two weeks and one year will be provided both an average of the acute time series and appropriate modeled exposure averages, and should determine the most appropriate exposure metric for their specific analysis.

1.8.1 Pre-Adjusted PM_{2.5} Exposure

Some users may prefer to conduct an analysis using PM_{2.5} that has been pre-adjusted for seasonal variability, to ensure that observed effects are PM_{2.5}-related rather than (say) temperature-related. This would primarily be of concern when studying acute outcomes that vary seasonally, and is primarily directed at outcomes which may be 'triggered' by unusually high deviations of PM_{2.5} from recent and typical levels. Pre-adjusted PM_{2.5} exposures are the residuals from a prediction model of PM_{2.5} that includes 6 degree of freedom per year b-splines on temperature and humidity, and 12 degrees of freedom per year on calendar time, and with categorical adjustment for day of week. The R code for pre-adjustment can be made available for analysts desiring a different model specification.

1.9. AQS Monitors Eligible for 'Nearest Monitor' Analysis (Averaging Period >= 1 Year)

Simplified spatial analyses can be conducted using the 'nearest monitor' average PM_{2.5} as the exposure estimate for a particular location. Generally we recommend using a modeled exposure for the primary hypothesis, but there are some cases in which analysts may want to compare results from nearest monitors. Nearest monitor analyses are not recommended for pollutants other than PM_{2.5} which vary significantly on the local scale, rather than the regional scale. Due to seasonality in air pollution trends, this analysis is usually recommended for exposure periods of one year or longer.

Monitors are typically considered eligible for inclusion in a nearest monitor analysis if they provide good coverage for the entire study period. Specific monitors have been identified for MESA Air regions from 1999-2012. Other regions and time periods will require monitor selection.

Distances to AQS monitors within 100 kilometers of each participant home were calculated via the spatial-statistics 'sp' package of the R language. The standard distance function of that package uses location latitudes and longitudes and the Great Circle Distance formula to calculate distances on a spherical surface without the need for planar projections. See Appendix C for the location of the code.

1.10. Grid Locations

In order to produce maps that reflect the spatial heterogeneity inherent in our exposure models, geographic covariates were calculated for grid points at several different scales. For city-specific models and maps, grid points were generated at three densities in each MESA Air area. Grids were designed to show a high level of detail in the areas with the greatest density of participant addresses (a dense grid at 500 m resolution) and to show less detail in outlying and rural areas (1 km resolution and 2 km resolution at the edges of the grid area). In addition, fine-scale grids at a 50 m resolution were generated in order to produce insets that highlight very small-scale variation in model predictions. For the national models, a grid with a 25 km resolution was created. This grid covers the continental US.

1.11. Geographic Covariates

The following table summarizes the geographic variables that are available and the sources of the data. Subsections follow that address each data source in more depth. Geographic covariates are typically not available outside the continental US.

Table 10. Available geographic information

Data Category	Source	Variable Name	Available Buffer Sizes
Airports ^a	NEI Database	m_to_airp, m_to_l_airp (large airport)	N/A
Coastline ^a	TeleAtlas	m_to_coast	N/A
Railroads ^a	TeleAtlas	m_to_rr	N/A
Railyards ^a	TeleAtlas	m_to_ry	N/A
City Hall ^{c,d}	Google Maps	m_to_main_cityhall, m_to_local_cityhall	N/A
Roads	TeleAtlas	ll_a<type>_s<radius>	50m, 100m, 150m, 300m, 400m, 500m, 750m, 1km, 1.5km, 3km, 5km
		m_to_a1 ^a , m_to_a2 ^a m_to_a3 ^a	N/A
Intersections	TeleAtlas	intersect_<type>_s<radius>	500m, 1km, 3km
		m_to_<type>_intersect ^a	N/A
Population (US)	US Census Bureau	pop_s <radius>	500m, 1km, 2km, 2.5km, 3km, 5km, 10km, 15km
Land use/ Commercial Land Use	MRLC 2006 National Landcover Dataset	rlu_<type>_p<radius>	50m, 100m, 150m, 300m, 400m, 500m, 750m, 1km, 1.5km ^g , 3km, 5km, 10km ^g , 15km ^g
	USGS historical source	lu_<type>_p<radius>	
		m_to_comm ^a	N/A
Ports ^a	National Geospatial Intelligence Agency	m_to_l_port, m_to_m_port, m_to_s_port	N/A
Emission Sources	NEI Database	em_<poll>_s<radius>	3km, 15km, 30km
Truck routes ^a	Bureau of Transportation Statistics	tl_s<radius>	50m, 100m, 150m, 300m, 400m, 500m, 750m, 1km, 1.5km, 3km, 5km, 10km, 15km
		m_to_truck	N/A
Impervious Surface	National Landcover Dataset	impervious_a<radius>	50m, 100m, 150m, 300m, 400m, 500m, 750m, 1km, 3km, 5km
Census Data	US Census	see Table 20	block, block group, and tract level

Table 10, continued

CALINE Long-Term Average ^d	Internal ⁱ	calinemod_lt_a<radius>, caline_alternate_a<radius> ^h	1.5km, 3km, 6km, 9km
Residual oil ^{c,e}	Environmental Defense Fund	m_to_oil, m_to_6oil	N/A
		oil_edf<oil grade>_s<radius>	100m, 150m, 300m, 500m, 750m, 1.5km, 3km
Bus routes ^{a,e}	NY Department of Transportation	bus_s<radius>	50m, 100m, 150m, 300m, 400m, 500m, 750m, 1km, 1.5km, 3km, 5km
		m_to_bus	N/A
Elevation ^f	National Elevation Dataset	elevation, elev_<radius>_<type>	1km, 5km
Urban Topography ^j	NYC PLUTO and City of Chicago building footprints	canyon_<type>	N/A
NDVI	University of Maryland	ndvi_<type>_a<radius>	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
Columnar NO ₂	UMN	satellite_NO2	N/A
Columnar NO ₂	UMN	no2_behr_<yyyy>	Moore neighborhood
Satellite CO		satco_<yyyy>	Years 2000-2016
Satellite PM _{2.5}		satpm25_<yyyy>	Years 1998-2014
Satellite NO ₂		satno2_<yyyy> satno2_<yyyy1>_<yyyy2>	Years 2005-2016
Satellite SO ₂		satso2_<yyyy>	Years 2005-2016
Satellite HCHO		sathcho_2005_2016	
MOVES emissions	Internal		N/A

^a Distances calculated to spatial features are truncated at 25 km

^c Distances calculated to spatial features are truncated- see covariate-specific section

^d Available only in MESA Air areas

^e Available only for New York City

^f Not buffered in the same way that other variables were

^j Available for NYC and Chicago

^g Available for USGS historical source only

^h Available only for LA/Riverside

ⁱ See the “Documentation of MESA Air Implementation of CALINE3QHR Model” for details on the inputs to the CALINE dispersion model

1.11.1 Sources of GIS data

Aside from MESA Air monitoring and home locations, geographic data are obtained from various external sources, such as TeleAtlas, the US Census Bureau, and US geological survey. All of these data are free of charge and can be acquired at any time, with the exception of the TeleAtlas data. The TeleAtlas database was obtained from the USEPA under their usage license via DVD. This DVD is currently stored at the EAC; contact the data manager if necessary. A table of other sources and their websites where data can be downloaded are listed below by data category. For detailed publication sources, publication dates and accuracy information of obtained data, please contact the EAC for the metadata document.

Table 11. Data sources

Source	URL/Contact
NEI Database	http://www.epa.gov/ttn/chief/net/2002inventory.html
TeleAtlas	http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm
US Census Bureau	http://arcdata.esri.com/data/tiger2000/tiger_download.cfm
US Geological Survey	http://water.usgs.gov/GIS/dsdl/ds240/index.html
National Geospatial Intelligence Agency	World Port Index (Pub 150) http://164.214.12.45/MSISiteContent/StaticFiles/NAV_PUBS/WPI/WPI_Shapefile.zip
EPA AQS	http://www.epa.gov/ttn/airs/airsags/detaildata/downloadaqdata.htm
IMPROVE	http://views.cira.colostate.edu/web/
NDVI	http://glcf.umd.edu/data/ndvi/
National Landcover	http://www.mrlc.gov/index.php
Bureau of Transportation Statistics	http://www.bts.gov/publications/national_transportation_atlas_database/
NYC PLUTO 2004	http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml
Chicago Building Footprints	https://data.cityofchicago.org/Buildings/Building-Footprints/w2v3-isjw
National Elevation Dataset	http://nationalmap.gov/elevation.html
Environmental Defense Fund	Contact EDF attorney Isabelle Silverman
UMN Columnar NO ₂ (census block centroids)	Contact Julian Marshall ³
Satellite PM _{2.5}	http://fizz.phys.dal.ca/~atmos/martin/?page_id=140
Satellite NO ₂	http://www.temis.nl/airpollution/no2.html
Satellite SO ₂	https://disc.gsfc.nasa.gov/datacollection/OMSO2_CPR_003.html
Satellite HCHO	https://disc.gsfc.nasa.gov/datasets/OMHCHO_V003/summary
Satellite CO	https://eosweb.larc.nasa.gov/datapool

1.11.2 Creation or Projection of Shapefiles from Raw Data Sources

All geographic variables must be calculated from shapefiles. Emissions data are downloaded as flat files and shapefiles are created from latitude and longitude at the EAC. USGS land use, TeleAtlas road data, census data for the year 2000, and NDVI images are downloaded as shapefiles from the source website. All shapefiles were re-projected into State Plane Zones and clipped with a 25 kilometer ‘buffer’ that extends beyond the boundaries of the State Plane Zone. This enables geographic calculations for points near the border of a particular zone.

1.11.3 Land use data

Variables for land use as percentage of a buffer were calculated from two sources. USGS polygon layers, generated by manually-intensive methods using aerial photography from the 1970s and 1980s were used to calculate the variables with the “lu” prefix. Rasters based on satellite data from the year 2006 were obtained from the Multi-Resolution Land Characteristics (MRLC) Consortium, and these were used to calculate the variables with the “rlu” prefix. In general, the EAC is recommending that the Raster

³ Novotny EV, MJ Bechle, DB Millet, and JD Marshall. 2011. "National satellite-based land-use regression: NO₂ in the United States". *Environmental Science & Technology*. 45 (10): 4407-14.

Land Use data be used for exposure models developed to reflect “current” exposures (e.g., from 1999 to the present) and the USGS data be used to calculate “historical” exposure (prior to 1999). Please note that is not advisable to create an exposure model that includes both sets of land use variables. Please contact the EAC with further questions on which set of land use variables to use.

Tables 9 and 10 contain the full lists of the possible land use designations, with rough equivalents between the two sources. More information about the USGS land use classifications can be found at <http://landcover.usgs.gov/pdf/anderson.pdf>. The stated positional accuracy for USGS land cover is approximately 200 meters. This affects all areas, and may produce unexpected results, especially within small buffers. Analysts are encouraged to scrutinize results related to these variables. The positional accuracy for the satellite-based rasters is 30 meters and these variables may be more reliable than those calculated from the USGS polygon files. More information about the satellite-based land use classifications can be found here: http://www.mrlc.gov/nlcd06_leg.php. However, analysts are cautioned that the ice and snow designations in this data source may potentially be inaccurate, as pavement is occasionally misclassified as this land use type.

As an additional note on the processing of the polygon files, the USGS organizes land cover data by grids. Multiple grids might be required for a single State Plane Zone. Land use grids are merged into a single shapefile and projected accordingly. Commercial land use is selected and exported as a separate data category for the distance to commercial land use calculation.

Table 12. Land use variable names (lu_<type>_p<radius>) for variables based on 1970s and 1980s aerial photography

Variable Description	Variable Name (<type>)	Raster Land Use Equivalent
Urban or Built-Up Land		
Residential	resi	dev_lo, dev_med
Commercial and services	comm	dev_hi
Industrial	industrial	dev_hi
Transportation, communications, and utilities	transport	dev_hi
Industrial and commercial complexes	industcomm	dev_hi
Other urban or built-up land	oth_urban	No specific variable ^b
Mixed urban or built-up land	mix_urban	No specific variable ^b
Agricultural Land		
Cropland and pasture	crop	Crop
Orchards, groves, vineyards, nurseries	grove	Crop
Confined feeding operations	feeding	Crop
Other agricultural	oth_agri	Crop
Rangeland		
Herbaceous rangeland	herb_range	Grass
Shrub and brush rangeland	shrub	Shrub
Mixed rangeland	mix_range	Shrub
Forest Land		
Deciduous forest land	forest	decid_forest
Evergreen forest land	green	evergreen
Mixed forest	mix_forest	mix_forest

Variable Description	Variable Name (<type>)	Raster Land Use Equivalent
Water		
Streams and canals	stream	water
Lakes	lakes	water
Reservoirs	reservoir	water
Bays and estuaries	bays	water
Wetland		
Nonforested wetland	nf_wetland	herb_wetland
Wetland	wetland	woody_wetland
Barren Land		
Beaches	beach	barren
Dry salt flats	dry_salt	barren
Sandy areas other than beaches	sandy	barren
Strip mines, quarries, and gravel pits	mine	barren
Transitional areas	transition	dev_hi, dev_med, dev_lo, shrub
Bare exposed rock	rock	barren
Mixed barren land	mix_barren	barren
Tundra		
Herbaceous tundra	herb_tun	grass, barren, shrub
Bare tundra	bare_tun	ice
Wet tundra	wet_tun	herb_wetland
Mixed tundra	mix_tun	No specific variable ^c
Perennial Snow or Ice		
Perennial snowfields	snowfield	ice
Glaciers	glacier	ice
Other		
Not specified (usually outside US boundaries: ocean, Mexico, or Canada)	unspec	
The minimum distance to an area designated "Commercial and Services" ^a	m_to_comm	

^a Provided as a "distance" variable, not as land use in a buffer

^b Mixture of developed, agricultural, and natural areas

^c Mixture of categories

Table 13. Land use variable names (rlu_<type>_p<radius>) for variables calculated from satellite-imagery rasters

Variable Description	Variable Name (<type>)
Open water	water
Perennial ice, snow	ice
Developed open space	dev_open
Developed low intensity	dev_lo
Developed medium intensity	dev_med
Developed high intensity	dev_hi
Bare rock, sand, barren, mine	barren
Deciduous forest	decid_forest
Evergreen forest	evergreen
Mixed forest	mix_forest
Shrubland	shrub
Grasslands, herbaceous vegetation	grass
Pasture, hay	pasture
Cultivated crops such as orchards, vineyards, grains	crop
Woody wetlands	woody_wetland
Emergent, herbaceous wetland	herb_wetland

1.11.4 TeleAtlas Road Data

TeleAtlas road data were obtained from the EPA for the year 2000. These road networks had duplicate road segments in some areas. That is, two road segments with the same name and other identifying information were occasionally found to lie one right on top of the other. These were considered digitization errors and were removed with ArcGIS via python code.

Feature class codes are used to categorize roadways. Limited access highways are designated as A1 roads. Other major roads, such as state and county highways without limited access, are designated as A2 or A3 roads. More detailed information about the road classification system used in our database can be found at <http://www.maris.state.ms.us/pdf/CFCCcodes.pdf>.

1.11.5 Distance to Road and Near Road Determination

The perpendicular distance from locations to A1, A2, and A3 roads was calculated, based on the TeleAtlas road network and the geocoding of addresses as specified in section 2.4. Locations are considered ‘near road’ if the distance to and A1 or A2 is less than 100 meters, or if the distance to an A3 is less than 50 meters.

1.11.6 Airports and Major Airports

Airport shapefiles and airport emission sums are obtained from NEI database. Runways of airports are subsetted, and centroids of runways were calculated. Emissions in tons and freight numbers were then merged to centroids of runways and projected into appropriate State Plane projections. Airports were classified as “major” according to the number of passengers per year served by the airport and by the freight tonnage. A value of approximately 8,000 passengers per year was determined to be the approximate center of the distribution of passenger volumes among all airports in the study area. Approximately 160,000 pounds per year was determined to be the approximate center of the distribution for freight. Airports reporting passenger volumes or freight tonnage above these values,

and all international airports, were considered major airports. Therefore, distance to major (large) airport (m_to_l_airp) was calculated separately from the distance to any airport (m_to_airp).

1.11.7 Coastlines, Railroads, and Rail Yards

Coastlines, railroads, and rail yard locations were obtained from the TeleAtlas geodatabase.

1.11.8 Ports

Port locations were obtained from the World Port Index from the National Imagery and Mapping Agency (now the National Geospatial Intelligence Agency). The designations of 'small', 'medium', and 'large' are made by that agency. The documentation says only that "the classification of harbor size is based on several applicable factors, including area, facilities, and wharf space. It is not based on area alone or on any other single factor."

1.11.9 Distance to Nearest Truck Route and Length of Truck Routes in Buffers

Truck route data was obtained from the National Transportation Atlas Database 2009 for data collected in 2008. The distance to the nearest truck route in meters is provided (m_to_truck). Truck route lengths in buffers were also summed (tl_s<radius>), in the same manner as road lengths.

1.11.10 Population

Population data at the block group level is obtained from US Census Bureau for the year 2000 and block group boundaries are extracted from TeleAtlas database. Data from these two sources are then merged together by blockgroup key to create a new shapefile. This shapefile is then split into a series of shapefiles based on its designated State Plane projection. New columns are added to calculate area of each block group polygon in km² and then to calculate the population density in number of persons per km². Points locations are buffered, and the population densities and areas of the block groups within the buffer are used to calculate the total number of individuals within certain radii (measured in meters). These variables appear in the database with names such as pop_s01000.

1.11.11 Emissions Data

EPA's Emission Inventory and Analysis Group prepares a national database, the National Emission Inventory (NEI), of air emissions information with input from numerous State and local air agencies, from tribes, and from industry. The NEI database includes estimates of facility-specific Criteria Air Pollutants (CAPs) and Hazardous Air Pollutants (HAPs) emissions, along with their source-specific parameters necessary for modeling, such as location and facility characteristics (stack height, exit velocity, temperature, etc.). The latest and most updated NEI data available is the third and final version of the [2002 NEI](#) data posted in January 2008. Using these data, the EAC sums the total tonnage of NO_x, SO₂, PM_{2.5}, PM₁₀, and CO emitted from short stacks within 3 km and from tall stacks between 3km and 15 km or 30 km of locations of interest. Facilities having stack heights of at least 30 meters (98.425 feet) are considered "tall" stack facilities. The rest are considered "short" stack facilities.

Table 14. Emissions variable names

Variable Description	Variable Name
Sum of major emissions from short stacks within 3 km	em_NOx_s03000, em_SO2_s03000, em_PM25_s03000, em_CO_s03000, em_PM10_s03000
Sum of major emissions from tall stacks within 15 km, minus the emissions from tall stacks within 3 km	em_NOx_s15000, em_SO2_s15000, em_PM25_s15000, em_CO_s15000, em_PM10_s15000
Sum of major emissions from tall stacks within 30 km, minus the emissions from tall stacks within 3 km	em_NOx_s30000, em_SO2_s30000, em_PM25_s30000, em_CO_s30000, em_PM10_s30000

1.11.12 Normalized Difference Vegetation Index

NDVI was obtained as a series of 16-day composite satellite images from the year 2006. The index was converted by the University of Maryland from the -1 to 1 scale to the 0-255 (pixel brightness) scale. On this scale, water has a value of approximately 50 and areas with dense vegetation have values around 200. For each location of interest, for each image, all pixels with a centroid within a certain distance of the location were averaged (radii included 250m, 500m, 1km, and 5km). For each buffer size, five summary numbers were calculated from the series of 23 averages for each location: the 25th, median, and 75th percentile of the entire year's series, the median of the expected 'high vegetation' season, defined as April 1 - September 30, and the median of the expected 'low vegetation' season, defined as the rest of the year.

Table 15. NDVI variable names

Variable Description	Variable Name
25 th percentile of entire year	ndvi_q25_a<radius>
Median of entire year	ndvi_q50_a <radius>
75 th percentile of entire year	ndvi_q75_a<radius>
Median of April through September	ndvi_summer_a<radius>
Median of January through March and October through December	ndvi_winter_a<radius>

1.11.13 Impervious Surface

Impervious surface was obtained from the Multi-Resolution Land Characteristics (MRLC) Consortium's National Landcover Dataset. Imperviousness was calculated by the MRLC from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) satellite imagery. Briefly, the MRLC uses regression tree methods and the reflectance of different wavelengths of energy that are measured by satellite to characterize land cover. Image data are from 2006. Imperviousness refers to the percentage of area in a pixel that is covered with an impervious surface, such as pavement or concrete. Covariates provided by the EAC are averages of pixel values within various radii.

1.11.14 CALINE Long-Term Average

Dispersion of air pollutants was modeled using the CALINE3QHCR line-source model. This model integrates road position, traffic volume, diurnal traffic patterns, and meteorology to produce an hourly estimate of the concentration of a generic, non-reactive pollutant at specific locations known as "receptors". These concentrations have arbitrary units, and are not comparable between different MESA Air regions. Hourly estimates were averaged over time to produce a long-term (2004-2006)

average concentration. This value can be used as a spatial covariate representing the average pollution exposure due to major-road traffic, relative to other locations in the same region.

In the EAC database Rev 4, slight modifications of CALINE3QHCR were implemented, based on prior experience and operational considerations. Datasets generated prior to Rev 4 contain “CALINE”-type covariates, whereas datasets generated at Rev 4 or later contain ‘CALINEMOD’-type covariates. The modifications included:

- The minimum distance between receptors and any road link was set to 20m, to reduce outliers due to positional-accuracy limitations.
- The minimum mixing height was set to 50m (CalPuff default)
- At any specific hour, only road links within +/-30 degrees of the upwind direction from the receptor were considered.
- The Python code preparing road link input files from shapefiles was standardized to include only links within 9km and +/-30 degrees, and to reduce run time and storage overhead.

In the greater Los Angeles and Riverside study areas, due to the geographical extent of MESA AIR cohort sub-regions (120km east to west) and the meteorological heterogeneity, separate CALINE values were calculated using meteorological stations situated at the area’s eastern and western edges. These are available for all monitoring locations, and for many other locations (see below).⁴ For tractability and reproducibility reasons, in the LA area the CALINE variable is designated as primary or “alternate” according to standard geographical boundaries. The designation is as follows:

- In LA, Ventura and Orange counties, the primary (“calinemod_lt_a0xxx”) variable is calculated using the western, LAX airport meteorology. The alternate variable (“calinemod_alternate_a0xxx”) is available for all monitoring locations, and for cohort and grid locations >15km from the coast or in State Plane Zone 6.
- In Riverside and San Bernardino counties, the primary variable is calculated using the eastern, March Air Force Base meteorology (10km SE of Riverside). The alternate variable is available for all monitoring locations, and for cohort and grid locations <25km from the coast, in State Plane Zone 5, or west of -117.7 degrees longitude.
- In other counties (if relevant), the primary variable is designated according to state plane – LAX for locations in Zone 5 and March AFB in Zone 6.

CALINE, auxiliary code, and further documentation are available in the CALINE directory of MESA-AIR’s Subversion repository.

1.11.15 Motor Vehicle Emissions Trends

Modeling was performed using the Motor Vehicle Emission Simulator 2014 (MOVES2014) software package⁵ for 10 parameters in 21 counties that cover the MESA Air study areas. These parameters were: CO, NO₂, NO_x, Total PM_{2.5}, PM₁₀ and PM_{2.5} from brake wear, PM₁₀ and PM_{2.5} from tire wear, EC, Non-EC PM, sulfate, aerosol H₂O, and total energy expended by vehicles.

Vehicle types were classified into two distinct categories. These categories were broadly labeled “Heavy Duty Vehicles” and “Light Duty Vehicles.” Only the contribution of vehicles travelling on urban restricted

⁴ Additional meteorological stations closer to the region’s center have been attempted, but at present none have been found with reliable hourly-resolution data for wind direction and speed.

⁵ EPA Office of Transportation and Air Quality

access and urban unrestricted access road types were considered in this analysis. Each possible combination of road type and vehicle type were paired and run separately.

Table 16. Classification of vehicle types

Light Duty Vehicles
Light Commercial Truck
Motor Home
Motorcycle
Passenger Car
Passenger Truck
Heavy Duty Vehicles
Combination Long-haul Truck
Combination Short-haul Truck
Intercity Bus
Refuse Truck
School Bus
Single Unit Long-haul Truck
Single Unit Short-haul Truck

Monthly values are available for the parameters by county for the years 1990 and 1999-2014, by vehicle/access category (light duty restricted access, light duty unrestricted access, heavy duty restricted access, heavy duty unrestricted access).

1.11.16 Residual Oil in New York City

Residual oil boilers are associated with high emissions of soot and of some elements (e.g., sulfur and nickel). Such boilers are common in the New York City metropolitan area. They are used in medium-to-large apartment buildings and in institutions such as hospitals and colleges. Most of the emission is usually through roof chimneys, rather than at street level. Residual oils graded 4 and 6 are the heaviest fractions of petroleum distillation, with 6 being heavier and dirtier.

Data were received from Environmental Defense Fund (EDF), who transferred the information to the public domain via a FOIA from New York City government. Data were cleaned of a few gross location and BTU-capacity errors. Boiler locations were geocoded by the MESA-Air EAC. Only boilers installed prior to 2007 were included (>95% of the original EDF dataset).

Table 17. Residual oil variable names

Variable Description	Variable Name
Meters to nearest residual oil grade 4 or 6 boiler; censored between 30m and 2000m	m_to_oil
Meters to nearest Residual Oil 6 boiler; censored between 30m and 2000m	m_to_6oil
Sum of total residual oil active heating capacity (in Mega-BTU per hour)	oil_edf<oil grade>_s<radius>

1.11.17 Distance to City Hall

This variable is available only for MESA regions, out to a distance of about 100-150km from each MESA city. Additionally, some challenges were encountered in converting the calculation to the psql process and in Rev 11 it is missing for many locations. It can be used as a proxy for the excess emissions due to congestion and increased activity around the center of a metropolitan area, above and beyond the emissions predicted by other variables such as population, land use or roads.

For the 3 smaller MESA cities (Baltimore, St. Paul, Winston-Salem), the variable provides the distance in meters to the nearest city hall of a city as large as the MESA city or larger. The distance is censored between 1 and 25 km. Therefore, the urban center is effectively approximated as a circle 2km in diameter, with all points inside the circle assumed equally "central". The reference city halls for these study areas are:

Baltimore - Baltimore, DC (White House), Philadelphia.

St. Paul - Minneapolis, St. Paul.

Winston-Salem - Winston-Salem, Greensboro, Charlotte, Durham, Raleigh, Roanoke (VA).

In this MESA region the minimum distance is 500m rather than 1km, since most of the reference cities are smaller, including Winston-Salem itself.

For the 3 larger MESA cities, distances to two locations were calculated. The "m_to_main_cityhall" variable measures the distance to the center of the main MESA city only. It is censored between 3km (2km in Chicago) and 50km. It is used as a proxy for the overall metropolitan-center plume. In addition, for these 3 cities the "m_to_local_cityhall" variable also includes city halls of edge cities (but, generally, not of continuous suburbs to the main city). Distances are censored between 1 and 25 km. The added cities are:

Chicago - Milwaukee, Madison, South Bend (IN).

Los Angeles - Riverside, San Bernardino, Ontario, Pomona, Long Beach, Santa Ana, San Diego, Bakersfield.

New York - Newark, Trenton, Bridgeport (CT), New Haven (CT).

The list of added cities was assembled to accommodate the location of agency and MESA monitors in the outlying areas that have been used in MESA Air exposure models. Values are missing for locations outside the main study areas and for locations geocoded to zip code centroids.

City halls were located via Google Maps, with geocodes rounded to the 3rd digit (~100m). Distances were calculated using R "sp" package function "spDistsN1", which calculates distances directly from the geocodes, assuming a spherical Earth.

1.11.18 Elevation

Each MESA location point had an elevation value in meters extracted from a National Elevation Dataset (NED)-based raster grid, provided by the USGS. In the continental US and Hawai’i, the resolution of the raster was 1 arc second. The resolution was 2 arc seconds for Alaska. A concentric set of twenty four elevation points were then generated around each location point for a statistical sample at two radii (1000 m. and 5000 m). Standard deviation and counts of “above” or “below” a threshold elevation (+/- 20 or 50 meters respectively) were calculated for each point utilizing the twenty four point sampling.

Table 18. Elevation variable names

Variable Description	Variable Name
Elevation above sea level in meters	elev_elevation
Standard deviation of elevation of twenty points surrounding location	elev_1k_stdev, elev_5k_stdev
Count of points (out of 24) within 20m of the same elevation	elev_1k_at_elev
Count of points (out of 24) more than 20m uphill or downhill of the location	elev_1k_above, elev_1k_below
Count of points (out of 24) within 50m of the same elevation	elev_5k_at_elev
Count of points (out of 24) more than 50m uphill or downhill of the location	elev_5k_above, elev_1k_below

A small number of other statistics were calculated regarding the points surrounding the locations. These statistics are not anticipated to be useful for general modeling applications and will not be distributed as part of the set of standard covariates. Users interested in creating a more sophisticated model for the effect of elevation should contact the EAC for more details.

1.11.19 Urban Topography

Building footprints, heights and parcel information were obtained for Chicago and New York City, from the respective city governments (see Table 11 for data source details). Statistics were calculated in an effort to characterize the degree to which each location in these two cities is situated in a “Street Canyon” (i.e., surrounded by buildings in a manner that significantly constrains air flow). If these data are missing for any specific location in NYC or Chicago, this indicates that no building was found within 60m of that location.

Generally speaking, individual urban topography variables do not directly quantify the pollution exposure exacerbation due to street canyons. They serve as building blocks for street canyon models currently under development and can be provided to analysts with a specific use for them. However, they are not intended to be distributed with the ‘standard set’ of modeling variables.

Table 19. Urban topography variable names

Variable Description	Variable Name
Distance in meters from the location's geocode to the closest building. That building is the reference for all subsequent calculations. If the geocode falls within a building's polygon, the distance is set at 1m. Maximum hit distance is 60m.	canyon_bldg_hitdist
Height of reference building in floors	canyon_bldg_hgt
Distance-weighted mean height of buildings on the same block on the same side of the street or of the buildings on the opposite side of the street (in floors). Open space is included as zero.	canyon_bldg_meanflrs, canyon_opp_meanflrs
The length of the block, measured as the distance in meters between the two road intersections defining it.	canyon_road_len
CFCC type of the road in the middle of the canyon, i.e. A2, A3, A4	canyon_road_type
Binary; 1 indicates that building is at the block's end	canyon_iscorner
Binary; 1 indicates that there is no opposite building	canyon_opp_missing

1.11.20 Census Data

Census data for the year 2000 were obtained from the US Census via the University of Washington Library. These demographic variables are included in the Census' Summary File 3 (SF 3). Data were compiled at the block (bk), block group (bg), and tract (tr) aggregating levels.

Table 20. 2000 Census variable names

Variable Description	Variable Name
Geographic Identifiers	
(Consolidated) Metropolitan Statistical Area Code, used by OMB for federal statistical agencies	cmsa
Race/Ethnicity (Self-Identified)	
% White alone	<level>_p_race_white**
% Black or African American alone	<level>_p_race_black
% American Indian and Alaska Native alone	<level>_p_race_native_amer
% Asian alone	<level>_p_race_asian
% Native Hawaiian and Other Pacific Islander alone	<level>_p_race_pacific
% Some other race alone	<level>_p_race_other
% Two or more Races	<level>_p_race_multi
% Hispanic	<level>_p_ethn_hisp
% Non-Hispanic	<level>_p_ethn_non_hisp
% Non-Hispanic Black	<level>_p_ethn_non_hisp_black
% Non-Hispanic White	<level>_p_ethn_non_hisp_white
Marital Status for the population 15+ (Self Identified)	
% Never married*	<level>_p_marital_never
% Now married*	<level>_p_marital_current
% Widowed*	<level>_p_marital_widow
% Divorced*	<level>_p_marital_divorce
Income	
Median Household Income (\$)*	<level>_med_inc_hshld
Mean Household Income (\$)*	<level>_mean_inc_hshld

Median Family Income (\$)*	<level>_med_inc_family
Mean Family Income (\$)*	<level>_mean_inc_family
% Households with income < \$15,000/year*	<level>_p_inc_hshld_le_15k
% Households with income >\$150,000/year*	<level>_p_inc_hshld_ge_15k
% of the population aged 5+ years who are below the poverty line*	<level>_p_poverty_individual
% of families who are below the poverty line*	<level>_p_poverty_family
% of households who are below the poverty line*	<level>_p_poverty_hshld
Employment	
% of the civilian population aged 16+ years that is unemployed*	<level>_p_unemp_all
% of the civilian male population aged 16+ years that is unemployed*	<level>_p_unemp_male
Occupation	
<i>% Employed civilian population 16+ in each of the following categories</i>	
Management, Professional, and related occupations*	<level>_p_emp_professional
Service occupations*	<level>_p_emp_service
Sales and office occupations*	<level>_p_emp_sales
Farming, fishing, and forestry occupations*	<level>_p_emp_farm_fish
Construction, extraction, and maintenance occupations*	<level>_p_emp_construction
Production, transportation, and material moving occupations*	<level>_p_emp_transport
Housing	
% Total housing units that are occupied	<level>_p_housing_occ
% Occupied housing units that are owner-occupied (vs. renter-occupied)	<level>_p_housing_owner_occ
% Occupied housing units with >1 person per room*	<level>_p_housing_crowded
Median value (\$) for specified owner-occupied housing units*	<level>_med_housing_value
% of the population that has lived at the same residence for 5+ years – tract only	tr_same_resi_5_yr
Household Size of Occupied Housing Units	
% of occupied housing units with a 1-person household	<level>_p_hshld_1
% of occupied housing units with a 2-person household	<level>_p_hshld_2
% of occupied housing units with a 3-person household	<level>_p_hshld_3
% of occupied housing units with a 4-person household	<level>_p_hshld_4
% of occupied housing units with a 5-person household	<level>_p_hshld_5
% of occupied housing units with a 6-person household	<level>_p_hshld_6
% of occupied housing units with a 7-or-more-person household	<level>_p_hshld_7plus
Urbanicity	
% of the population living in an urban area – tract only	tr_p_pop_urban
% of the population living in a rural area – tract only	tr_p_pop_rural
Education	
% Adults 25+ with < High School*	<level>_p_edu_less_hs
% Adults 25+ with ≥ High School*	<level>_p_edu_ge_hs
% Adults 25+ with High School diploma or equivalency*	<level>_p_edu_hs
% Adults 25+ with Some College, No Degree*	<level>_p_edu_some_college
% Adults 25+ with Associate's Degree*	<level>_p_edu_aa
% Adults 25+ with Bachelor's Degree*	<level>_p_edu_ba
% Adults 25+ with Master's Degree*	<level>_p_edu_ma

% Adults 25+ with Professional School Degree*	<level>_p_edu_professional
% Adults 25+ with Doctorate Degree*	<level>_p_edu_phd
* Not available at the block level	
** Where level is either block (bk), block group (bg), or tract (tr)	

Census data is also available for 2010. After the 2000 Census, the Census Bureau began administering the long-form questionnaire on a rolling (yearly) basis through the American Community Survey (ACS). The Census only releases ACS data for small areas in five-year aggregations, and so we provide ACS data from the period 2006 through 2010 for tracts and block groups. Block-level data are more restricted, and therefore short-form (SF1) 2010 data are provided. 2010 data was retrieved through the National Historical Geographic Information System⁶ (<http://www.nhgis.org/>) and from Social Explorer (<http://www.socialexplorer.com/>). All dollar figures were reported in 2010 inflation-adjusted dollars. More variables, years, and geographic areas are publicly available through these websites and through <http://www.census.gov/>.

Table 21. Year 2010 census variable names

Variable Description	Variable Name	Availability
Basic Demographics		
Total Population	<level>_tot_pop_2010*	All
Area (land only) in Square Kilometers	<level>_land_area_sqkm	All
% Male	<level>_p_sex_male	All
% Female	<level>_p_sex_female	All
% Under 5 years	<level>_p_age_0_5	All
% 5 to 9 years	<level>_p_age_5_9	All
% 10 to 14 years	<level>_p_age_10_14	All
% 15 to 17 years	<level>_p_age_15_17	All
% 18 to 24 years	<level>_p_age_18_24	All
% 25 to 34 years	<level>_p_age_25_34	All
% 35 to 44 years	<level>_p_age_35_44	All
% 45 to 54 years	<level>_p_age_45_54	All
% 55 to 64 years	<level>_p_age_55_64	All
% 65 to 74 years	<level>_p_age_65_74	All
% 75 to 84 years	<level>_p_age_75_84	All
% 85 years and over	<level>_p_age_85_plus	All
Race/Ethnicity		
% White alone	<level>_p_race_white	All
% Black or African-American alone	<level>_p_race_black	All
% American Indian and Alaska native alone	<level>_p_race_native_amer	All
% Asian alone	<level>_p_race_asian	All
% Native Hawaiian and other Pacific Islander alone	<level>_p_race_pacific	All

⁶ Minnesota Population Center. *National Historical Geographic Information System: Version 2.0*. Minneapolis, MN: University of Minnesota 2011.

% Some other race alone	<level>_p_race_other	All
% Two or more races	<level>_p_race_multi	All
% Hispanic	<level>_p_ethn_hisp	All
% Non-Hispanic	<level>_p_ethn_non_hisp	All
% Non-Hispanic: white alone	<level>_p_ethn_non_hisp_white	All
% Non-Hispanic: black or African-American alone	<level>_p_ethn_non_hisp_black	All
Native/Foreign Born		
% Native born	<level>_p_born_native	TR
% Foreign born	<level>_p_born_foreign	TR
% of Foreign born: naturalized citizen	<level>_p_born_foreign_naturalize d	TR
% of Foreign born: not a citizen	<level>_p_born_foreign_noncitizen	TR
Households and Families		
Households	<level>_households	All
Families	<level>_families	All
Housing units	<level>_housing	All
Housing units: occupied	<level>_housing_occ	All
Housing Characteristics		
% Occupied housing units: owner occupied	<level>_p_housing_occ_owner	All
% Occupied housing units: renter occupied	<level>_p_housing_occ_renter	All
% Occupied housing units: House heating fuel: gas (utility, bottled, tank, or LP gas)	<level>_p_housing_fuel_gas	TR, BG
% Occupied housing units: house heating fuel: electricity	<level>_p_housing_fuel_elec	TR, BG
% Occupied housing units: house heating fuel: fuel oil, kerosene, etc.	<level>_p_housing_fuel_oil	TR, BG
% Occupied housing units: house heating fuel: coal, coke or wood	<level>_p_housing_fuel_wood	TR, BG
% Housing units: 1 unit	<level>_p_housing_units_1	TR, BG
% Housing units: 2	<level>_p_housing_units_2	TR, BG
% Housing units: 3 or 9	<level>_p_housing_units_3_9	TR, BG
% Housing units: 10 to 49	<level>_p_housing_units_10_49	TR, BG
% Housing units: 50 or more	<level>_p_housing_units_50_plus	TR, BG
% Housing units: Mobile home	<level>_p_housing_units_mobile	TR, BG
% Housing units: Boat, RV, van, etc.	<level>_p_housing_units_other	TR, BG
Median year structure built	<level>_housing_med_yr_built	TR, BG
Owner-occupied housing units: median value	<level>_med_housing_value	TR, BG
Income		
Median household income (In 2010 Inflation adjusted dollars)	<level>_med_inc_hshld	TR, BG
Average household income (In 2010 Inflation adjusted dollars)	<level>_mean_inc_hshld	TR, BG
Median family income (In 2010 Inflation	<level>_med_inc_family	TR, BG

adjusted dollars)		
Average family income (In 2010 Inflation adjusted dollars)	<level>_mean_inc_family	TR, BG
% Households with Income Less than \$15,000	<level>_p_inc_hshld_le_15k	TR, BG
% Households with Income Less than \$25,000	<level>_p_inc_hshld_le_25k	TR, BG
% Households with Income Less than \$50,000	<level>_p_inc_hshld_le_50k	TR, BG
% Households with Income Less than \$100,000	<level>_p_inc_hshld_le_100k	TR, BG
% Households with Income Less than \$150,000	<level>_p_inc_hshld_le_150k	TR, BG
% Households with Income Less than \$200,000	<level>_p_inc_hshld_le_200k	TR, BG
Marital Status		
% Population 15 years and over: never married	<level>_p_marital_never	TR, BG
% Population 15 years and over: now married (not including separated)	<level>_p_marital_current	TR, BG
% Population 15 years and over: separated	<level>_p_marital_sep	TR, BG
% Population 15 years and over: widowed	<level>_p_marital_widow	TR, BG
% Population 15 years and over: divorced	<level>_p_marital_divorce	TR, BG
Employment		
% Population 16 years and over: in labor force	<level>_p_labor_force	TR
% Civilian population 16 years and over in labor force: unemployed	<level>_p_unemp_all	TR
% Civilian male population 16 years and over in labor force: unemployed	<level>_p_unemp_male	TR
% Employed civilian population 16 years and over: agriculture, forestry, fishing and hunting, and mining	<level>_p_emp_farm_fish	TR
% Employed civilian population 16 years and over: construction	<level>_p_emp_construction	TR
% Employed civilian population 16 years and over: manufacturing	<level>_p_emp_manufacturing	TR
% Employed civilian population 16 years and over: wholesale trade	<level>_p_emp_trade_whole	TR
% Employed civilian population 16 years and over: retail trade	<level>_p_emp_trade_retail	TR
% Employed civilian population 16 years and over: transportation and warehousing, and utilities	<level>_p_emp_transportation	TR
% Employed civilian population 16 years and over: information	<level>_p_emp_information	TR
% Employed civilian population 16 years and over: finance and insurance, and real estate and rental and leasing	<level>_p_emp_financial	TR
% Employed civilian population 16 years and over: professional, scientific, and management, and administrative and waste management services	<level>_p_emp_pro	TR

% Employed civilian population 16 years and over: educational services, and health care and social assistance	<level>_p_emp_edu_health	TR
Employed civilian population 16 years and over: arts, entertainment, and recreation, and accommodation and food services	<level>_p_emp_food_leisure	TR
% Employed civilian population 16 years and over: public administration	<level>_p_emp_public	TR
% Employed civilian population 16 years and over: Other services, except public administration	<level>_p_emp_other	TR
Education		
% Civilian population 16 to 19 years: not high school graduate, not enrolled (dropped out)	<level>_p_edu_teen_dropout	TR, BG
Population 25 years and over: less than high school	<level>_p_edu_less_hs	TR, BG
Population 25 years and over: high school graduate (includes equivalency)	<level>_p_edu_ge_hs	TR, BG
Population 25 years and over: some college	<level>_p_edu_some_college	TR, BG
Population 25 years and over: bachelor's degree	<level>_p_edu_bs	TR, BG
Population 25 years and over: master's degree	<level>_p_edu_ma	TR, BG
Population 25 years and over: professional school degree	<level>_p_edu_professional	TR, BG
Population 25 years and over: doctorate degree	<level>_p_edu_phd	TR, BG
Poverty Status		
Population for whom poverty status is determined: Under 1.00	<level>_p_poverty_income_ratio_0_99	TR, BG
Population for whom poverty status is determined: 1.00 to 1.99	<level>_p_poverty_income_ratio_100_199	TR, BG
Population for whom poverty status is determined: 2.00 and over	<level>_p_poverty_income_ratio_200_plus	TR, BG
Families with income below poverty level	<level>_p_poverty_families	TR, BG
Urbanicity		
Urban/rural designation	urban_rural	BK

* Where level (see last column) is either block (bk), block group (bg), or tract (tr)

1.11.21 Distance to Bus Route

Bus route data was obtained through contact with the New York Department of Transportation and included bus route information for the New York City area including routes in New York, New Jersey, and Connecticut. The distance to the nearest bus route in meters is provided (m_to_bus). Bus route lengths in buffers were also summed (bus_s<radius>), in the same manner as road lengths.

Bus routes for other areas of the US will be added in Rev 12. These data were downloaded from GTFS Feeds (<http://transitfeeds.com/>) in October 2016. GIS software was used to reproject into state plane zones, dedup links, and remove unwanted transportation types (ferry, light rail).

1.11.22 Satellite Data: PM2.5, NO2, SO2, HCHO, CO

Satellite-based estimates of air pollution for PM2.5, NO2, SO2, CO, and formaldehyde (HCHO) were obtained.

Annual satellite-based estimates of ground-level PM2.5 (1998-2014) at 0.1° were obtained from a previously published, publicly available data set (van Donkelaar et al., 2016). Briefly, five aerosol optical depth (AOD) satellite retrievals were combined with (1) satellite-based measurements of vertical aerosol profiles, (2) modeled AOD and ground-level PM2.5 from a global chemical transport model (GEOS-Chem), and (3) ground-based AOD measurements from the aerosol robotic network (AERONET) to estimate annual ground-level PM2.5 on a 0.1° grid (van Donkelaar et al., 2016).

We obtained daily L2 surface-level CO from the Measurements of Pollution in The Troposphere (MOPITT) sensor on NASA’s Terra satellite for years 2001-2016 (Deeter et al., 2017). Tropospheric NO2, SO2, and HCHO were derived from daily measurements obtained by the Ozone Monitoring Instrument (OMI) onboard the NASA Earth Observing System (EOS)-Aura satellite. Measurements were screened for quality based on cloud cover, illumination, and data flags and averaged temporally following a standard method for satellite data that considers pixel values within a buffer. The resolution of the final, processed rasters is provided in the table below. More details are available from Matt Bechle.

The value of the covariate in each case is the value of the pixel/grid cell in which the location lies.

Table 22. Satellite data methods

Pollutant	Years	Resolution	Instrument	Surface/Column
PM _{2.5}	1998 – 2016	0.1°	Multiple instruments	Surface
NO ₂ ^a	2005 – 2016	0.1°	OMI	Column
SO ₂	2005 – 2016	0.25°	OMI	Column
HCHO ^b	2005 – 2016	0.1°	OMI	Column
CO	2001 – 2016	0.25°	MOPITT	Surface

^aboth 1-year and 3-year averages calculated; ^blong term (12-year) average only

1.11.23 Method of Covariate (Variable) Calculation

Calculations were performed using the PostGIS spatial extension to PostgreSQL. Relevant scripts are located in Q:\neogeo.

All distance calculations are truncated at 25 kilometers, except for distance to A2 or A3 which were truncated at 20 kilometers. For example, the distance to the nearest coastline will be 25,000 meters for all locations in the Winston-Salem study area, as well as those locations in the Los Angeles study area that are greater than 25 kilometers from the coast. Data users should be aware that this will be true whether the location is 25.1 kilometers from the coast or 100 kilometers from the coast.

1.11.24 Data Quality

The EAC does not perform QC on the source shapefiles that were downloaded from third-party data providers. Data quality as reported by the data generating agency is disclosed in the metadata document and discussed in the MESA Air Quality Assurance Project Plan (see Appendix C).

1.12. Exposure Model Outputs

Modeled pollutant estimates may lag behind location data corrections by one database revision. Therefore, locations may not be assigned the estimate produced for an old location that lies more than 10m from the updated location.

For pollutants modeled at the two-week time resolution, an annual average will not be distributed for a participant that has fewer than 25 two-week predictions available for the residences at which they lived during the time period of interest. This could happen if, for example, a participant lived outside the MESA Air modeling areas during half of the year prior to their exam.

Table 23. Availability of predictions in MESA Air and SPIROMICS modeling regions. For participant locations, predictions are generated for all addresses geocoded to intersections or exact locations. Models are maximum-likelihood based, unless otherwise indicated. Cells indicate the database version providing the underlying covariate data.

Pollutant	Dates Available and Temporal Resolution	MESA Air	Regional Grids	Fine-Scale Grids	SPIROMICS Air
PM _{2.5} (µg/m ³)	01/1999 – 12/2013, 01/1999 – 12/2016 (SPIROMICS only) 2-week resolution ^c	Rev 8	Rev 8	Rev 8	Rev 11
NO _x (ppb)	01/1999 – 12/2013, 01/1999 – 12/2016 (SPIROMICS only) 2-week resolution ^c	Rev 8	Rev 8	Rev 8	Rev 11
NO ₂ (ppb)	01/1999 – 12/2013, 01/1999 – 12/2016 (SPIROMICS only) 2-week resolution ^c	Rev 8	Rev 8	Rev 8	Rev 11
LAC (10 ⁻⁵ m ⁻¹)	2005 – 2009 at 2-wk resolution and spatial-only representing 2006 – 2008	Rev 8	Rev 8	Rev 8	----
O ₃ (ppb)	01/1999 – 12/2013, 01/1999 – 12/2016 (SPIROMICS only) ^e 2-week resolution	Rev 9	Rev 9	Rev 9	Rev 11
LUR-based Coarse PM _{10-2.5} (µg/m ³), Cu (ng/m ³), P (ng/m ³), Si (µg/m ³), Zn (ng/m ³)	Spatial only (meant to represent the 2009 annual average)	Rev 3, available only for participants in Chicago, Winston-Salem and St. Paul ^b	Rev 3	----	----
Individual-level PM _{2.5} exposure (integrating infiltration and time-location) (µg/m ³)	Specific to residence at time of Air Questionnaire	Rev 5 ^d	----	----	----
Pragmatic PM _{2.5} (AKA “Interim Pragmatic”) ^a (µg/m ³)	1999 – 2006 at 2-wk resolution; available only for baseline addresses	Prior to development of the EAC database	----	----	----
Pragmatic PM _{2.5} (AKA “Final Pragmatic”) ^a (µg/m ³)	1999 – 2009 at 2-wk resolution	Rev 0	----	----	----

^a Not recommended for current use; superseded by more recent models.

^b MESA Coarse modeling regions differ from standard MESA Air regions. MESA Coarse predictions are available within 25 km of monitors.

^c MESA ST model predictions from 3/2012 – 12/2013 are based on extrapolation of temporal trends based on AQS data released after model development. These extensions were new in Rev 9.

^d This Rev is the time that the infiltration questionnaire table was created. Final exposures integrate the current version of ST model predictions.

^e Ozone predictions from the SPIROMICS models start in 1/2002 for Winston-Salem, 1/2006 for Salt Lake City, and 4/2009 for Ann Arbor.

Table 24. Availability of national model predictions. All national models are available at annual average resolution only, and only addresses within the continental US are included for any cohorts at this stage. Cells indicate the database version providing the underlying covariate data.

Pollutant	Years Available (all annual averages)	MESA Air	Sister, Two Sister	CHS/Other	WHI	National (Large-Scale) Grid
PM _{2.5} (µg/m ³)	1999 – 2015	Rev 11	Rev 11	----	Rev 11	Rev 11
PM _{2.5} (µg/m ³) (historical)	1980 – 2010	Rev 9	----	----	Rev 10	Rev 9
PM ₁₀ (µg/m ³)	1990 – 2014	Rev 11	Rev 11	Rev 11	Rev 11	Rev 11
PM _{2.5} species (µg/m ³): As, Cu, EC, Ni, OC, S, Si, SO ₄ ²⁻ , SO ₂ , NO ₃ ⁻ , V, Cr ^a	07/2009 – 06/2010	Rev 5	----	----	Rev 10	----
Satellite NO ₂ (Includes Satellite-based NO ₂) (ppb)	1990 – 2014	Rev 11	Rev 11	Rev 11	Rev 11	Rev 11

^aChromium is inconsistent in two ways: First, this pollutant was modeled using Rev 4 data, and second, Cr predictions are not currently available for MESA Air participants.

1.12.1 NO_x, NO₂, and PM_{2.5} Likelihood Model Predictions

The NO_x, NO₂, and PM_{2.5} predictions use a spatio-temporal modeling methodology originally developed by Paul D. Sampson. It was first described in Fuentes et al.⁷, and adapted to MESA Air's data systems by Sampson et al.⁸. The model is optimized via maximum likelihood, as developed and described by Szpiro et al.⁹ Both monitoring data and geographical variables were extracted from MESA Air Exposure Assessment Center's database. All models are based on Version 7 (July 2013) data for predictions through March 2013. We additionally provide predictions through 12/2013 using AQS data released after model development. To do this, we extended the spatiotemporal trends using constrained penalized splines. This approach ensures that predictions from previous database versions are unaffected by these model extensions. However, we note that the extrapolated predictions are much less variable than those available before 03/2012 and suggest that analysts exercise caution when using these extrapolations. Please contact the Data Team for further documentation.

It is worth noting that, in addition to the data described in this document, and provided as part of the MESA Air Data Request System (V7), models for pollutants NYC and Rockland include NYCCAS monitoring data. NYCCAS collected NO, NO₂, and PM_{2.5} in New York City over a period of two years between December 2008 and December 2010. Samples were collected at 150 sites for 7 to 8 two-week periods (one per season per year) over the two year period. Five reference locations, one in each NYC borough, collected two-week samples for the entire two-year period¹⁰.

The open-source R statistical analysis language was used, with the core model-fitting functions taken from the "SpatioTemporal" R package¹¹, authored by Lindstrom et. al.¹² and maintained by Lindstrom. The package is available on CRAN. Predictions were generated at all MESA Air participant addresses that were valid from 1999-2012 (last prediction available is 3/28/2012), were within a modeling area, and were geocoded to an 'exact', 'block', or 'intersection' location. See appendix for modeling area and model performance statistics.

1.12.2 O₃ Likelihood Model Predictions

The O₃ models were developed by Meng Wang using equivalent methods to those used in the previous section. This model was based on Version 8 data for predictions through December 2013.

Again, the open-source R statistical analysis language was used, with the core model-fitting functions taken from the "SpatioTemporal" R package¹³, authored by Lindstrom et. al.¹⁴ and maintained by

⁷ Fuentes, Monteserrat and Guttorp, Peter and Sampson, Paul D., 2006. Using transforms to analyze space-time processes. In: Finkenstadt, B. and Held, L. and Isham, V., *Statistical Methods for Spatio-Temporal Systems*, CRC/Chapman and Hall, pp. 77-150.

⁸ Paul D. Sampson, Adam A. Szpiro, Lianne Sheppard, Johan Lindström, and Joel D. Kaufman, "Pragmatic Estimation of a Spatio-Temporal Air Quality Model with Irregular Monitoring Data" (November 30, 2009). *UW Biostatistics Working Paper Series*. Working Paper 353. Available at: <http://www.bepress.com/uwbiostat/paper353>.

⁹ Szpiro, Adam A., Sampson, Paul D., Sheppard, Lianne, Lumley, Thomas, Adar, Sara and Kaufman, Joel, 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics* 21, 606-631.

¹⁰ The New York City Department of Health and Mental Hygiene, Queens College Center for the Biology of Natural Systems, and Zev Ross Spatial Analysis.

¹¹ The NO₂ model was developed with version 0.6.0 of the package, and updated NO_x with version 0.9.2

¹² Johan Lindstrom, Adam Szpiro, Paul D. Sampson, Silas Bergen and Assaf P. Oron (2012). SpatioTemporal: Spatio-Temporal Model Estimation. R package version 0.9.2. <http://CRAN.R-project.org/package=SpatioTemporal>

¹³ The NO₂ model was developed with version 0.6.0 of the package, and updated NO_x with version 0.9.2

¹⁴ Johan Lindstrom, Adam Szpiro, Paul D. Sampson, Silas Bergen and Assaf P. Oron (2012). SpatioTemporal: Spatio-Temporal Model Estimation. R package version 0.9.2. <http://CRAN.R-project.org/package=SpatioTemporal>

Lindstrom. The package is available on CRAN. Predictions were generated at all MESA Air participant addresses that were valid from 1999-2012 (last prediction available is 1/1/2014), were within a modeling area, and were geocoded to an 'exact', 'block', or 'intersection' location. See appendix for modeling area and model performance statistics.

1.12.3 SPIROMICS-Specific Likelihood Model Predictions

Predictions were obtained using methods similar to those utilized in MESA-Air (described briefly in Section 1.12.1 above) following the operational details of the hierarchical spatiotemporal model in Keller et al. Monitoring data and geographical covariates were extracted from MESA Air Exposure Assessment Center's database for seven SPIROMICS cities (Baltimore, MD, New York, NY, Winston-Salem, NC, Los Angeles, CA, San Francisco, CA, Salt Lake City, UT, and Ann Arbor, MI, the first four of which were also MESA cities). After covariate pre-processing based on the recommendations in Keller et al., core model fitting was performed using SpatioTemporal, an R package available on CRAN. Predictions are based on a draft of Rev 11 of the data (DR0268, filled 5/22/2017), and cover the period 01/01/1999 – 12/28/2016 at SPIROMICS Air participant addresses. See the appendix for a summary of source monitors, modeling areas and model performance.

1.12.4 Light Absorption Coefficient (LAC) Predictions

LAC predictions are available at MESA participants' residential locations. A spatio-temporal model for LAC was developed based on MESA Air monitoring data collected between July 2005 and July 2009, NYCCAS data from 2008 – 2010 for NYC and Rockland, and geocovariates from Rev 7 of the database. The model was fit using the SpatioTemporal R package, and predictions were averaged over the timeframe of MESA Air monitoring (2006-2008). The spatial averages from the model are intended to represent long term exposures to black carbon. See appendix for model performance statistics.

1.12.5 Pragmatic PM_{2.5} Predictions

The pragmatic PM_{2.5} predictions were calculated using R and the methods described in "Pragmatic Estimation of a Spatio-Temporal Air Quality Model with Irregular Monitoring Data"¹⁵. These are available as two-week integrated predictions for ambient PM_{2.5} concentrations. "Initial" pragmatic predictions were generated for the "at exam" addresses for MESA Classic participants at Exam 1 and MESA Air Family and New Recruits for their baseline exam (Exam 3 or 4). "Final" pragmatic predictions were generated at all MESA Air participant addresses that were valid from 1999-2009, were within a modeling area, and were geocoded to an 'exact' or 'intersection' location. Besides encompassing more addresses and a longer timeframe, the road network used was changed from the Census source to the TeleAtlas source. Additional covariates, such as NDVI, and larger buffers of existing covariates, were available for the partial least squares regression. Site-varying neighborhood radii were used in the local random effects modeling strategy to compute trend functions. Finally, the Hawkins and Cressie variogram estimator was used for space-time residuals. See appendix or citation for model performance statistics.

¹⁵ Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45(36), 6593-6606.

Keller JP, Olives C, Kim SY, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environ Health Perspect* 2015; 123: 301-309.

1.12.6 National Model PM_{2.5}, PM₁₀, and NO₂ Predictions

The national models for PM_{2.5}, PM₁₀ and NO₂ are universal kriging models with partial least squares used to select relevant components for the mean regression. In Rev 9, we additionally include an alternative national NO₂ national model that includes as an independent predictor variable columnar satellite NO₂ measurements. This model showed improved performance compared to the model without satellite data and is the recommended exposure. The nation was divided into three regions based on topography. Each model was fit as a whole using maximum likelihood, with each region having its own set of estimated parameters for each pollutant and year. That is, for each year of annual average monitoring data, models were separately estimated (meaning that for each year PLS components, PLS coefficients, and variograms were estimated separately). Predictions were generated for all participant residence locations. The method was developed by Paul Sampson et. al. and implemented by Michael Young and Josh Keller¹⁶. PM_{2.5} (1999-2011) is based on Version 5 (August 2012) data, PM₁₀ (1990-2010) is based on Version 6 (February 2013) data, NO₂ (1990-2012) is based on Version 7 (August 2013), and satellite-enhanced NO₂ (1990-2012) is based on Version 8 data. Models through 2014 or 2015 were added in Rev 11. See appendix for model performance statistics.

1.12.7 National Model Historical PM_{2.5} Predictions

The model for historical PM_{2.5} Predictions relies on the MESA Air spatio-temporal modeling framework applied to annual average concentrations. See section 2.12.1 for a summary of these methods. Models were built using Version 6 covariates (DR0110). Temporal trends were back-casted using line extrapolation of trends estimated from 1999-2012 data. Historical prediction models were developed by Sun-Young Kim.¹⁷ See appendix for model performance statistics.

1.12.8 National Model for As, Cr, Cu, EC, Ni, NO₃⁻, OC, S, Si, SO₂, SO₄²⁻ and V, Predictions

Methods used for the national PM_{2.5} model (excluding the regional interactions) were applied to As, EC, OC, Si, S, Cr, Cu, Ni, V, SO₂, nitrate, and sulfate data covering the one year period between July 2009 and June 2010. This period was chosen to maximize the number of whole-year averages, which was necessary to accommodate analysis method changes that were implemented at STN monitors. Covariates and AQS data came from Rev 5 (DR0096) for all pollutants except Cr (year 2009, covariates from Rev 4, DR0076). Predictions were generated for WHI-OS locations by Silas Bergen. See appendix for model performance statistics.

1.12.9 Coarse PM Land Use Regression Predictions

Model predictions for the MESA Coarse PM Study were developed for PM_{10-2.5}, Cu, Si, P, and Zn using monitoring data collected at Coarse Snapshot locations (see section 2.3.1) and covariates from Rev 2 (DR0056). Land use regression models were selected using a separate exhaustive search for each study area and pollutant for the model with the lowest RMSE. Variable selection for the candidate models is described in Zhang et. al. (submitted). Briefly, models were designed to accommodate monitoring data from two rounds of sampling with interaction terms for season. The variable list was pared down from the available set described in section 2.10 to include those covered adequately by both the monitored locations and participant locations, then further reduced to 15 using LASSO. Kriging was not supported

¹⁶ Paul David Sampson, Mark Richards, Adam A Szpiro, Silas Bergen, Lianne Sheppard, Timothy V Larson, Joel D Kaufman, 2013. A Regionalized National Universal Kriging Model using Partial Least Squares Regression for Estimating Annual PM_{2.5} Concentrations in Epidemiology. *Atmospheric Environment* 75, 383-392.

¹⁷ Kim SY, Olives C, Sheppard L, Sampson PD, Larson TV, Keller JP, Kaufman JD. Historical prediction modeling approach for estimating long-term concentrations of PM_{2.5} in cohort studies before the 1999 implementation of widespread monitoring. *Environmental health perspectives*. 2017 Jan;125(1):38.

by the monitoring data and was not used. Predictions for locations lying more than 25 km from a monitoring location or with covariate values lying outside the range observed in the monitoring locations (with a 10% tolerance) were flagged. By default, these predictions are excluded from datasets provided to analysts. Note that this means that ‘missing’ predictions will vary by pollutant. See appendix for model performance statistics.

1.12.10 Alternative National NO₂

NO₂ predictions are available nationally at census block centroid for the national NO₂ model developed by Novotny et. al.¹⁸ Briefly, this is a land use regression model that depends primarily on satellite data for column NO₂ and impervious surface.

1.12.11 Individual-level Exposures to Ambient PM_{2.5}

Ambient source exposures incorporate the likelihood-based PM_{2.5} predictions, predicted infiltration fractions, and time-location questionnaire data. The fraction of PM_{2.5} that infiltrates indoors (F_{inf}) is impacted by temperature (as temperature affects behavior such as window opening and heat and air conditioning use). The temperature used when calculating F_{inf} is based on the two-week average temperature observed at a central monitor in each study area. A ‘warm season’ model is applied when the average temperature exceed 18 degrees Celsius, and the “summer” time-location patterns are used to calculate the percent of time spent indoors during a typical week. The infiltration model was developed by Ryan Allen at Simon Fraser University¹⁹. This is a regression model based on MESA Air monitoring data and Air Questionnaire information. Each two-week PM_{2.5} prediction is multiplied by the infiltration fraction and the percent of time spent indoors and this is added to the PM_{2.5} prediction multiplied by the percent of time spent outdoors. These two-week individual-level exposures can be aggregated up to any time scale desired.

1.13. Meteorological Data

Meteorological (“met”) data (temperature, humidity, wind speed, wind direction, ceiling height, dew point, sea level pressure, station level pressure, and visibility) are downloaded from the National Oceanic and Atmospheric Administration’s (NOAA) National Climatic Data Center (NCDC) at <http://www7.ncdc.noaa.gov/CDO/dataproduct>, as are the locations of the weather stations. Meteorological data are available nationally for the years 1979 through December 2017. Weather stations are identified by their Weather Bureau Army Navy (WBAN) and Cooperative Station (COOP) identification numbers. Meteorological data can be provided from the nearest met monitor or from a single station in each MESA Air community (including Riverside, CA; Rockland County, NY; and the metropolitan areas of the six cities). Because the data outside the MESA Air study areas has not been thoroughly reviewed for completeness and data quality issues, we recommend that “nearest met monitor” data be used with caution. For health analysis requests, we will provide the “community met monitor” data as the default.

These data are available at daily time resolution. For a subset of timepoints, data are reported at multiple altitudes at the same location. These data are all considered “surface” level data by NOAA. Whenever multiple recordings from the same location are available during the same hour, recordings are ranked by altitude at which they were recorded, and a single datapoint is selected and maintained in

¹⁸ Novotny EV, MJ Bechle, DB Millet, and JD Marshall. 2011. "National satellite-based land-use regression: NO₂ in the United States". *Environmental Science & Technology*. 45 (10): 4407-14.

¹⁹ Allen RW, Adar SD, Avol E, Cohen M, Curl CL, et al. 2012 Modeling the Residential Infiltration of Outdoor PM_{2.5} in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environ Health Perspect* doi:10.1289/ehp.1104447

the EAC database. The “ground level” recording is retained if available. The next choice is a “sea level” recording, which is often a conversion of the ground level recording to sea level conditions. The last choice would be a recording taken at an unknown elevation, reported for aviation, but still considered “surface level”. Even with these rules, users should be aware that coverage in time is not perfect. No further cleaning has been performed, and no QA/QC rules have been set for meteorological data. Data are formatted and uploaded to the EAC database ‘as-is’.

Temperature and dewpoint are reported in degrees Celsius. Relative humidity is reported as a percentage. Wind speed is reported in m/s. Wind direction is reported as an angle, with zero degrees representing a wind blowing from due north. The daily prevailing wind was calculated using vector averaging, with calm wind hours considered to be the 0 vector (no direction reported, speed reported as 0-1.5 m/s). Vectors with no direction and wind speeds above 1.5 m/s were considered missing data and dropped from the averaging. Visibility and ceiling height are reported in meters. Pressure is reported in hectopascals. Daily values for all meteorological components met daily completeness rules outlined in section 2.2.4.

Visibility was truncated at 10 miles due to a method change at most sites by the year 2001. Noon visibility is reported alongside noon relative humidity and noon weather conditions. Weather conditions are reported as a string of binary indicators for the presence of airborne dust, fog, and precipitation.

1.13.1 B-Spline Variables for Temperature, Humidity, and Calendar Time

For studies of the chronic effects of air pollution on outcomes which may show seasonal variability, an alternative to adjusting directly for temperature and humidity is to adjust for b-spline (basis spline) variables. Users should be aware that each spline generates a large number of variables. However, this allows for very fine and flexible control of seasonal variability. We typically distribute these as variables from b-splines with 6 degrees of freedom per year (i.e. 6 variables per year).

2. Exposure Assessment Core Data Request System

2.1 Placement of a Request

Data requestors are strongly encouraged to complete a Statistical Analysis Plan before requesting data from the EAC. A template for a Statistical Analysis Plan is included in Appendix A. Once a data requestor has a clear idea of the data that he or she will need, that person should open an electronic Exposure Assessment Core Data Request.

There are two possible forms that a requester can use. If an analyst is conducting an epidemiological analysis, in which they will be analyzing the relationship between health outcome data and exposure estimates (either exposure model outputs, or estimates based on EPA's AQS data, or "distance-to-roadway" metrics), they should use this form: <https://catalyst.uw.edu/webq/survey/agassett/236007>

If they are either developing an exposure model or doing any analyses that would require the use of raw monitoring or geographical data, then this form is more appropriate: <https://catalyst.uw.edu/webq/survey/agassett/236008>

Instructions are provided within the form. Please contact Amanda Gassett at agassett@uw.edu with questions or feedback, or if you are uncertain of which form to use. Analysts that are not certain what data to request are encouraged to consult with a member of the EAC data team.

Data requests will be tracked by the project manager and within the database. The Q:\ drive will contain a directory for data requests, and a subdirectory for each request with the output from the original electronic request and the queries used to generate the datasets.

2.2 Maps of Participant-Identifying Information

In some cases, data analysts may desire to include a map of participant locations or exposure modeling predictions in their papers or presentations. It is important to remember that participants' residential locations are considered participant-identifying information and are protected in accordance with the DMDA. Any maps that will be displayed publically or viewed by anyone (including the analyst) who has not read and signed the DMDA must have participant locations jittered so that the actual residential locations are not identified. Other data may be included on the maps as desired. The outline of the desired map can be described in the final text box in the electronic data request form.

2.3 Multiple Users of Requested Data, or Accessing an Existing Request

In some cases, multiple analysts will be working with the same data and the same datasets. Data users are discouraged from sharing data amongst themselves. Data security issues are raised when participant identifying information is transmitted between users, and the EAC prefers to control the transmission of data for the protection of participant data. Furthermore, all data issued by the EAC is subject to caveats, corrections, and updates. The EAC can only ensure that analysts receive the appropriate communications related to the datasets that they are using if the EAC can track which analysts are using which datasets for which projects. Data users are encouraged to contact the EAC if they wish to use existing datasets. In most cases, these requests can be filled with very little delay.

2.4 Fulfillment of a Request

The electronic data request tool generates a report of the request and transmits it to the project manager. As soon as the project manager reads the request, she will send the data requestor a confirmation email with a summary of the data that the EAC data team expects to provide. The original request will be assigned an identification number and attached as a pdf. All members of the EAC data team will receive a carbon copy of this message. The project manager will bring the request to the EAC data team meeting. The details of the request are discussed in the meeting, and the project manager will email the requestor with any clarifying questions.

Data requestors will receive a notification from the QA Officer when the request is complete. The notification will contain a summary of the data that was requested, a brief description of each file produced that includes the number of records that the files contain, and any caveats or additional information that the data team deems necessary. The data requestor should retrieve and review the datasets promptly, and notify the data team of any problems. If no problems are noted, the data requestor should send email confirmation to the project manager or QA Officer that the request can be closed.

Requests will not be updated automatically when new database versions are released, but users will be notified when a new version of the data is available. Data users whose analyses are still in progress should contact the EAC for updates.

Table 25. Summary of data request file locations

Type of File	Type of Data User	File Location	Storage Timeframe
Final, contains participant locations	Final user, without access to MAIDS	Transmitted by email as a password-protected zipped file	Emails should be deleted as soon as possible
Final, no participant locations	Final user, without access to MAIDS	Transmitted by email as a zipped file	Emails should be deleted as soon as possible
Final, contains participant locations	Final user, with MAIDS access	M:/MAIDS/Monitoring Database/Active Data Request - sensitive /<request>/ <filenames>	7 days
Final, no participant locations	Final user, without MAIDS access	M:/MAIDS/Active Data Request - Non sensitive /<request>/ <filenames>	7 days
Interim	EAC data team member	Q:/eac_database/requests/<request>	Indefinite

2.4.1 Data Request Fulfillment: Work Flow Details for Internal Users

Roles for each data request are established in the EAC data team meeting. Once a request is confirmed, one member will be responsible for writing the SQL queries to extract, compile, and process the data into the requested format. A folder will be generated on Q:/ that will contain a copy of the original request, a copy of the SQL queries that generate the final dataset, and will be a temporary storage space for files related to the request. These folders should be generated from the bash shell using `“./make_request.sh <name> <description> <file format>”`.

The request will be filled using the bash shell script, "fill_request.sh". This script is located at /var/local/QUTE/eac_database/requests, and can be accessed from the bash shell with the command "./fill_request.sh <request_id_number>". This script runs the queries to generate datasets, creates a file of 'metadata' which contains the meanings of column names, and creates a record in the request table. This table, housed inside the database, logs the name of the requestor, the date the request was filled, and the query used to fill the request.

A different member of the EAC data team will check the requested datasets for completeness and correctness. Once the dataset generator notifies this person that the datasets are ready for the QA check, the QA auditor will perform their inspection and transfer the files to the appropriate location on the server. The QA auditor will notify the project manager and QA officer that the request is complete, and the QA officer will notify the data requestor. Once the data requestor expresses satisfaction with the data product, the request can be closed.

Appendices

A. Statistical Analysis Plan (SAP)

Analysis Plan

Working Title:

Overview/Purpose:

General Scientific Question(s):

Specific Scientific Question(s) (e.g. hypotheses):

Outcomes of Interest:

Predictors of Interest:

Potential Confounders or Adjustment Variables:

Other Data Specifics (e.g. time period, subgroup):

Data request (date, number):

Type of Analysis:	Hypothesis testing	Estimation
	Hypothesis screening	Modeling
	Hypothesis generating/exploratory	Method evaluation
	Descriptive	

Analysis Approach and Special Issues:

List of Tables: (or note location of draft tables)

Plan of Action:

Responsibilities and deadlines:

- Paper outline
- Initial analyses
- Introduction
- Methods
- Results
- Discussion
- Tables and Figures
- Follow up analyses
- Final Draft

Names and roles (authors, co-authors, Data Core staff):

Revision History:

B. List of Acronyms and Abbreviations

CAP – Criteria Air Pollutant	NOAA – National Oceanic and Atmospheric Administration
CC – Coordinating Center	NO ₂ – Nitrogen dioxide
C-CAR – Center for Clean Air Research	NO _x – Oxides of nitrogen, including NO ₂ and NO
COOP – Cooperative Station	NPACT – National Particle Components Toxicity
CRAN – Comprehensive R Archive Network	O ₃ – Ozone
DMDA – Data and Materials Distribution Agreement	PM – Particulate matter
EAC – Exposure Assessment Core	PM _{2.5} – Particulate matter <= 2.5 µg in diameter
EC/OC – Elemental carbon and organic carbon	PM ₁₀ – Particulate matter <= 10 µg in diameter
FIPS – Federal Information Processing Standards	POC – Parameter Occurrence Code
FRM – Federal Reference Method	ppm – parts per million
GIS – Geographic Information System	ppb – parts per billion
HAP – Hazardous Air Pollutant	QA – Quality Assurance
HEI – Health Effects Institute	QC – Quality Control
IMPROVE – Interagency Monitoring of Protected Visual Environments	S - Sulfur
LAC – Light Absorption Coefficient	SAP – Statistical Analysis Plan
LOD – Limit of Detection	Se - Selenium
MAIDS – MESA Air Intermediate Data Server	Si - Silicon
MESA – Multi-Ethnic Study of Atherosclerosis	SOP – Standard Operating Procedure
µg/m ³ – micrograms per cubic meter	SO ₂ – Sulfur dioxide
NAAQS – National Ambient Air Quality Standards	SPCS – State Plane Coordinate System
NEI – National Emissions Inventory	STN – Speciation Trends Network
NDVI – Normalized Difference Vegetation Index	VOC – Volatile Organic Compound
NIST – National Institute of Standards and Technology	XRF – X-ray fluorescence
NO – Nitrogen oxide	WBAN – Weather Bureau Army Navy
	WHI - OS – Women’s Health Initiative - Observational Study

C. Referenced Documents and Code Locations

Referenced Documents

Description	Directory	Filename
Quality Assurance Project Plan (MESA Air)	M:\MAIDS\QA\QAPP\	QAPP_040709.doc
Final QA/QC Report (MESA Air)	M:\MAIDS\QA\QA report for ESAC\	QAQC_Committee_Report_042310.doc
Metadata document	M:\MAIDS\Documentation\GIS\Meta data	Metadata_08042009.doc
Field SOPs	M:\MAIDS\QA\SOPs\Field\Final\ Current SOPs\PDFs	All documents in directory
Lab SOPs	M:\MAIDS\QA\SOPs\Labs\Final\PDFs	All documents in directory
Documentation of MESA Air Implementation of CALINE3QHR Model	CALINEDoc2.doc	Subversion repository, ExposureModel\CALINE\
Monitor Issue Log	M:\MAIDS\QA\AQS_QA	Monitor_Issue_Log.xls
QA/QC Report (SPIROMICS Air)	P:\QA\QA_report	QAQC_report_14July16.docx (in progress)
QA/QC Report (CCAR)	W:\UW CCAR\Project 1-5, Cores Folders\Project 5\QC\QAQC_report	
QMP (CCAR)	W:\UW CCAR\QA QC	Quality Management Plan
QA/QC Report (Coarse)		

Useful scripts and codefiles

Purpose	Filename	Language	Directory
Parses pump files for TSI Sidepack 530 pumps	get_pump_data.sas	SAS	H:\ My Documents\sas_code
Reads and processes HOBO files for indoor temperature and relative humidity	get_hobo_data.sas	SAS	H:\ My Documents\sas_code
Removes duplicate road segments from TeleAtlas shapefiles	cat_and_dedup_roads.py	Python	H:\GIS\pycode\mesa\
Calculates covariates for point locations	covar.py	Python	H:\GIS\pycode\mesa\
Parses AQS 'input transaction' formatted files for monitors	ReadAQS_Monitors.sas	SAS	M:\MAIDS\EPA monitoring\Programs
Parses AQS concentration files	ReadAQS_NO2.sas ReadAQS_Nox.sas ReadAQS_PM25.sas	SAS	Subversion repository, Data_Processing_Codes\MesaAQS_Raw directory
Read in and combine data tables in the NEI database	Read Sources.sas	SAS	M:\MAIDS\Emissions Data\Programs\
Parses and formats raw temperature and humidity data	ReadMet.sas	SAS	M:\MAIDS\MeteorologicalData\Programs
Nearest Monitor Calculation	nearest.R	R	Subversion repository, ExposureModel\ SimpleModels\
Updates metadata table	metadata_tbl.py	Python	Q:\eac_database\code
Creates request folder	make_request.sh	Bash	Q:\eac_database\requests
Generates data request datasets and metadata	fill_request.sh	Bash	Q:\eac_database\requests

Locations of raw monitoring data

Contents	Filename	Directory
Field log data	MESA_Samples_V1_2008_10_10.mdb	M:\MAIDS\COC
Teflon mass measurements	MESA_Air_FilterMassEntry_v1_2.mdb	M:\MAIDS\GravLACAnalysis\database
Teflon LAC measurements	MESA_Air_Reflectometry_v1_1.mdb	M:\MAIDS\GravLACAnalysis\database
Ion Chromatography measurements for Ogawas (NO ₂ , SO ₂ , and O ₃)	IC<YYYYMMDD_##>.xls	M:\MAIDS\AnalyticLab\Edited IC Data for Import Jim
Ultraviolet Spectroscopy (For NO _x)	UV<YYYYMMDD_##>.xls	M:\MAIDS\AnalyticLab \Edited UV Data for Import Jim
Infiltration questionnaire	InfiltrationObservaton_v_1_09_10_08.mdb	M:\MAIDS\Field\Databases
Time Activity Diary for personal sampling	TimeActivity_v_3.0.mdb	M:\MAIDS\Field\Databases
Chain of Custody (sample handling and shipping history)	MESA_Samples_V1_2008_10_10.mdb	M:\MAIDS\COC

D. Suggested Citations

Data Description	Suggested Citation
Airport and point source emission volumes ^a	USEPA, Emission Inventory Group, 2002 National Emissions Inventory Database [digital data set] (2006). U.S. Environmental Protection Agency: Washington, DC. Available FTP: http://www.epa.gov/ttn/chief/net/2002inventory.html [accessed Aug 2008].
Road, railroad, railyard, and airport locations, coastlines	TeleAtlas, TeleAtlas Dynamap 2000 [CD_ROM] (2000). TeleAtlas, Lebanon, NH.
Population density	U.S. Department of Commerce, Census Bureau, TIGER/Line Shapefiles (2001). U.S. Department of Commerce, Census Bureau: Washington, DC. Available FTP: http://arcdata.esri.com/data/tiger2000/tiger_download.cfm , redistributed by TeleAtlas Dynamap 2000 [CD_ROM] (2000). TeleAtlas, Lebanon, NH.
Land use	Price, C.V., Nakagaki, N., Hitt, K.J., and Clawges, R.C., Enhanced Historical Land-Use and Land-Cover Data Sets of the U.S. Geological Survey, U.S. Geological Survey Digital Data Series 240 [digital data set] (2006). Available: http://pubs.usgs.gov/ds/2006/240 [accessed Jun 2009].
Port locations and characteristics ^b	National Imagery and Mapping Agency (NIMA), Ports of the Wider Caribbean, from NIMA World Port Index (2002). Distributed by the World Resources Institute: Washington, DC. Available FTP: http://164.214.12.145/pubs/pubs_j_wpi_sections.html [accessed Jun 2009].
EPA AQS data	USEPA, Air Quality System Data: Query AQS Data [digital data set] (2011). U.S. Environmental Protection Agency: Washington, DC. Available: http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm [accessed June 2013].
Meteorological data	National Climatic Data Center, Hourly Global Surface Data [digital data set] (2010). National Climatic Data Center: Asheville, NC. Available: http://www7.ncdc.noaa.gov/CDO/dataproduct [accessed June 2013].
NDVI ^a	Carroll, M.L., C.M. DiMiceli, R.A. Sohlberg, and J.R.G. Townshend, 250m MODIS Normalized Difference Vegetation Index [digital data set] (2008). University of Maryland, College Park, Maryland. All available days, 2006. Available: http://glcf.umiaccs.umd.edu/data/ndvi/ .
NYCCAS Data	The New York City Department of Health and Mental Hygiene, Queens College Center for the Biology of Natural Systems, and Zev Ross Spatial Analysis.
U.S. Census Data	Census 2000 Summary File 3—United States/ prepared by the U.S. Census Bureau, 2002.
U.S. Truck Routes	National Transportation Atlas Database [digital data set] (2009). Bureau of Transportation Statistics: Washington, DC. Available: http://www.bts.gov/publications/national_transportation_atlas_database .
NY Bus Routes	Quodomine, R. 2013. Personal communication (email with E. Spalt, University of Washington). New York Department of Transportation, Albany, NY.
Impervious Surface	NLCD 2006 Percent Developed Imperviousness [digital data set] (2006). Multi-Resolution Land Characteristics Consortium: Sioux Falls, SD. Available: http://ims.cr.usgs.gov/webappcontent/mrlc/nlcd2006_downloads.php .

Suggested citations, continued

MESA Air Monitoring Method	Cohen, Martin A., Adar, Sara D., Allen, Ryan W., Avol, Edward, Curl, Cynthia L., Gould, Timothy, Hardie, David, Ho, Anne, Kinney, Patrick, Larson, Timothy V., Sampson, Paul, Sheppard, Lianne, Stukovsky, Karen D., Swan, Susan S., Liu, L.-J. Sally, Kaufman, Joel D., Approach to Estimating Participant Pollutant Exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environmental Science & Technology 2009, 43 (13), 4687-4693.
----------------------------	---

^a This link no longer works or this exact file is no longer available online. Please refer to table 10 for current links.

^b This file is no longer available online, and the National Imagery and Mapping Agency no longer exists. The current version (the Twentieth Edition) is produced by the National Geospatial Intelligence Agency.

E. Known Data Quality Issues

Issue	Extent of Data Affected	Expected Resolution
Overlapping polygons in land use shapefiles (>100% land use totals)	Locations along the coast of NC (primarily Sisters and WHI locations)	Manually fix shapefiles
Because the Oregon truck routes were determined by a different method than the rest of the US, truck route line lengths are doubled along A1 highways	Locations near A1 highways in Oregon (i.e. Sisters, WHI, AQS monitors)	Manually fix shapefiles
Distance to city hall calculations not correctly implemented in psql		

F. Quick Reference for Averaged Exposure Variable Names

Analysts may consider exposure periods ranging from a single day or week prior to an exam up to a number of years. Some may consider multiple exposure periods. Standard naming conventions help distinguish these different exposure periods quickly and easily. The set formats for these variable names are:

<pollutant>_<model>_<t1>_<t2>_<interval>_<reference>_<weighting>
 <pollutant>_<model>_<e1>_<e2>_<rounding>_<reference>_<weighting>
 <pollutant>_<model>_<yymm1>_<yymm2>_<weighting>
 <pollutant>_<model>_year_<year>_<weighting>
 MET_<meteorology>_<t1>_<t2>_<interval>_<reference>_<weighting>

A table of the options that appear between brackets is included below. Characters without brackets appear as literals.

Pollutants include all of the gases, elements, and PM sizes that have been mentioned throughout this document. Models include both very sophisticated spatio-temporal models as well as ‘simple’ models such as ‘nearest monitor’ for PM_{2.5}. The numbers t1 and t2 indicate the span of the averaging period, while ‘interval’ indicates the units of time to which t1 and t2 refer. Much of time, analysts are

interested in a particular event, such as an exam, a stroke, or a diagnosis of disease. This is the reference point. Finally, some analysts will need the exposure averaged or weighted by the locations where participants actually lived. Others will need the average at a particular location over the entire time period, regardless of whether the participant lived at that location during the entire time period. Meteorological measurements can be averaged over similar time frames and will have similar names to indicate the averaging period. However, 'weighting' will not be indicated, as there is only one meteorological time series in each study area.

Table of parameter options

Pattern Code	Description	Values	Meaning
pollutant	Pollutant of interest	PM25, PM10, Si, S, EC, OC, NOx, NO2, O3, etc.	
model	Pollution model type	prag lik nat acute near	pragmatic likelihood national (long term) acute regional time series nearest monitor
meteorology	Meteorological variable	temp rh windspd winddir	temperature humidity wind speed wind direction
t1, t2	Timepoints that define beginning and end of the exposure averaging period	Integers from 0 – 99 “a” followed by an integer 1-99	Intervals before reference Intervals after reference
e1, e2	Exam numbers	integer 1-5	Exam numbers between which average is calculated
rounding	Whether time between exams is rounded to nearest whole year	rnd trc N/A	Round (nearest whole) Truncate (round down) Exact time
interval	Units of time	dy wk qt yr	Day week quarter year
reference	Reference point, or event that defines what the exposure is prior to	enrol event exam ct us mri	enrollment event (MI, stroke, diagnosis) clinic exam
weighting	Indicates whether the average is weighted over all addresses at which participants lived during the exposure period	wght stat wsp statsp	weighted by all addresses static, usually refers to ‘at exam’ address weighted, spatial-only model static, spatial-only model
yymm1, yymm2	Fixed endpoints of exposure period	0001 – 1112	two-digit year and two-digit month (beginning of month in endpoint 1 through end of month in endpoint 2)

G. MESA Air Modeling Regions

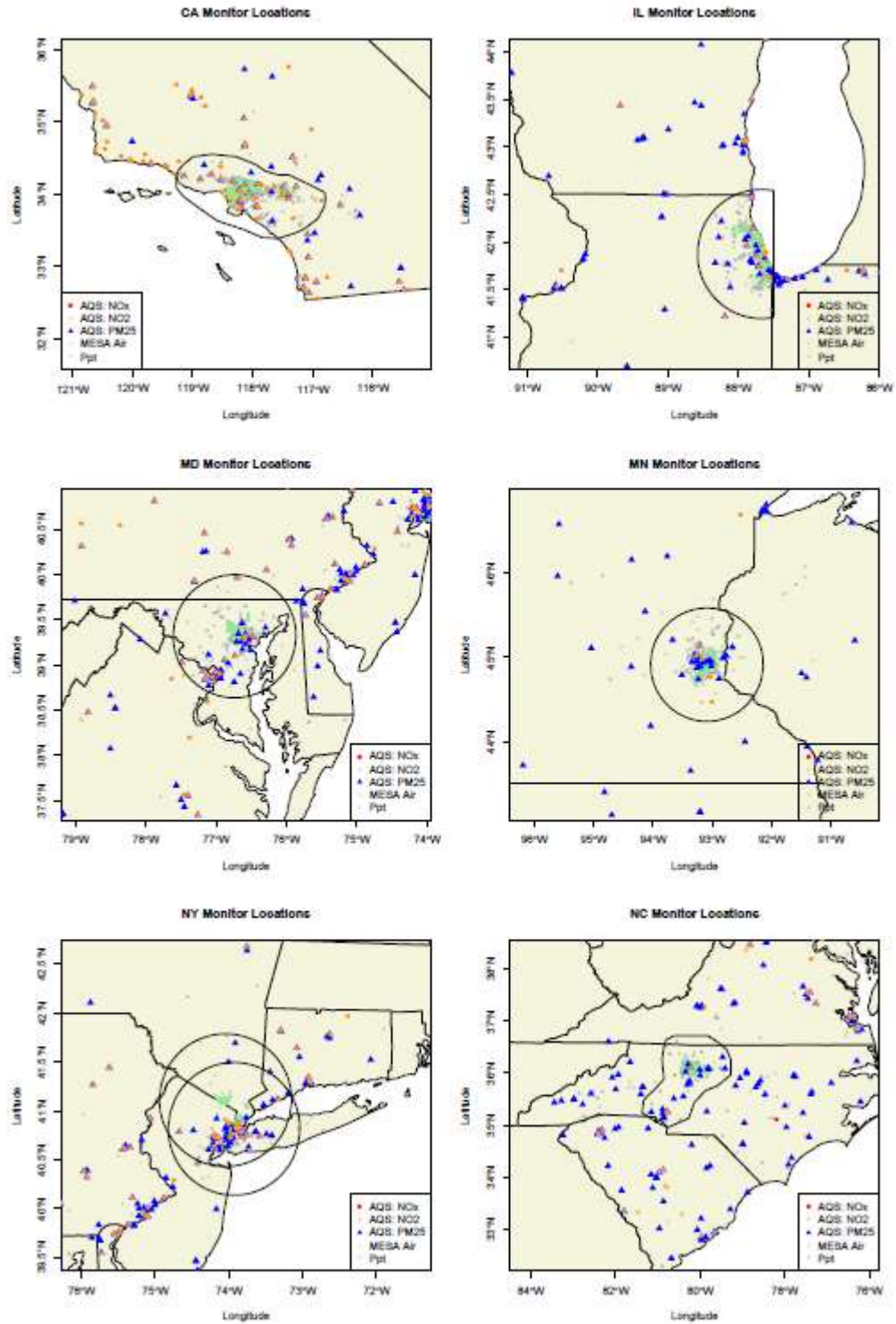
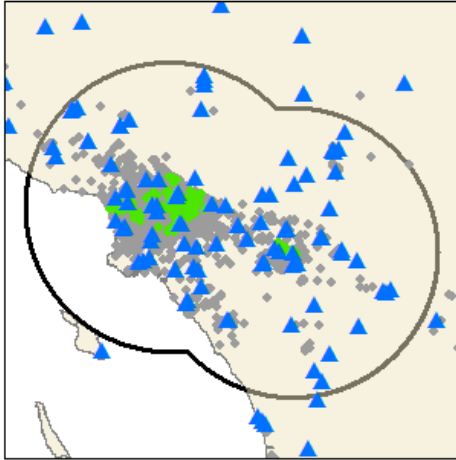


Figure 1: Modeling Regions for the Six MESA cities

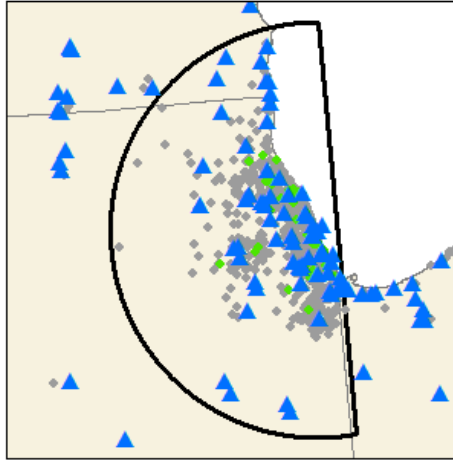
1

I. Ozone Modeling Regions

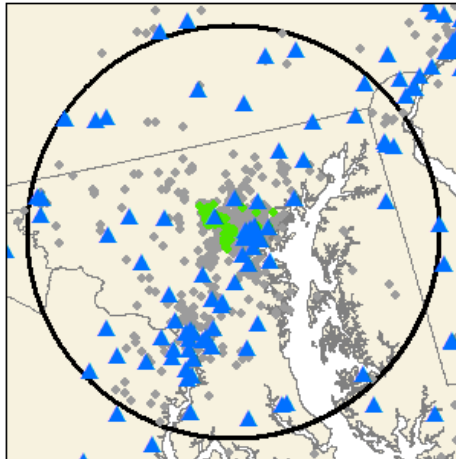
CA



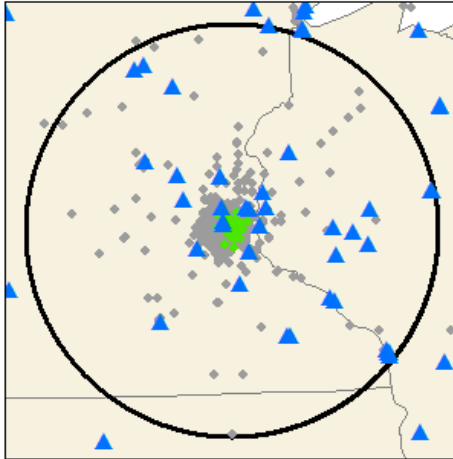
IL



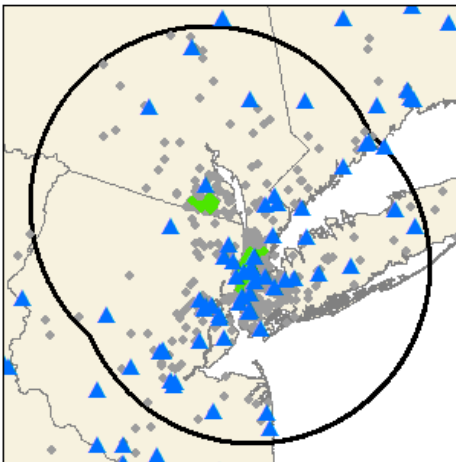
MD



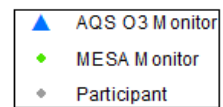
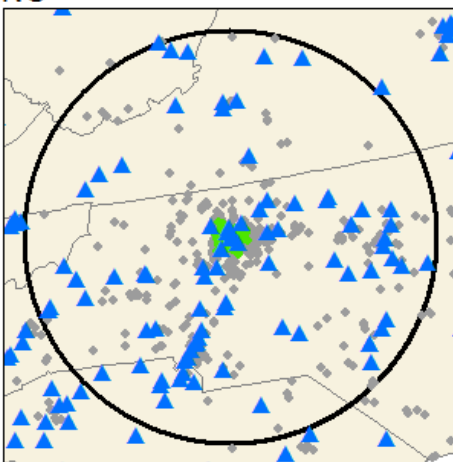
MN



NY



NC



J. Performance Statistics for Exposure Models

Table 26. Cross-validation measures of predictive accuracy for site means at monitoring locations for likelihood-based exposure models of NO₂, NO_x, PM_{2.5}, and LAC. Leave-one-out cross-validation was used for AQS and fixed sites and ten-fold cross-validation was used for home sites. New York models include NYCCAS reference sites with AQS and fixed sites and NYCCAS distributed sites with home sites. Units for RMSE are ppb (NO₂ and NO_x), µg/m³ (PM_{2.5}), and 10⁻⁵m⁻¹ (LAC). R²_{reg} represents the regression R². Table taken from Keller et al.²⁰

City	AQS and MESA Fixed Sites			MESA Home Sites		
	RMSE	R ² _{CV}	R ² _{reg}	RMSE	R ² _{CV}	R ² _{reg}
NO_x						
Baltimore	3.32	0.92	0.96	3.98	0.92	0.92
Chicago	3.88	0.87	0.91	6.08	0.59	0.59
Los Angeles	6.74	0.87	0.87	5.69	0.91	0.92
New York	8.85	0.61	0.89	16.66	0.50	0.50
St. Paul	1.69	0.98	0.98	3.58	0.83	0.84
Winston-Salem	5.46	0.00	0.94	3.74	0.60	0.63
NO₂						
Baltimore	0.76	0.96	0.97	1.47	0.90	0.90
Chicago	1.51	0.87	0.97	3.31	0.45	0.48
Los Angeles	2.23	0.88	0.89	3.13	0.77	0.78
New York	1.86	0.92	0.93	3.82	0.78	0.78
St. Paul	1.27	0.93	0.94	1.24	0.87	0.87
Winston-Salem	0.95	0.85	0.98	1.41	0.74	0.75
PM_{2.5}						
Baltimore	0.42	0.90	0.90	1.24	0.84	0.86
Chicago	0.59	0.78	0.82	1.43	0.80	0.80
Los Angeles	1.28	0.83	0.84	2.92	0.77	0.78
New York	0.59	0.91	0.91	2.80	0.54	0.56
St. Paul	0.60	0.45	0.84	1.78	0.78	0.79
Winston-Salem	0.44	0.89	0.90	1.00	0.85	0.86
LAC						
Baltimore	0.096	0.87	0.91	0.127	0.78	0.79
Chicago	0.045	0.86	0.92	0.108	0.61	0.62
Los Angeles	0.114	0.70	0.93	0.266	0.69	0.71
New York	0.147	0.75	0.79	0.329	0.51	0.52
St. Paul	0.043	0.91	0.92	0.074	0.69	0.69
Winston-Salem	0.020	0.94	0.99	0.088	0.64	0.65

²⁰ Keller JP, Olives C, Kim SY, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environ Health Perspect* 2015; 123: 301-309.

Table 27: Number of monitors and cross-validated measures of predictive accuracy for exposure models of NO₂, NO_x, PM_{2.5}, and O₃ for SPIROMICS Air. Leave-one-out cross-validation was used for AQS and fixed sites and ten-fold cross-validation was used for home sites.

City	Modeling region radius (km)	Number of monitors				AQS and fixed sites		Home sites	
		A	F	H/O	S	R_{reg}^2	R_{mse}^2	R_{reg}^2	R_{mse}^2
NO _x									
Baltimore	75	8	4	115	194	0.90	0.88	0.87	0.90
Los Angeles	90	28	7	186	263	0.85	0.85	0.89	0.90
San Francisco	75	15	0	32	100	0.73	0.70	0.71	0.82
New York	90	12	3	144	258	0.80	0.66	0.76	0.52
Salt Lake City	90	6	0	35	104	0.78	0.75	0.81	0.90
Ann Arbor	90	4	0	31	102	0.70	0.52	0.57	0.90
Winston-Salem	150	3	4	201	216	0.79	0.71	0.61	0.61
NO ₂									
Baltimore	75	11	4	115	194	0.82	0.82	0.90	0.92
New York	90	18	3	145	258	0.87	0.87	0.83	0.79
San Francisco	75	21	0	31	99	0.88	0.87	0.91	0.92
Los Angeles	90	29	7	186	265	0.88	0.87	0.88	0.92
Salt Lake City	90	6	0	35	104	0.80	0.75	0.88	0.95
Winston-Salem	150	2	4	207	216	0.63	0.54	0.73	0.82
Ann Arbor	90	5	0	31	102	0.67	0.61	0.75	0.93
O ₃									
Baltimore	75	25	0	88	91	0.93	0.90	0.79	0.85
New York	90	27	0	116	98	0.87	0.81	0.78	0.82
San Francisco	75	31	0	32	99	0.85	0.85	0.83	0.64
Los Angeles	90	33	0	182	95	0.82	0.71	0.78	0.90
Salt Lake City	175	21	0	34	104	0.82	0.82	0.80	0.84
Winston-Salem	200	49	0	191	95	0.68	0.68	0.68	0.55
Ann Arbor	200	34	0	30	102	0.85	0.84	0.60	0.82
PM _{2.5}									
Baltimore	150	34	5	108	0	0.89	0.88	0.92	0.93
New York	150	90	2	105	0	0.87	0.82		0.39
San Francisco	150	23	0	32	0	0.88	0.86		0.67
Los Angeles	100	25	2	88	0	0.86	0.83		0.89
Salt Lake City	120	21	0	35	0	0.93	0.89		0.71
Winston-Salem	175	52	4	121	0	0.89	0.89		0.88
Ann Arbor	175	46	0	31	0	0.88	0.88		0.29

A – AQS, F – Fixed, H/O – Home outdoor, S – Snapshot, R_{reg}^2 – Regression-based R-squared, R_{mse}^2 – MSE-based R-squared

Table 28: Model Performance using Cross Validation at AQS+Fixed Sites for O₃ Spatio-Temporal Models.

City	¹ AQS and MESA Fixed Sites					
	Overall CV _{MSE} R ²	Overall CV _{REG} R ²	² Spatial CV _{MSE} R ²	² Spatial CV _{REG} R ²	³ Temp CV _{MSE} R ²	³ Temp CV _{REG} R ²
⁴ LA	0.70	0.78	0.41	0.73	0.60	0.88
Baltimore	0.88	0.88	0.75	0.78	0.77	0.90
Chicago	0.82	0.84	0.59	0.66	0.83	0.89
New York	0.76	0.84	0.35	0.85	0.61	0.90
⁵ W-S	0.65	0.73	0.02	0.17	0.69	0.89
St Paul	0.81	0.82	0.39	0.66	0.68	0.90

1. Leave one out cross validation (LOOCV) at AQS and fixed locations;
2. Median CV_{MSE} and CV_{REG} R² based on annual averages at each AQS and fixed location across years, which reflects spatial prediction ability of the models;
3. Median CV_{MSE} and CV_{REG} R² between predictions and observations at two-week time points across the entire study period for individual sites.
4. Los Angeles basin including Los Angeles and Riverside
5. Winston-Salem

Table 29. 10-fold cross-validated R² and RMSE by year for national NO₂ model. All metrics are on the square root scale (sqrt(ppb)).

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
R²	0.83	0.86	0.86	0.87	0.89	0.85	0.86	0.86	0.87	0.87	0.85	0.86
RMSE	0.54	0.48	0.46	0.44	0.40	0.46	0.42	0.42	0.40	0.40	0.40	0.40
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
R²	0.88	0.89	0.85	0.85	0.85	0.82	0.83	0.84	0.82	0.85	0.79	0.83
RMSE	0.39	0.35	0.37	0.37	0.39	0.41	0.39	0.39	0.39	0.36	0.41	0.34
Year	2014	2015										
R²	0.83	0.83										
RMSE	0.35	0.35										

Table 30. 10-fold cross-validated R² and RMSE by year for national satellite-based NO₂ model. All metrics are on the square root scale (sqrt(ppb)).

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
R²												
RMSE												
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
R²												0.84
RMSE												0.34
Year	2014	2015										
R²	0.85	0.84										
RMSE	0.33	0.34										

Table 31. 10-fold cross-validated R² by year for national PM₁₀ model. All metrics are on the square root scale (sqrt(μg/m³)).

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
R²	0.58	0.62	0.59	0.63	0.54	0.58	0.51	0.52	0.47	0.53	0.53	0.51
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010			
R²	0.55	0.52	0.49	0.50	0.44	0.42	0.45	0.61	0.40			

Table 32. 10-fold cross-validated R² by year for national PM_{2.5} model. All metrics are on the square root scale (sqrt(μg/m³)). Year 2000 published in Sampson et al, Atmospheric Env, 2013.²¹

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
R²	0.88	0.88	0.89	0.88	0.87	0.88	0.91	0.88	0.89	0.85	0.86	0.77	0.73
Year	2012	2013	2014	2015									
R²	0.82	0.82	0.83	0.84									
RMSE	0.18	0.19	0.19	0.17									

Table 33. Cross-validated R² and RMSE for EC, OC, Si, and S PM_{2.5} Components. All metrics are on the square root scale. Published in Bergen et al, EHP 2013.²²

	EC	OC	Si	S
R²	0.79	0.60	0.36	0.63
RMSE	0.11	0.22	0.10	0.13

²¹ Paul David Sampson, Mark Richards, Adam A Szpiro, Silas Bergen, Sheppard Lianne, Timothy V Larson, Joel D Kaufman, 2013. A Regionalized National Universal Kriging Model using Partial Least Squares Regression for Estimating Annual PM_{2.5} Concentrations in Epidemiology. Atmospheric Environment 75, 383-392.

²² Bergen S, Sheppard L, Sampson PD, Kim SY, Richards M, Vedal S, Kaufman JD, Szpiro AA. 2013. A national prediction model for PM_{2.5} component exposures and measurement error-corrected health effect inference. *Environ Health Perspect* 121:1017-1025; <http://dx.doi.org/10.1289/ehp.1206010>

Table 34. Cross-validation statistics of the historical PM_{2.5} models for 1999-2010 by year and region.

Estimated trend		Linear trend from FRM/IMPROVE ^a PM _{2.5}	
Cross-validation statistics		R ²	MSE
Year/region	N ^b		
All ^c	1,460 (10,800)	0.87	2.08
1999	523	0.86	3.29
2000	865	0.85	2.38
2001	988	0.86	2.32
2002	1,054	0.84	2.39
2003	969	0.85	2.13
2004	980	0.86	1.99
2005	940	0.88	2.08
2006	898	0.86	1.87
2007	937	0.86	1.85
2008	902	0.82	1.82
2009	884	0.80	1.61
2010	860	0.83	1.63
East ^c	1,056 (7,956)	0.86	1.19
Mountain West ^c	239 (1,594)	0.59	3.84
West Coast ^c	165 (1,250)	0.84	5.50

- a. FRM = Federal Reference Method; IMPROVE = Interagency Monitoring of Protected Visual Environment; CASTNet = Clean Air Status and Trends Network; WBAN = Weather-Bureau-Army-Navy
b. Number of sites (Number of observations when different from the number of sites)
c. Annual averages from 1999 through 2010

Table 35. 10-fold cross-validated R^2 on native and model scale for As, Du, Ni, SO₄, So₂, NO₃, V, and Cr. * All components except for Cr were developed using covariates from Rev 5. Cr was built using Rev 4 covariates.

Pollutant	R^2	R^2 (native)	Modeling scale
As	0.84	0.59	log
Cu	0.79	0.26	log
Ni	0.75	0.48	log
SO ₄	0.91	0.87	sqrt
SO ₂	0.37	0.32	sqrt
NO ₃	0.86	0.82	sqrt
V	0.80	0.59	log
Cr*	0.74	0.07	log

Table 36. Model performance (10-fold cross validated R^2 and RMSE) for PM_{10-2.5} mass ($\mu\text{g}/\text{m}^3$) and species concentrations (ng/m^3) using land use regression (LUR). Table adapted from Zhang et al, Under review at EHP, 2013.²³

City	CV Measure	Total Mass	Cu	P	Si	Zn
Chicago, IL	R^2	0.68	0.65	0.50	0.68	0.73
	RMSE	1.16	2.29	3.88	82.10	10.63
St Paul, MN	R^2	0.51	0.86	0.68	0.93	0.40
	RMSE	2.33	0.61	4.14	72.60	4.44
Winston Salem, NC	R^2	0.41	0.51	0.76	0.48	0.36
	RMSE	1.09	0.93	3.95	73.10	1.89
All cities	R^2	0.47	0.73	0.55	0.43	0.65
	RMSE	1.84	0.00	0.00	0.15	0.01

²³ Zhang K, Larson TV, Gasset A, Szpiro AA, Daviglius M, Burke GL, Kaufman JD, Adar SD. Characterising spatial patterns of airborne coarse particulates (PM_{10-2.5}) mass and chemical components in three cities: the Multi-Ethnic Study of Atherosclerosis. Under review at Environmental Health Perspectives.

Table 37. Leave-one-out cross validation RMSE of PM_{2.5} using “Pragmatic Model” at all sites and at fixed sites on native scale (ug/m3). * N is the number of sites used for modeling. Modified from Table 2 in Sampson et al, Atmospheric Environment, 2011.²⁴

Site	All Sites				MESA Air Fixed Sites			
	N*	Mean	SD	RMSE	N	Mean	SD	RMSE
CA	31	17.84	4.47	2.42	7	20.64	3.04	1.98
IL	51	14.52	1.83	1.32	7	14.59	1.93	1.86
MN	44	10.68	1.49	0.88	3	10.61	0.72	1.05
MD	44	14.90	1.23	1.05	5	15.77	0.90	1.24
NY	48	13.66	1.58	1.21	3	14.70	2.52	1.75
NC	33	14.38	0.90	0.95	4	14.84	0.33	1.02

²⁴ Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45(36), 6593-6606.