Research Report 238

Ambient Air Pollution and COVID-19 in California

Michael Kleeman et al.

Appendix A. Supplemental Information for Chapter 3: Development of Chronic and Subchronic Exposure Fields

---

Appendix A was reviewed by the HEI Review Committee and has been lightly edited for spelling, grammar, punctuation, and cross-references to the main report.

Correspondence may be addressed to Dr. Michael Kleeman, University of California, Davis, Department of Civil and Environmental Engineering, 1 Shields Avenue, Davis, CA 95616; email: *mjkleeman@ucdavis.edu*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not necessarily reflect the views of the Agency; thus, no official endorsement by it should be inferred. It also has not been reviewed by private-party institutions, including those that support the Health Effects Institute, and may not reflect the views or policies of these parties; thus, no endorsement by them should be inferred.

CONTENTS

DAILY LAND USE REGRESSION MODELS DEVELOPMENT

Land use regression (LUR) model development for the HEI grant is integrated with funding from the California Air Resources Board (CARB) for the entire state. The following steps describe the development of such models and surfaces for the state of California for the pollutants nitrogen dioxide (NO2) and fine particulate matter (PM2.5) over the study period 2016–2020.

Develop comprehensive data sources through multiple platforms

The data sources were acquired and processed on multiple platforms through two major scripting languages. They included statewide daily traffic data for California highways, daily remote sensing data, daily weather data, parcel-level land use, detailed land cover data, biweekly vegetation index, and data on impervious surfaces and tree canopy.

R scripting on a workstation of 128 GBs of memory and 32 TBs of storage space: In California, traffic detectors covered 12.52% of highway segments. We used road type category criteria of the nearest neighbor to derive daily roadway traffic for the entire state of California, and those derived roadway traffic data were converted into daily traffic surfaces of 30-m resolution for the years 2016-2020. We also incorporated parcel-level land use data from 58 counties for the 40 million people in California in our modeling process for a spatial resolution of 30 m. The parcel-level land use data included agricultural, residential, commercial, industrial, governmental, and institutional uses as well as open land, parks, and recreational facilities. We also had daily remote-sensing data from the Ozone Monitoring Instrument for NO_2 at 25-km spatial resolution. Other potential predictors of 30-m spatial resolution included elevation (digital elevation model), distance to coast, distance to ports, and distance to highway roadways.

Google Earth Engine JavaScript scripting: We also included comprehensive land cover data (16 classes of 30-m resolution data, such as forest, shrubland, and developed land) from the USGS (US Geological Survey) NLCD (national land cover database), biweekly NDVI (normalized difference vegetation index; 250-m resolution) data from the NASA (National Aeronautics and Space Administration) MODIS (Moderate Resolution Imaging Spectroradiometer) instrument, tree canopy (30-m resolution) and impervious surface (30-m resolution) data from NLCD. TBs of daily Aerosol Optical Depth for $PM_{2.5}$ at 1-km resolution were also acquired from the NASA MAIAC (multi-angle implementation of atmospheric correction) algorithm. Further, we processed TBs of daily meteorological conditions data of 4-km resolution (called gridMet) provided by the University of Idaho. The data included maximum and minimum temperature, precipitation accumulation, downward surface shortwave radiation, wind velocity, maximum and minimum relative humidity, and specific humidity.

Generate buffered distance statistics on 30-m spatial resolution potential predictors

A series of buffered distance statistics of 50-5000 m at an interval of 50 m was created for the potential spatial predictors with a spatial resolution of 30 m (except for the traffic, which had buffered distance statistics of 50-2000 m). They included R scripting for daily traffic and parcel-level land use data and Google Earth Engine scripting for land cover data, % impervious surfaces, and tree canopy data. For each predictor (e.g., industrial land use), a total of 100 buffered distance statistics (i.e., covariates) were generated (40 for traffic). For all the potential predictors, including buffered and non-buffered predictors, we generated about 2200 covariates in predicting daily pollutant concentrations of a pollutant. This increases the chance of identifying the optimal distance impact of a predictor and

helps improve model performance. However, this also creates high-dimensional covariates that are highly correlated. To solve this issue, we applied a data reduction strategy to reduce the number of covariates used in predicting pollutant concentrations.

Apply a data reduction strategy to reduce the potential number of predictors

To reduce the number of covariates and avoid high collinearity between them for LUR modeling, we first created a correlation coefficient matrix between a pollutant (a response variable) and all the covariates (predictors). The covariate of the highest correlation with the pollutant was first selected and maintained as part of the reduced dataset. Then, the correlation coefficients between the first selected covariate and all the remaining covariates were calculated, and those covariates with an absolute correlation coefficient with the first selected covariate greater than or equal to 0.9 were removed. The covariate of the second-highest correlation with the pollutant from the remaining covariates was then selected and maintained as part of the reduced dataset. This process continued until no covariates could be further selected and maintained in the reduced dataset (i.e., all the chosen covariates for the reduced dataset had a correlation coefficient smaller than 0.9 between them). After applying the data reduction strategy, we maintained the number of covariates in an LUR model to be less than 150.

Integrate three types of air pollution measurements into a single modeling framework

In our modeling process, we incorporated data into a single modeling framework from multiple air pollution measurement instruments, including those from government continuous monitoring across California; our fixed sites saturation monitoring[1] in Los Angeles, Alameda, and Sacramento counties; and Google Streetcar mobile monitoring across San Francisco Bay (counties of Alameda, San Francisco, and San Mateo), Los Angeles County, and Central Valley regions (see: https://www.google.com/earth/outreach/special-projects/air-quality/). Those government continuous monitors are inherently sparsely distributed and do not typically have significant spatial autocorrelation[2] in our modeling process. The fixed sites saturation monitoring in our research was designed through a location-allocation algorithm, and it also does not have significant spatial autocorrelation. The Google Streetcar mobile measurements for each region are highly spatially autocorrelated because of the intense sampling of air pollutants on its road network. To reduce spatial autocorrelation of air pollutants measured from the Google Streetcar, we applied a location-allocation algorithm[3] to select 150 road segments for each of the four regions: Alameda and Contra Costa, San Francisco and San Mateo, Los Angeles, and the Central Valley. Each region had (1) 50 road segments selected from locations within 500 m of highways allowing truck traffic or within 500 m of major California ports (i.e., goods movement corridors or GMCs), (2) 50 road segments selected from locations within 500 m of highways not allowing truck traffic or within 300 m of major roadways (i.e., non-goods movement corridors or NGMCs), and (3) locations not encompassed in the first and second parts (i.e., control areas or CTRLs).

To integrate three types of air quality measurements into a single modeling framework, we divided each type (e.g., Google Streetcar mobile monitoring) or its sub-type (e.g., Google Streetcar mobile monitoring in Los Angeles) of air quality monitoring data equally into 10 folds and then merged corresponding folds of data into a large 10-fold dataset, with each fold having an equal presentation of the three types and corresponding sub-types of air quality monitoring data. The equal presentation of 10-fold data was then used in a v-fold, out-of-sample, cross-validation technique for LUR modeling.

Develop daily LUR models through v-fold, out-of-sample, cross-validation machine learning techniques

In developing daily LUR models for the three pollutants, we aimed to develop the models at their finest spatial resolution of 30 m. We also aimed to identify the optimal distance of impact for a potential predictor, and the models needed to be able to deal with multicollinearity among predictors and reduce model overfit. Further, we wanted to avoid excessive predictors in the final models and allowed a maximum of 20 predictors (in addition to four seasons) in an LUR model. Given those considerations, we applied the D/S/A machine learning algorithm in modeling daily pollutant concentrations.[4] The D/S/A machine learning algorithm is an aggressive model search algorithm, which iteratively generates polynomial generalized linear models based on the existing terms in the current "best" model and the following three steps: (1) a deletion step, which removes a term from the model, (2) a substitution step, which replaces one term with another, and (3) an addition step, which adds a term to the model. The search for the "best" estimator starts with the base model specified with "formula": typically, the intercept model, except when the user requires the number of terms to be forced in the final model. The original sample is randomly partitioned into V equal-sized subsamples before searching through the statistical model space of polynomial functions. Of the V subsamples, a subsample is retained as the validation data for testing the model, and the remaining V-1 subsamples are used as training data. The cross-validation process is then repeated V times, with each of the V subsamples used exactly once as the validation data.

The advantage of this method over the leave-one-out cross-validation technique is that single outliers have less impact on the prediction errors and, compared to repeated random sub-sampling, all observations in the V-folds are used for both training and validation, and each observation is used for validation once. With each iteration, an independent validation dataset is used to assess the performance of a model built using a training dataset. This technique minimizes over-fitting to the data to maximize the probability that the models will predict well at locations that have not been sampled. In addition, the D/S/A algorithm can deal with both linear and non-linear associations. However, for simplicity of model development and the clear interpretation of the predictors selected for a model, we limited the predictors to only linear terms (requiring the maximum sum of powers in each variable to be 1) and disallowed any interaction.

We developed daily LUR models for NO_2 and $PM_{2.5}$ across California at a spatial resolution of 30 m (about 3 GBs of storage space required for a daily raster). To save storage space, the daily surfaces were built for a spatial resolution of 100 m (about 400 MBs), which still maintains the ability to identify the small area variations of pollutant concentrations.

Apply the global production chain technique for model development and surface construction

One of the approaches of the global production networks and value chains (GVCs)[5] uses specialization to parcel out parts production to developing countries where the raw materials are acquired and then import those made parts back to a developed country for final product assembly and sales. Because most predictors had TBs of storage space (e.g., daily meteorological data and daily remote sensing AOD data), we applied a process like GVCs to first estimate predictor statistics separately. Next, the data from Google Earth Engine was processed through Google Earth Engine, and the data stored on campus workstations were processed through respective hardware. The individual predictor statistics calculated from separate platforms were then merged into a single file on another workstation and used in the D/S/A machine learning algorithm to develop land use models for the three pollutants. The same technique was used to build the air pollutant surfaces. Each of the final selected predictors was converted into corresponding daily surfaces through its acquisition platform and by having its model

regression coefficient and/or the buffered statistics information included. All the individual surfaces were then transferred into another workstation, and a Python script was developed to derive daily air pollutant surfaces for the state. We used Python scripting in generating final surfaces, because it does not limit memory usage in a program. Instead, it allocates as much memory as a program needs until the computer is out of memory. With 128 GB of memory and 32 TBs of storage space, we designed a Python script that generated surfaces day by day. Once a daily pollutant surface was generated, it was outputted into a physical location and removed from memory. This process continued until all the daily surfaces were generated.

The final developed LUR models explain 80% of the variance in NO2 concentrations (see Table A1) and 65% of the variance in PM2.5 concentrations (see Table A2).

Table A1. Daily NO2 Model for the State of California

Coefficient	Estimate	Std. Error	Statistic	*P*-Value
Season [Fall]	41.31525359	1.27380322	32.43456526	**<0.001**
Season [Spring]	37.88236015	1.27857987	29.62846605	**<0.001**
Season [Summer]	37.84886991	1.30048646	29.10362477	**<0.001**
Season [Winter]	42.05274917	1.25767472	33.43690420	**<0.001**
Vegetation index (NDVI)	-0.00018156	0.00001679	-10.81666470	**<0.001**
Week [Weekend]	-2.32441019	0.03671236	-63.31410266	**<0.001**
Distance to ports (m)	-0.00000584	0.00000024	-23.88315011	**<0.001**
NO_2 from OMI	7.327461e-16	5.426231e-18	135.03776831	**<0.001**

VKT (350 m)	0.00006147	0.00000071	86.71347177	**<0.001**
Developed high intensity (ha) (5000 m)†	0.00017474	0.00000231	75.62641371	**<0.001**
Minimum relative humidity (%)	-0.12444801	0.00102569	-121.3315915	**<0.001**
Wind velocity at 10 m (m/s)	-0.93918093	0.01122917	-83.63763975	**<0.001**
Roadway area (ha) (50 m)	6.29933371	0.10347870	60.87565365	**<0.001**
Minimum temperature (K)	-0.09471578	0.00443069	-21.37721199	**<0.001**
Percent impervious (%) (50 m)	0.01781697	0.00091568	19.45772955	**<0.001**
Developed low-intensity (ha) (400 m)	0.01218760	0.00020753	58.72813594	**<0.001**
Shrubs (ha) (3250 m)	-0.00009070	0.00000304	-29.82997246	**<0.001**
Water (ha) (50 m)	-1.93161136	0.07029621	-27.47817093	**<0.001**

Developed open space (ha) (50 m)	-0.19145437	0.01075227	-17.80594787	**<0.001**
Residential (ha) (350 m)	-0.07513870	0.00236946	-31.71130369	**<0.001**
Precipitation amount (mm, daily total)	0.04020234	0.00378600	10.61868762	**<0.001**
Wetlands (ha) (550 m)	-0.02732793	0.00117326	-23.29238367	**<0.001**
Observations	N = 162,570			
R^2 / R^2 adjusted	0.796 / 0.796			

NDVI = Normalized Difference Vegetation Index; OMI = Ozone Monitoring Instrument; VKT = Vehicle Km Traveled.

[†]The content in the first pair of parentheses is the unit of analysis, and the content in the second pair of parentheses is the distance of the buffer.

Table A2. Daily PM2.5 Model for the State of California

Coefficient	Estimate	Std. Error	Statistic	*P*-Value
Season [Fall]	90.21122937	1.04163758	86.60519819	**<0.001**
Season [Spring]	88.15829132	1.04862676	84.07022866	**<0.001**
Season [Summer]	89.58297126	1.06433106	84.16833306	**<0.001**
Season [Winter]	90.66738819	1.02822904	88.17820237	**<0.001**
AOD	0.03232083	0.00014388	224.64103333	**<0.001**
Wind velocity at 10 m (m/s)	-0.91353396	0.00935806	-97.62006415	**<0.001**
Roadway area (ha) (5000 m)[§]	0.00057177	0.00002919	19.58814640	**<0.001**
Minimum temperature (K)	-0.27202880	0.00361147	-75.32355385	**<0.001**
Minimum relative humidity (%)	-0.10749589	0.00109883	-97.82789738	**<0.001**
DEM (m)	-0.00355748	0.00006678	-53.26864451	**<0.001**
Industrial (ha) (1850 m)	0.00997592	0.00032603	30.59788510	**<0.001**

Distance to ports (m)	0.00001155	0.00000027	42.05332038	**<0.001**
Residential (ha) (850 m)	0.00904293	0.00040292	22.44333719	**<0.001**
VKT (350 m)	0.00000772	0.00000073	10.62487610	**<0.001**
NDVI	-0.00035052	0.00001309	-26.76815385	**<0.001**
Barren land (ha) (3000 m)	-0.00073488	0.00002057	-35.73111153	**<0.001**
Shrubs (ha) (200 m)	-0.01737372	0.00087123	-19.94171252	**<0.001**
Location category‡	-0.39053840	0.02256090	-17.31041212	**<0.001**
Developed open space (ha) (4950 m)	-0.00007838	0.00000264	-29.65769646	**<0.001**
Unknown land use (ha) (450 m)	-0.04719305	0.00195384	-24.15395733	**<0.001**
Agricultural (ha) (50 m)	-2.88221319	0.13664311	-21.09300113	**<0.001**
Observations	310720			
R^2 / R^2 adjusted	0.653 / 0.653			

VKT = Vehicle Km Traveled; NDVI = Normalized Difference Vegetation Index; DEM = Digital Elevation Model; AOD = Aerosol Optical Depth, and monthly median values were used.

‡Location category: 1 = GMC; 2 = NGMC and 3=CTRL.

§The content in the first paired parentheses is the unit of analysis; the content in the second pair of parentheses is the circular buffer distance.

CHEMICAL TRANSPORT MODEL

Simulations for the year 2016 were carried out across California, using the source-oriented University of California, Davis-California Institute of Technology (UCD-CIT) regional air quality model.

A moving sectional bin approach is used[1] so that particle number and mass can be explicitly conserved, with particle diameter acting as the dependent variable.

The emissions of particle source tracers are empirically set to be 1% of the total mass of primary particles emitted from each source category, so they do not significantly change the particle radius and the dry deposition rates. For a given source, the simulated concentration of the artificial tracer directly correlates with the amount of PM mass emitted from that source in that size bin. The corresponding number concentration (*num*) attributed to source *i* can be calculated using Equation (1)

$$num_i = \frac{tracer_i \times 100}{\frac{\pi}{6} Dp^3 \rho}, \qquad \text{(Equation 1)}$$

where $tracer_i$ represents the artificial tracer mass in size bin *i*, *Dp* is the core particle diameter, and ρ is the core particle density. Core particle properties are calculated by removing any condensed species to better represent the properties of the particles when they were emitted. More details describing the source apportionment technique in the UCD/CIT model are provided in previous studies.[2–6]

A total of 50 particle-phase chemical species are included in each size bin. Gas-phase concentrations of oxides of nitrogen (NOx), volatile organic compounds (VOCs), oxidants, ozone, and semi-volatile reaction products were predicted using the SAPRC-11 chemical mechanism.[7] Phase change for inorganic species occurs using a kinetic treatment for gas-particle conversion[8] driven towards the point of thermodynamic equilibrium.[9] Phase change for organic species is also treated as a kinetic process, with vapor pressures of semi-volatile organics calculated using the 2-product model.[10]

UCD/CIT model calculations were carried out using three nested model domains with 24-km, 4-km, and 1-km horizontal spatial resolution over the study domain. Sixteen telescoping levels were used in the vertical dimension, with a thickness of 30 m at ground level and 1000 m at the top height of 5 km.

Meteorological Model

Hourly meteorology inputs to drive the regional chemical transport model at 24-km, 4-km, and 1-km resolution in the year 2016 were simulated using the Weather Research and Forecasting (WRF) v3.4 model (https://www.mmm.ucar.edu/models/wrf). WRF model vertical resolution was 31 vertical layers from the ground level to the top pressure of 100 hPa. Initial and boundary conditions for meteorological simulations were taken from the North American Regional Reanalysis, which has a spatial resolution of 32 km and a temporal resolution of 3 h. The Yonsei University (YSU) boundary layer vertical diffusion scheme[11] and Pleim-Xiu land surface scheme[12] were adopted in this study. Four-dimensional data assimilation was applied to anchor the model predictions to observed meteorological patterns.

Emission Inventories

The year 2016 area source and point source emission inventories used in the current study were provided by the CARB, with several modifications. Fugitive dust emissions were replaced by an online dust model[13] based on the wind speed and soil moisture predicted by the WRF model. This change corrects the positive bias in dust emissions and $PM_{2.5}$ mass noted by Hu et al.[14,15] A major point source of unpaved road dust at MAGTFTC/MCAGCC Twentynine Palms military facility in San Bernardino County was converted to an area source over a 9-km^2 region around the base. Food cooking emissions in GAI 6069 (Victorville in San Bernardino County) were reduced by a factor of three so that the per capita emissions from food cooking activities were similar to those in Los Angeles County.

Area source emissions inventories provided by CARB had a spatial resolution of 4 km. Area source emissions with a spatial resolution of 1 km were created for major sources using spatial surrogates processed with the Spatial Allocator software maintained by the U.S. EPA. This software summarizes the spatial surrogates listed in Table A3 to downscale 4-km CARB area emissions to 1 km, accounting for 80% of the statewide area source emissions.

Table A3. Spatial Surrogates Used to Downscale 4-km CARB Area Emissions to 1 km

Surrogate	Description	Data Source
302	Industrial-related/industrial employment	See details in reference paper: DOI 10.1016/j.atmosenv.2020.117665
441	Total population	
587	Off-road construction equipment	
588	On-road construction equipment	
621	Service & Commercial employment	
651	Single-family housing	
720	Farm road VMT	California Air Resources Board (CARB) provided
190	Forestland	
530	Residential Gas Heating	
660	Unpaved road	
100	All airports	
140	Commercial airports	
382	Military airports	
610	Secondary paved road	Tiger/Line shapefile, S1400 + S1630 + S1640
480	Primary Road	Tiger/Line shapefile, S1100 + S1200
570	Residential heating – wood	California Air Resources Board (CARB) shapefile
560	Restaurants	Food service market dataset from ESRI (NACIS 7225)

Mobile Source Emissions

Three spatial surrogates were created to downscale mobile emissions to 1-km resolution, including gasoline mobile, diesel mobile, and tire/brake wear. Explicit traffic counts collected by the U.S. Highway Performance Monitoring System were used to distribute the majority of the tailpipe emissions to highways and other principal arterial roads. MacDonald et al.[16] showed that ~70% of gasoline, and ~80% of diesel vehicle fuel consumption in California occurs on roads with traffic count information. Emissions on these roads can be represented by VMT (i.e., traffic count × road length). The remaining ~30% of gasoline and ~20% of diesel vehicle activity can use road length as a spatial surrogate. This approximate treatment for the residual portion of the tailpipe emissions was done separately for urban and rural areas to ensure rural emissions were not overestimated.[17] 90% of the unmonitored gasoline and diesel activity occurs in urban areas, with the balance in rural areas. The final mobile gasoline and diesel surrogates were calculated using the equations:

$$Gasoline\ mobile\ surrogate = 70\% \times (AADT \times road\ length)_{normalized} + 30\% \times (road\ length\ without\ traffic\ counts)_{normalized}$$

$$Diesel\ mobile\ surrogate = 80\% \times (Truck\ AADT \times road\ length)_{normalized} + 20\% \times (truck\ road\ length\ without\ traffic\ counts)_{normalized}$$

$$(Truck)\ Road\ length\ without\ traffic\ counts = 90\% \times urban\ road\ length + 10\% \times rural\ road\ length.$$

Tire and brake wear emissions were estimated as a fixed fraction of tailpipe emissions for all engine types. The 2016 CARB emissions inventories[18] specify that gasoline/diesel emissions account for 86% to 14% of total mobile emissions. Thus, the tire and brake wear spatial surrogate was calculated using the equation:

$$Tire\ \&\ brake\ wear\ surrogate = 86\% \times (Diesel\ mobile\ surrogate)_{normalized} + 14\% \times (Gasoline\ mobile\ surrogate)_{normalized}$$

Data sources used for traffic surrogates are listed in Table A4.

Table A4. Data Sources Used for Traffic Surrogates

Description	Data Source
Gasoline vehicle traffic count – Average Annual Daily Traffic (AADT)	https://www.fhwa.dot.gov/policyinformation/hpms/shapefiles.cfm, accessed August 2020
Diesel vehicle traffic count – Truck AADT (with three or more axles)	Caltrans
Road shapefiles	https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html, accessed August 2020
Truck road network as defined in the Freight Analysis Framework	https://ops.fhwa.dot.gov/freight/freight_analysis/faf/, accessed August 2020).

Soil NOx

Candidate soil NOx emissions were included in the calculations based on a biogeochemical model combined with fertilizer application rates.[19] Soil NOx emissions varied by month of the year, based on the effects of temperature on the biogeochemical cycle. Sensitivity studies carried out across years between 2000 and 2015 indicate that the inclusion of soil NOx emissions improves the accuracy of model predictions for gas-phase ozone and particulate nitrate.[20]

Biogenic Emissions

Biogenic emissions were generated using the Model of Emissions of Gases and Aerosols from Nature (MEGANv2.1) based on the meteorological fields generated using the WRF model. The gridded geo-referenced emission factors and land cover variables required for MEGAN calculations were created using the MEGAN v2.1 pre-processor tool and the ESRI_GRID leaf area index, as well as plant functional type files available at the Community Data Portal.[21]

Wildfires

Daily values of wildfire emissions were generated using the Global Fire Emissions Database (GFED).[22] Wildfire emissions were assigned the same particle size and composition distribution as routine biomass combustion. Typical wildfire plumes rise to 6-10 km in the atmosphere, depending on the intensity of the fire and the local meteorological conditions.[23] Wildfire plumes were injected at the top of the model domain at a height of approximately 5 km in the current simulations.

Wildfire emissions were represented using the GFED, which uses satellite images of burned areas, combined with vegetation maps to estimate smoke released each day during wildfires.[24] Spatial resolution of GFED emissions inventories is 0.25 degrees. Smoke from these fires impacted cities throughout central California, as plumes were trapped within the Central Valley. Wildfire emissions were assigned particle size and composition profiles based on measurements during biomass burning experiments.[25]

BIAS CORRECTION

Predicted monthly averaged $PM_{2.5}$ concentrations were compared to measured $PM_{2.5}$ concentrations at all available monitoring sites across the study domains for the entire duration of the study year 2016. Summary statistics were calculated to characterize CTM performance, including the correlation coefficient (*R*), mean fractional error, mean fractional bias (MFB), mean error, mean bias (MB), and root mean square error. $PM_{2.5}$ predictions were moderately correlated with measured concentrations ($R > 0.5$ at more than half of the monitoring sites), the predicted concentrations exceeded measured concentrations by a factor of approximately 50% (average MFB=0.549). This overprediction is likely caused by an under-prediction of vertical mixing and dilution associated with the combination of updates to the WRF model v3.4 and the incorporation of non-local transport terms into the aerosol advection/diffusion algorithms.

Figure A1 illustrates the distribution of *R* and MFB values across the 40 monitoring sites in the study domain.

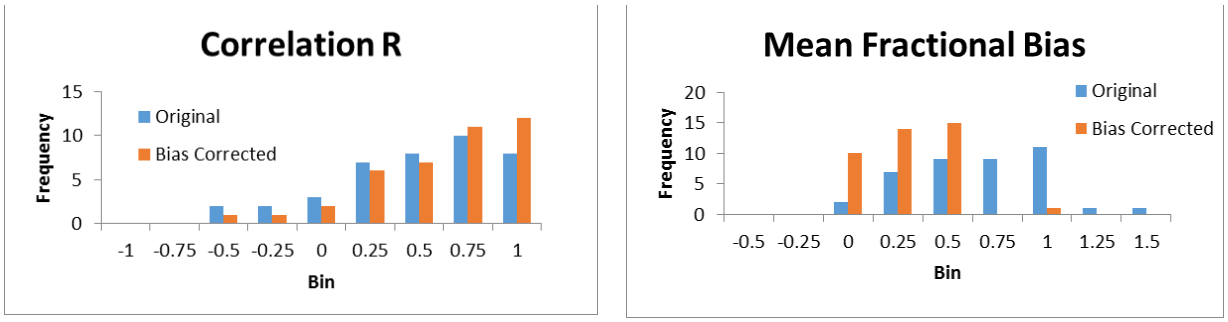

Figure A1. Summary of performance statistics for PM2.5 mass after bias correction. Ideal values are R = 1 and MFB = 0.

Bias corrections were only applied to primary PM species components emitted directly into the atmosphere in the particle phase. Concentrations of secondary PM components predicted by the CTM were not adjusted, because the measurements at the limited number of speciation sites suggested that secondary components were not over-predicted to the same extent as total mass. Bias corrections were also not applied to gas-phase species, such as O_3 and NO_2, because these species are formed from chemical reactions in the atmosphere that have a non-linear dependence on atmospheric mixing in which increasing concentrations of some species, such as NO, can decrease concentrations of other species, such as O_3. The spatial pattern of the gas-phase concentrations should be approximately correct in the current analysis, but future studies should correct the mixing in the meteorological fields and repeat the CTM calculations to remove bias in all species.

CHEMICAL TRANSPORT MODEL RESULTS

Figure A2 shows the location of $PM_{2.5}$ monitoring locations in the core of the study domain. Figures A3-A6 show the time series of predicted $PM_{2.5}$ mass concentrations and measured concentrations across the counties within the study domain. Model predictions have been bias corrected using the methods described in previous sections. Model predictions at most locations are generally in reasonable agreement with measured concentrations. Overall, $PM_{2.5}$ predictions have a slight positive bias.

Figure A5c shows that predicted $PM_{2.5}$ concentrations are 2-4 times higher than measured values at the monitoring site near Victorville, California (population 121,902) in San Bernardino County. This overprediction results from overestimated emissions in this urban location. Food cooking emissions were scaled down to match per capita values in Los Angeles County, but emissions from other area sources were not rescaled. Given the small population in Victorville, this isolated overprediction in $PM_{2.5}$ concentrations should not have a large influence on study results.

Seasonal patterns in both predicted and measured $PM_{2.5}$ concentrations are modest. Most residences in the study region use natural gas or electricity for home heating during winter months, and so the much higher winter concentrations associated with residential wood combustion are generally absent at most sites except around Bakersfield (see, for example, Figure A6a–d). Modest increases in concentrations are observable in winter and summer months, due to more stagnant atmospheric conditions compared to spring and fall months.

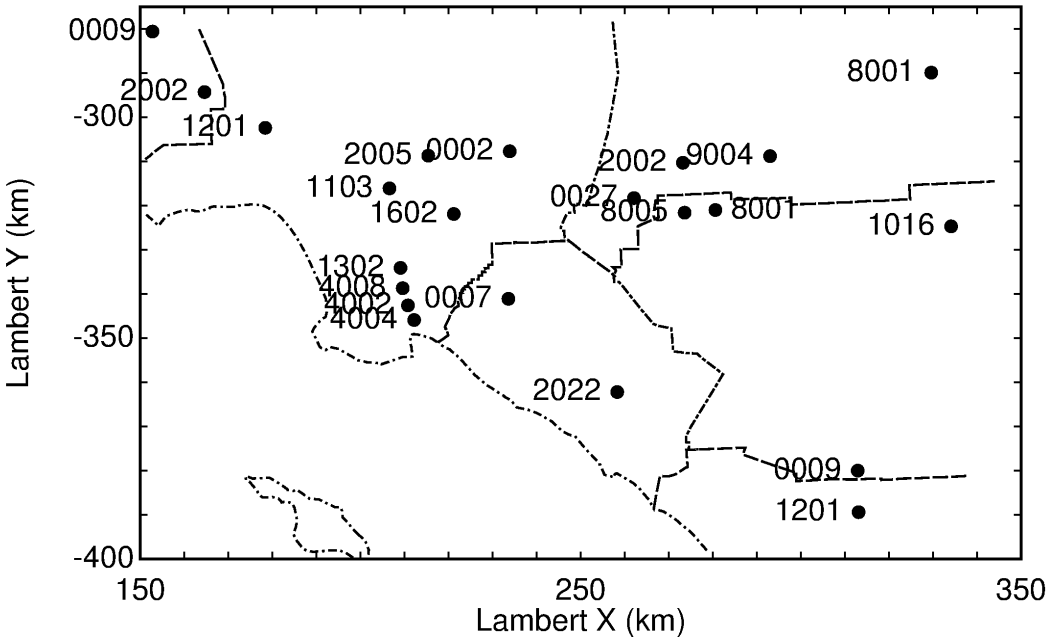

Figure A2. Locations of $PM_{2.5}$ mass monitoring sites around the central portion of the study domain, which contains the majority of the study population. Full site codes shown in subsequent figures are preceded by the state identification number (California=06) and the county FIPS code (Ventura=061, Los Angeles=037, Orange=059, San Diego=073, San Bernardino=071, Riverside=065).

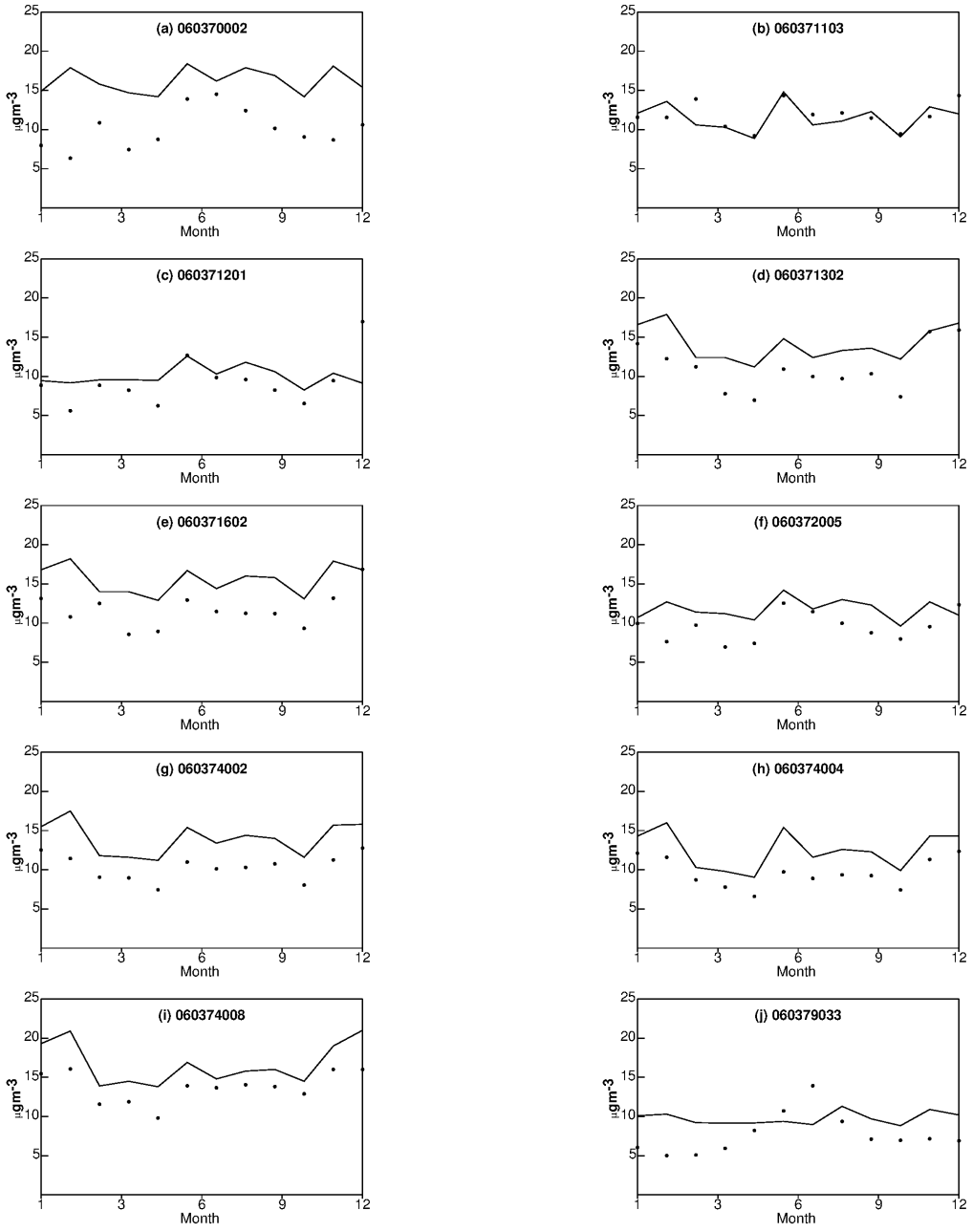

Figure A3. Time series of predicted (solid line) vs measured (dots) monthly average $PM_{2.5}$ mass concentrations at measurement locations in Los Angeles County. All model concentrations have been bias corrected. Measurement site codes correspond to names designated by the U.S. EPA monitoring network.

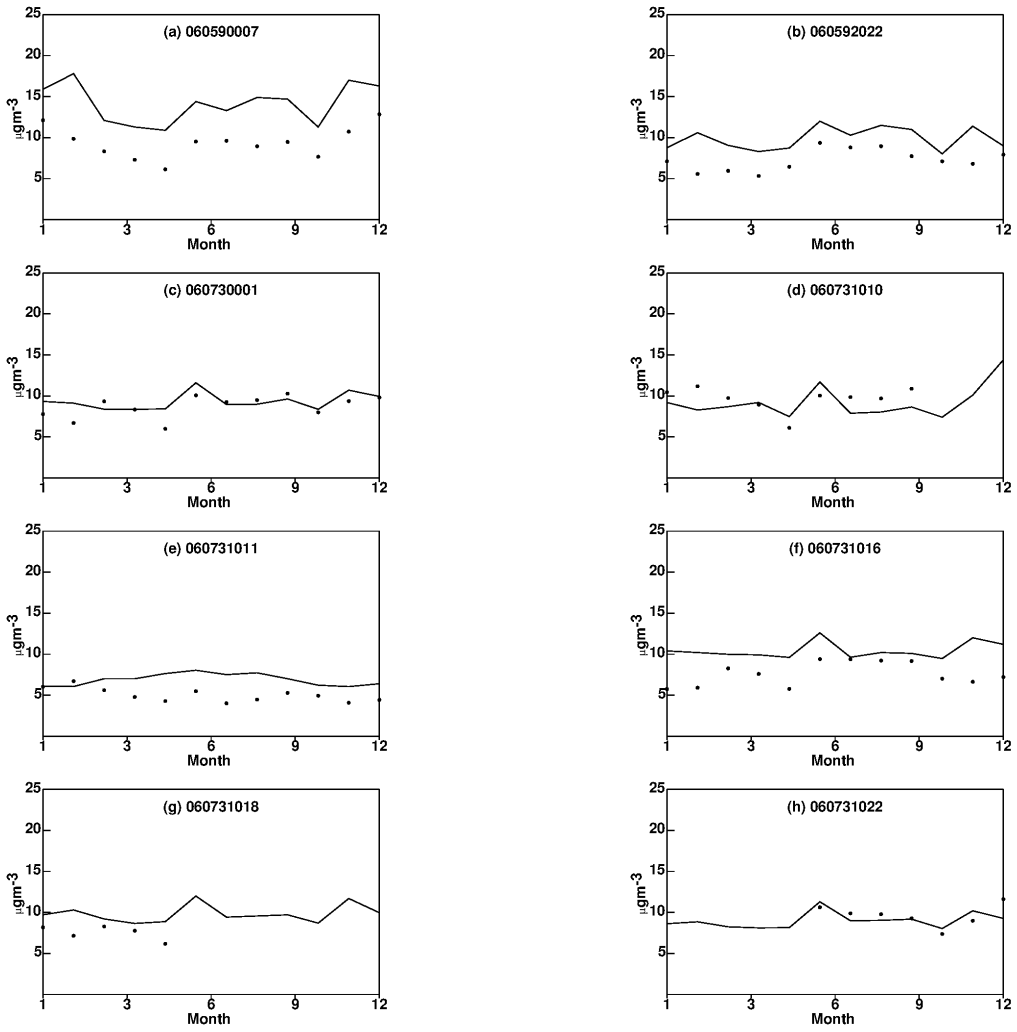

Figure A4. Time series of predicted (solid line) vs measured (dots) monthly average $PM_{2.5}$ mass concentrations at measurement locations in Orange County and San Diego County. All model concentrations have been bias corrected. Measurement site codes correspond to names designated by the US EPA monitoring network.

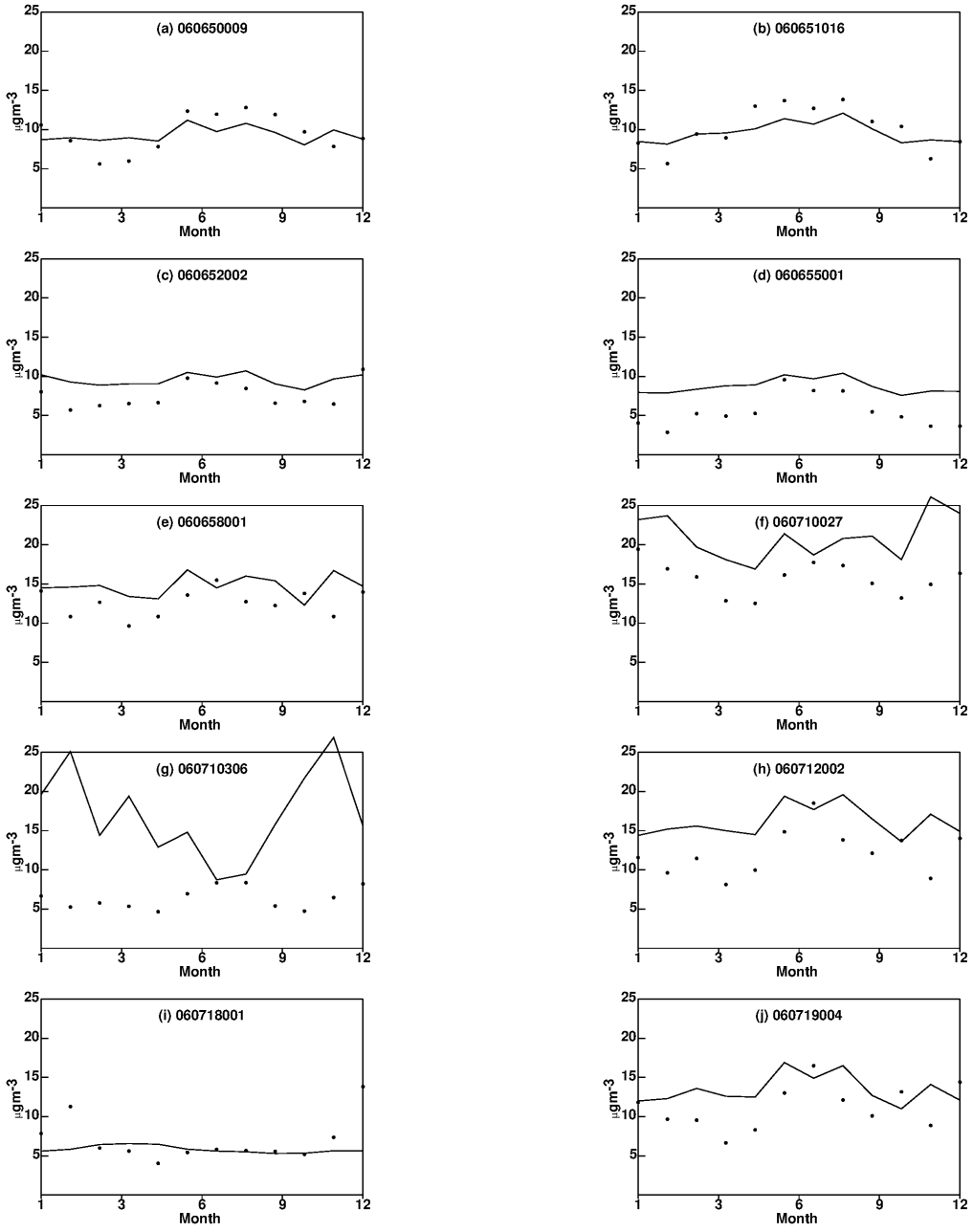

Figure A5. Time series of predicted (solid line) vs measured (dots) monthly average $PM_{2.5}$ mass concentrations at measurement locations in Riverside County and San Bernardino County. All model concentrations have been bias corrected. Measurement site codes correspond to names designated by the US EPA monitoring network.

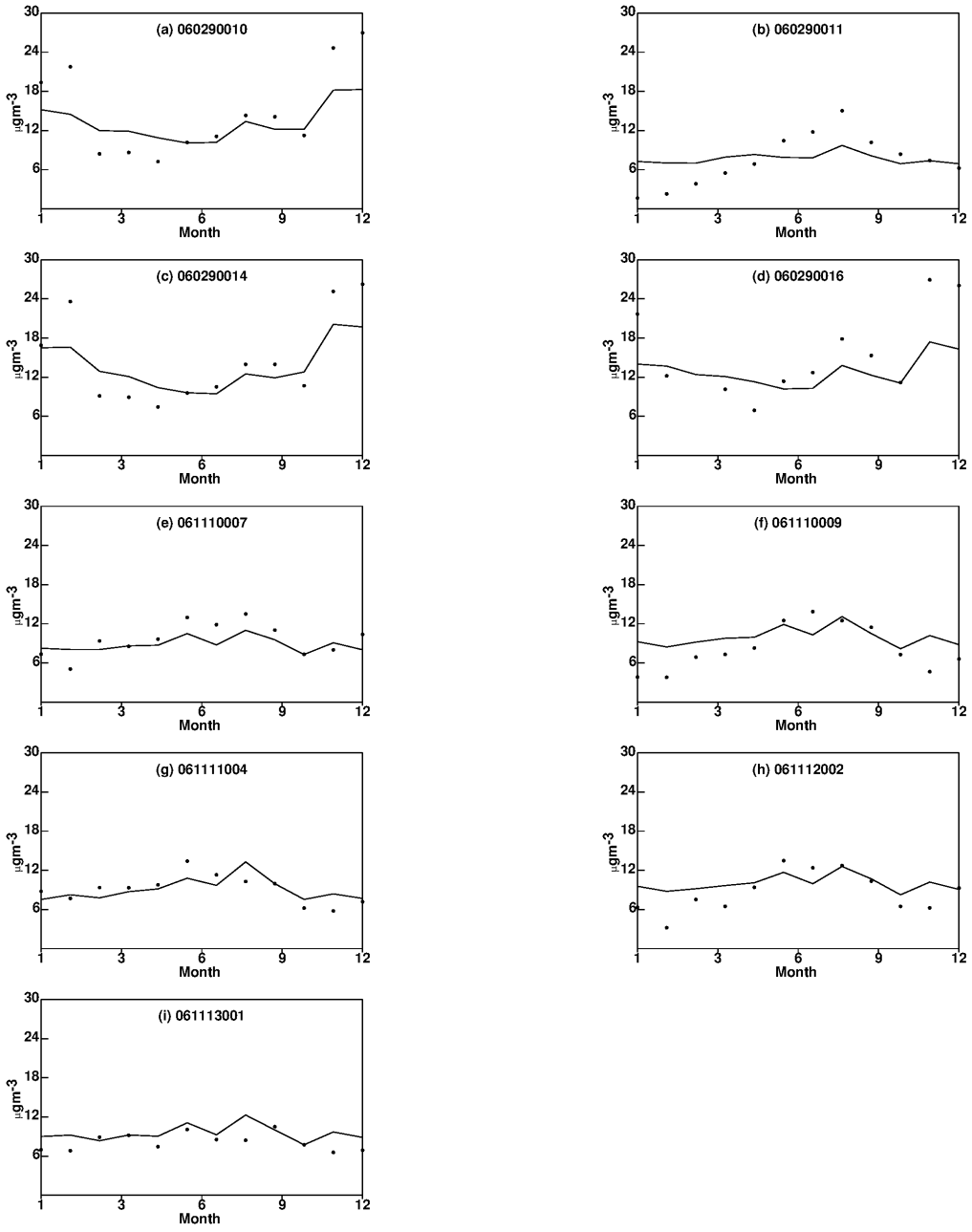

Figure A6. Time series of predicted (solid line) vs measured (dots) monthly average $PM_{2.5}$ mass concentrations at measurement locations in Kern County and Ventura County. All model concentrations have been bias corrected. Measurement site codes correspond to names designated by the US EPA monitoring network.

Figure A7 displays the predicted ground-level daily maximum 1-hr average O_3 concentration averaged during each season of the year 2016. The scale in each subpanel is adjusted based on seasonal trends, with the highest concentrations in the summer and the lowest concentrations during the winter. O_3 concentrations generally increase moving from west to east (downwind) in the air basin. Maximum summer concentrations occur in the mountains north of Los Angeles, where anthropogenic NOx emissions mix with biogenic VOC emissions, leading to enhanced O_3 formation. As noted previously, gas-phase concentrations were not bias corrected in the current study, and so the displayed concentrations may reflect errors associated with under-predicted wind speeds.

Figure A8 illustrates the predicted ground-level $PM_{2.5}$ mass exposure fields over the study domain during each season of the year 2016. The scale has been adjusted to show concentrations over major population centers. The maximum predicted $PM_{2.5}$ concentrations over military airports (circled) are off scale, but this does not significantly affect population-weighted exposures. Maximum $PM_{2.5}$ mass concentrations occur east of central Los Angeles in San Bernardino County. Elevated concentrations of $PM_{2.5}$ mass are also predicted to occur along major transportation corridors connecting the Port of Los Angeles and the Port of Long Beach with distribution centers in San Bernardino County.

Figure A9 illustrates the predicted ground-level $PM_{2.5}$ elemental carbon (EC) exposure fields over the study domain during each season of the year 2016. EC is a primary pollutant directly emitted from diesel engines and from gas direct injection gasoline engines. The pattern of EC concentrations, therefore, follows major transportation corridors, with a maximum value once again occurring over distribution centers in San Bernardino County. Increased stagnation in the atmosphere during winter and summer months leads to higher EC concentrations, compared to spring and fall months.

Figure A10 illustrates the predicted ground-level $PM_{2.5}$ nitrate concentrations over the study domain during each season of the year 2016. Nitrate is a secondary pollutant formed from atmospheric chemical reactions involving precursor NOx emissions. Regional nitrate patterns are generally more distributed than regional patterns of EC (compare Figure A9 to Figure A10. Maximum nitrate concentrations generally occur over a broad area east (downwind) of central Los Angeles. Concentrations are generally higher in the colder winter months because nitrate can evaporate in warmer months.

Figure A11 illustrates the predicted ground-level $PM_{2.5}$ concentrations associated with primary particulate matter emitted from on-road diesel engines. As expected, the spatial pattern generally follows major transportation corridors, with a noticeable maximum at distribution centers in San Bernardino County. The seasonal pattern of the primary on-road diesel particulate matter is similar to the seasonal pattern for EC (see Figure A9).

Ultrafine particles with a diameter less than 0.1 μm can be emitted directly (primary pollutant) or formed in the atmosphere through either condensation or nucleation processes (secondary pollutant). The $PM_{0.1}$ concentration fields illustrated in Figure A12 show evidence of both pathways. Fall and winter concentrations are highest over distribution centers in San Bernardino County as a result of primary emissions from activities related to the movement of goods. $PM_{0.1}$ concentrations in the spring are highest over the Port of Los Angeles, given the +conversion of sulfur emissions to sulfuric acid that subsequently partitions to the particle phase. $PM_{0.1}$ concentrations during summer are highest in the foothills of the mountains to the north of Los Angeles, where anthropogenic and biogenic emissions mix. Overall, the $PM_{0.1}$ mass exposure fields have the greatest seasonal variability of all the considered pollutants.

Figure A13 illustrates O_3 and $PM_{2.5}$ mass concentrations in the years 2016, 2019, and 2020 to verify the stability of the concentration patterns over time. O_3 concentrations in all years increase from background concentrations along the California coast to close towards peak concentrations north and east

of central Los Angeles. O_3 concentrations in the southern portion of the San Joaquin Valley are lower than concentrations in surrounding regions because of the influence of fresh NOx emissions in all years. $PM_{2.5}$ mass concentrations in all years are elevated over the populated regions of the South Coast Air Basin surrounding Los Angeles, with higher concentrations generally observed east of central Los Angeles. Hotspots are predicted around military bases, which are generally located in less populated locations that will not have a strong influence on the epidemiological analysis. Year-to-year differences in the exposure fields are caused by El Niño-Southern Oscillation weather patterns, the location of wildfires, and behavior shifts associated with COVID-19.

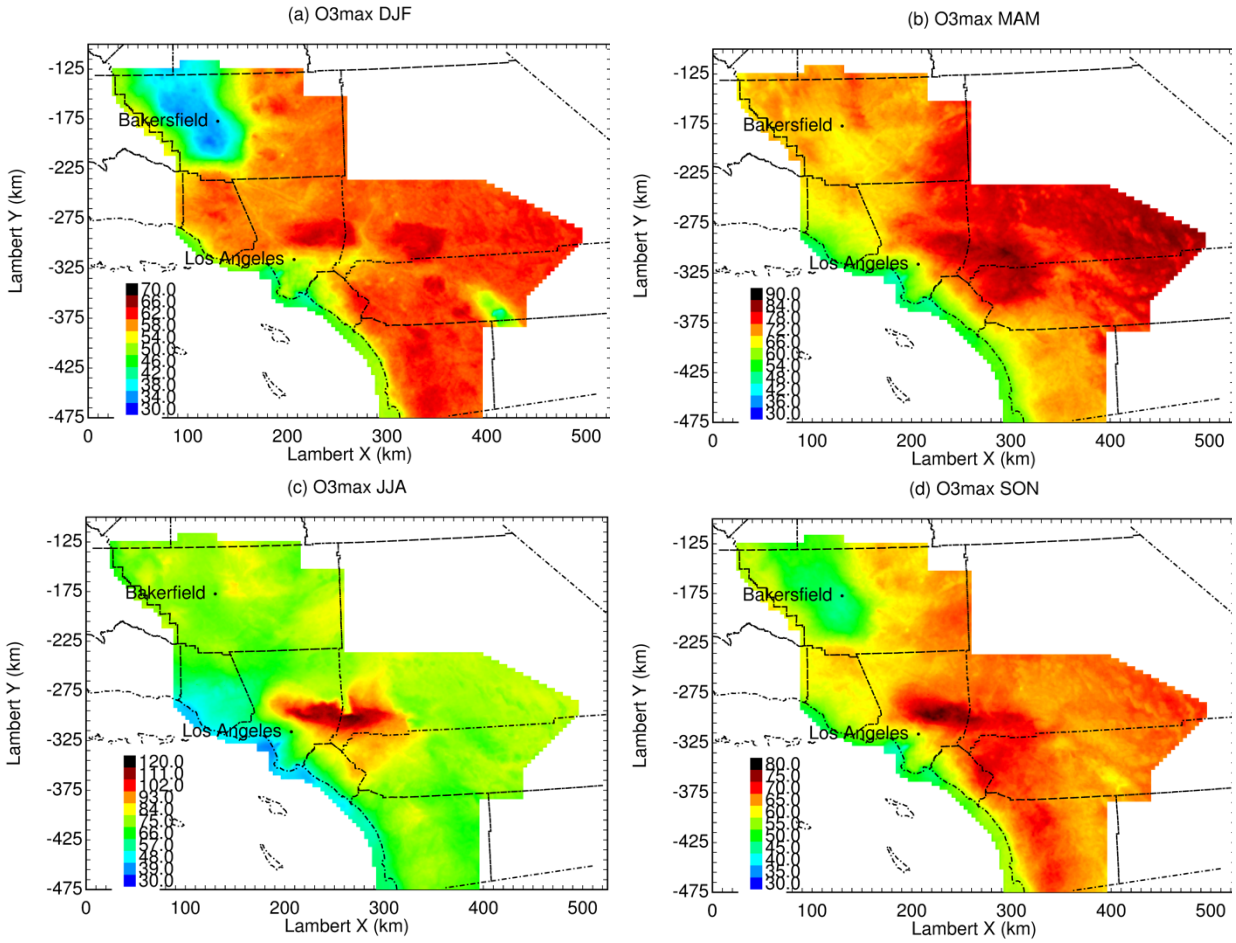

Figure A7. Predicted O_3 maximum exposure fields during four seasons in the year 2016. All units are ppb. DJF = December January February; JJA = June July August>; MAM = March April May; SON = September October November.

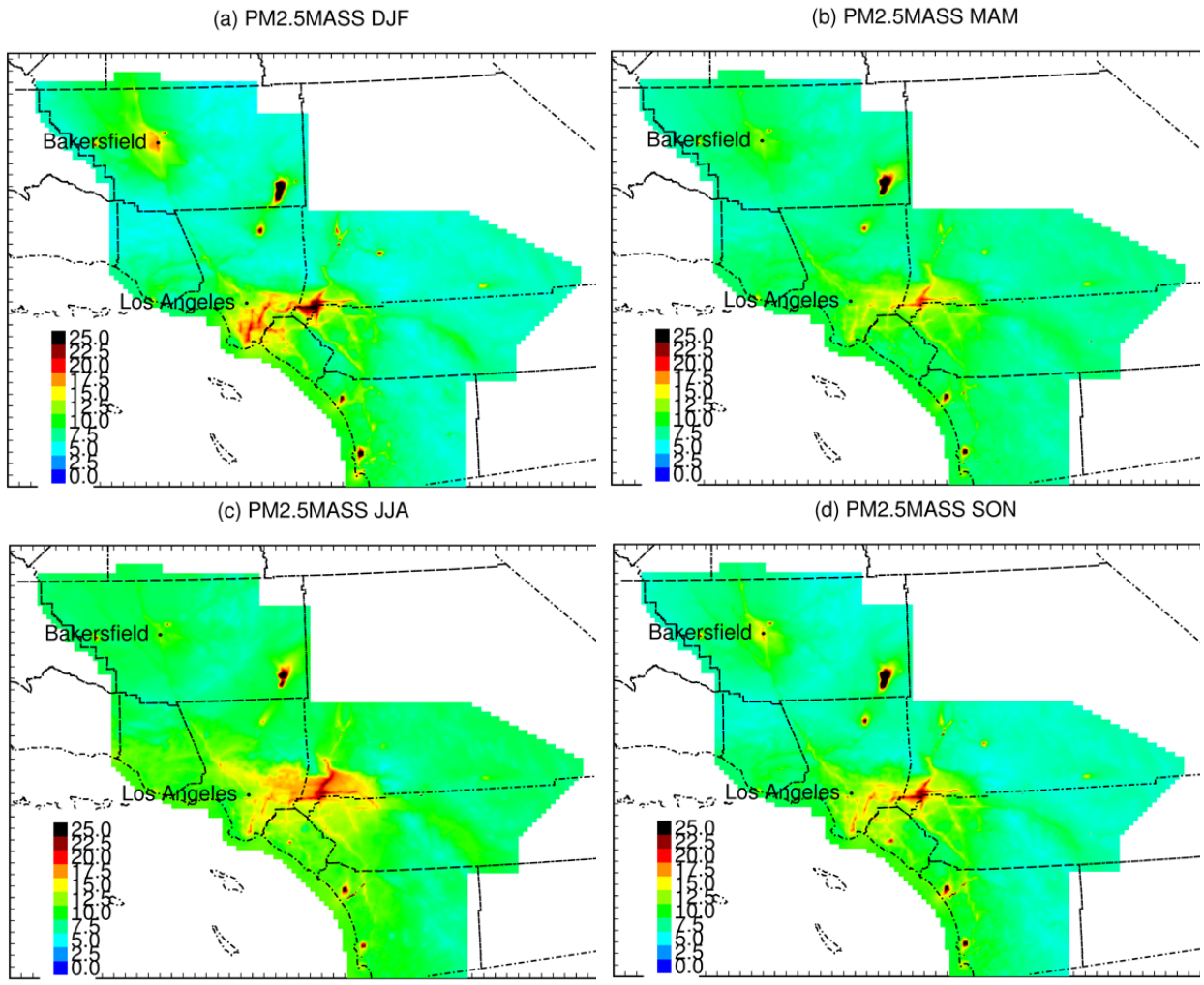

Figure A8. Predicted $PM_{2.5}$ mass exposure fields during four seasons in the year 2016. All units are µg m-3. DJF = December January February; JJA = June July August; MAM = March April May; SON = September October November.

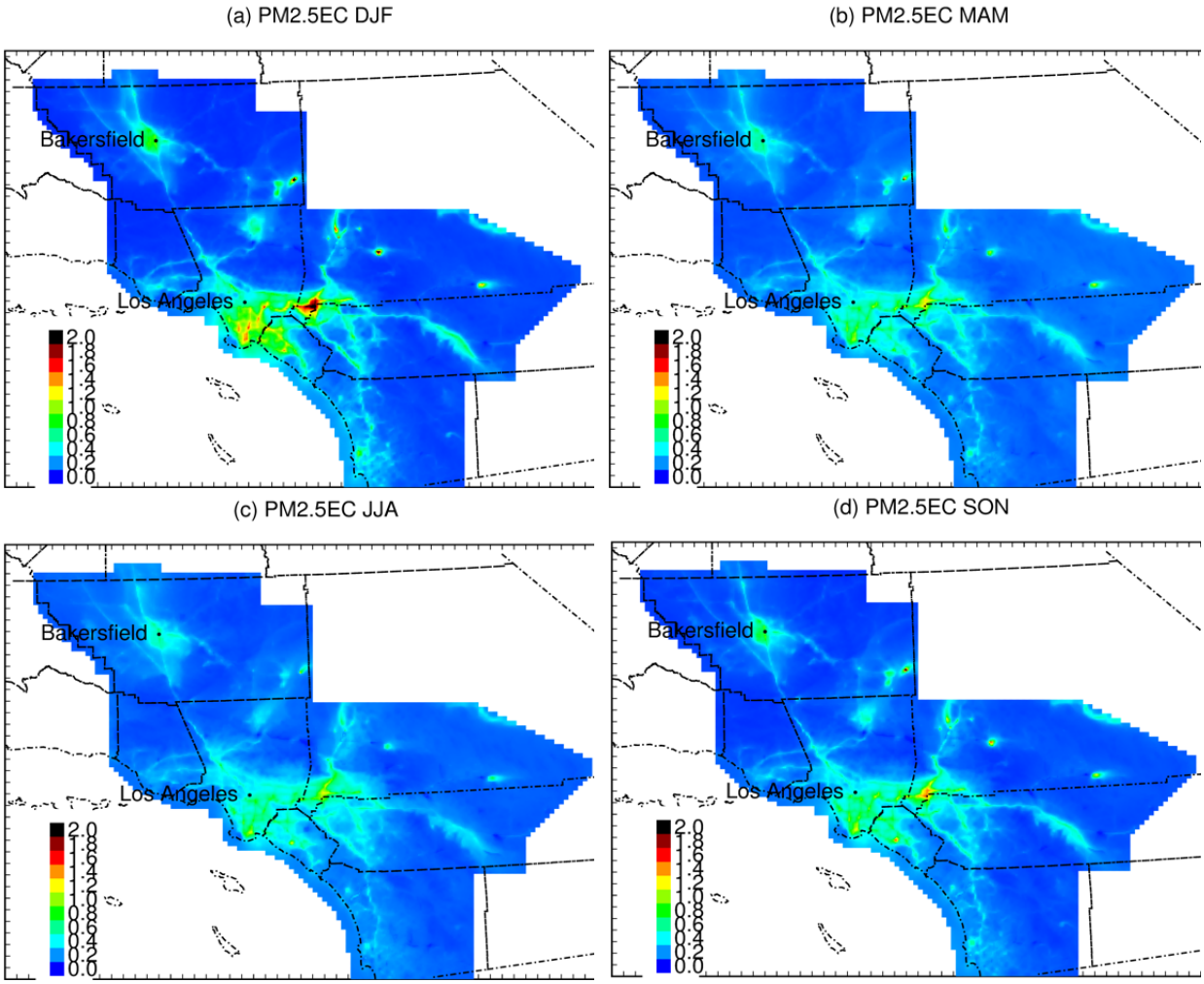

Figure A9. Predicted $PM_{2.5}$ elemental carbon exposure fields during four seasons in the year 2016. All units are µg m-3. DJF = December January February; JJA = June July August; MAM = March April May; SON = September October November.

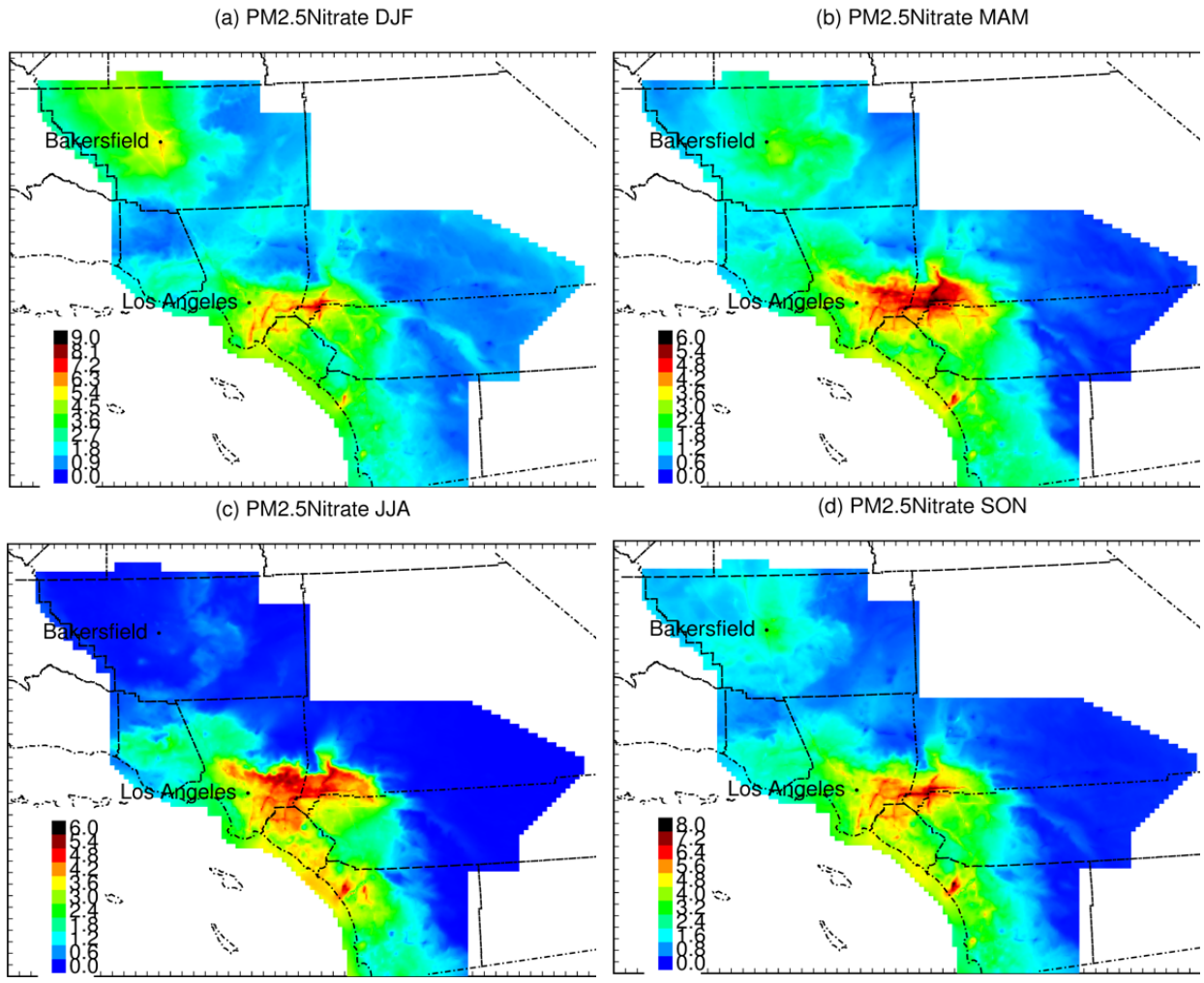

Figure A10. Predicted $PM_{2.5}$ nitrate exposure fields during four seasons in the year 2016. Note the different maximum values in different seasons. All units are µg m-3. DJF = December January February; JJA = June July August; MAM = March April May; SON = September October November.

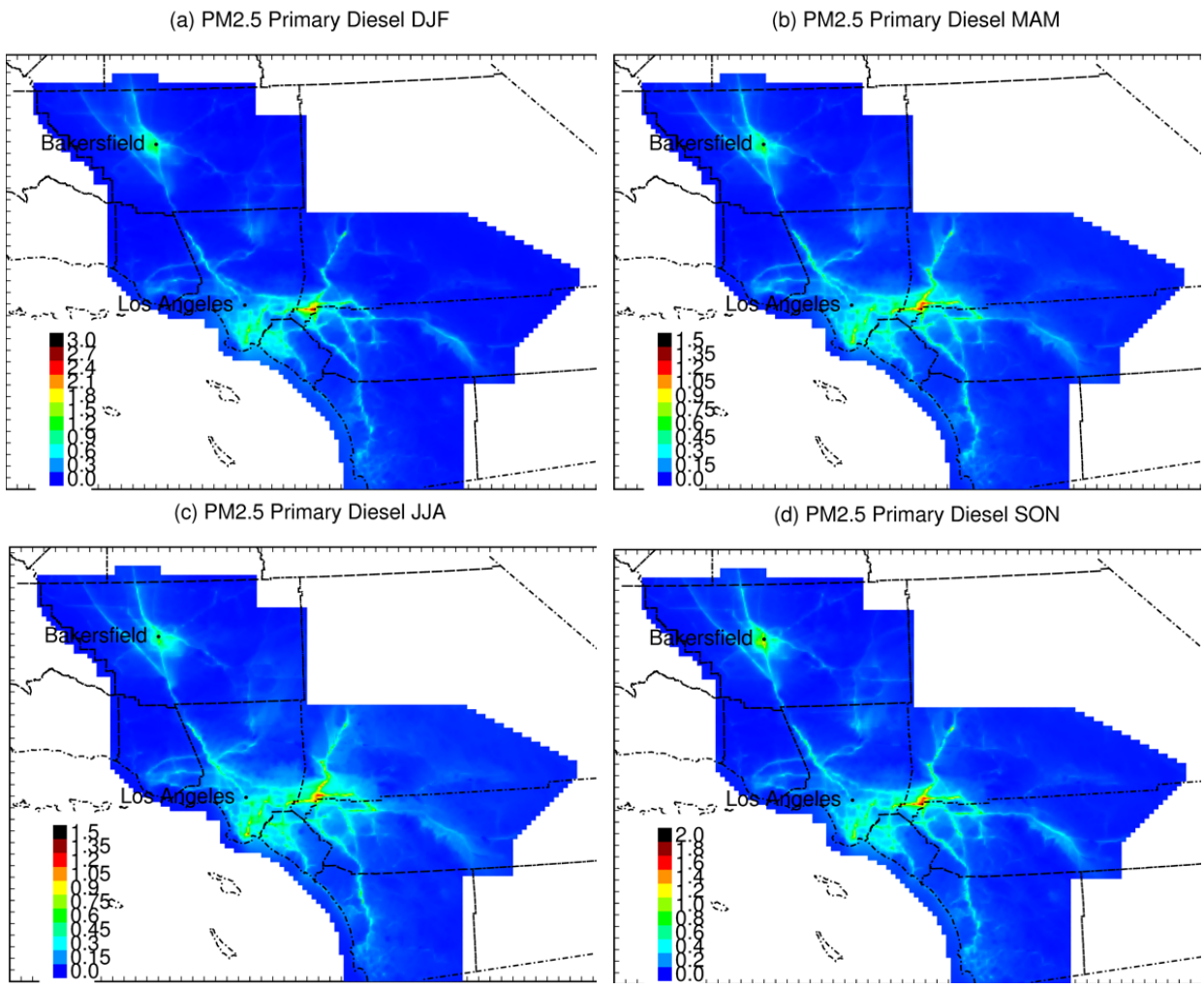

Figure A11. Predicted diesel primary $PM_{2.5}$ mass exposure fields during four seasons in the year 2016. Note the different maximum values in different seasons. All units are µg m-3. DJF = December January February; JJA = June July August; MAM = March April May; SON = September October November.

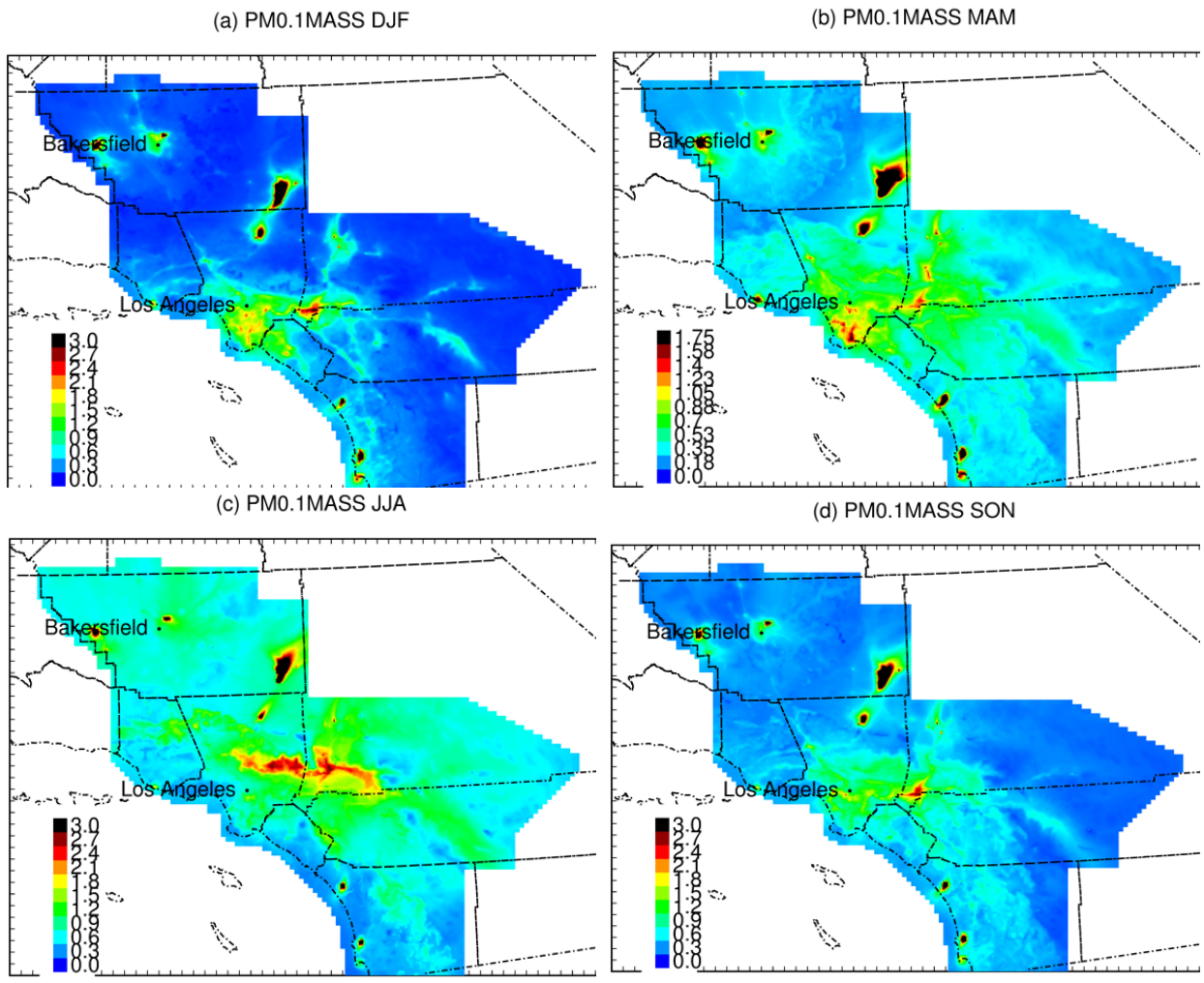

Figure A12. Predicted $PM_{0.1}$ mass exposure fields during four seasons in the year 2016. Note the different maximum values in different seasons. All units are µg m-3. DJF = December January February; JJA = June July August; MAM = March April May; SON = September October November.

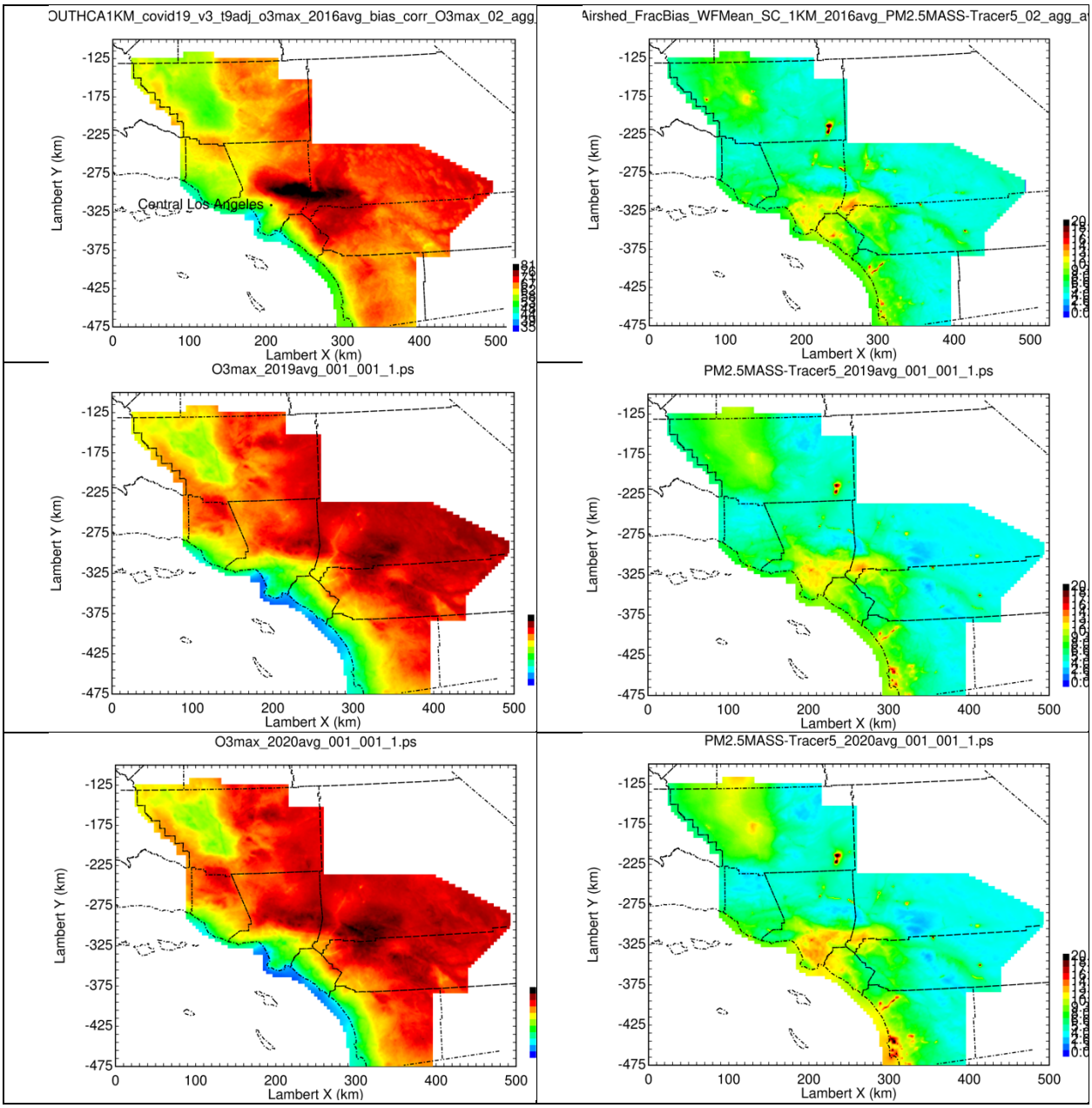

Figure A13. Comparison of exposure concentrations in 2016, 2019, and 2020. Left column: annual average of daily 1-hr maximum O_3 concentrations (ppb). Right column: annual average of $PM_{2.5}$ mass concentrations ($\mu g/m^3$) from all sources except biomass combustion. Results for 2016 are displayed in the top row; results for 2019 are displayed in the center row, and results for 2020 are displayed in the bottom row.

REFERENCES

Daily Land Use Regression Models Development

1. Su, J. G.; Meng, Y.-Y.; Chen, X.; Molitor, J.; Yue, D.; Jerrett, M. Predicting differential improvements in annual pollutant concentrations and exposures for regulatory policy assessment. *Environ Int* 2020, *143*, 105942.

2. Legendre, P. Spatial autocorrelation: trouble or new paradigm? *Ecology* 1993, *74* (6), 1659-1673.

3. Kanaroglou, P. S.; Jerrett, M.; Morrison, J.; Beckerman, B.; Arain, M. A.; Gilbert, N. L.; Brook, J. R. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment* 2005, *39* (13), 2399-2409.

4. Beckerman, B. S.; Jerrett, M.; Serre, M.; Martin, R. V.; Lee, S.-J.; van Donkelaar, A.; Ross, Z.; Su, J.; Burnett, R. T. A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States. *Environmental Science & Technology* 2013, *47* (13), 7233-7241. DOI: 10.1021/es400039u (accessed 2013/08/14). Su, J. G.; Jerrett, M.; Meng, Y. Y.; Pickett, M.; Ritz, B. Integrating smartphone-based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. *Sci Total Environ* 2015, *506*, 518-526. DOI: 10.1016/j.scitotenv.2014.11.022.

5. Jones, L.; Demirkaya, M.; Bethmann, E. Global value chain analysis: concepts and approaches. *J. Int'l Com. & Econ.* 2019, 1.

Chemical Transport Model

1. Kleeman MJ, Cass GR, Eldering A. Modeling the airborne particle complex as a source-oriented external mixture. *Journal of Geophysical Research: Atmospheres*. 1997;102(D17):21355-21372. doi:10.1029/97JD01261

2. Hu J, Zhang H, Ying Q, Chen SH, Vandenberghe F, Kleeman MJ. Long-term particulate matter modeling for health effect studies in California – Part 1: Model performance on temporal and spatial variations. *Atmospheric Chemistry and Physics*. 2015;15(6):3445-3461. doi:10.5194/acp-15-3445-2015

3. Yu X, Venecek M, Hu J, et al. *Sources of Airborne Ultrafine Particle Number and Mass Concentrations in California*. Aerosols/Atmospheric Modelling/Troposphere/Chemistry (chemical composition and reactions); 2018. doi:10.5194/acp-2018-832

4. Ying Q, Fraser MP, Griffin RJ, Chen J, Kleeman MJ. Verification of a source-oriented externally mixed air quality model during a severe photochemical smog episode. *Atmospheric Environment*. 2007;41(7):1521-1538. doi:10.1016/j.atmosenv.2006.10.004

5. Ying Q, Lu J, Allen P, Livingstone P, Kaduwela A, Kleeman M. Modeling air quality during the California Regional PM10/PM2.5 Air Quality Study (CRPAQS) using the UCD/CIT source-oriented air quality model – Part I. Base case model results. *Atmospheric Environment*. 2008;42(39):8954-8966. doi:10.1016/j.atmosenv.2008.05.064

6. Ying Q, Lu J, Kaduwela A, Kleeman M. Modeling air quality during the California Regional PM10/PM2.5 Air Quality Study (CPRAQS) using the UCD/CIT Source Oriented Air Quality Model – Part II. Regional source apportionment of primary airborne particulate matter. *Atmospheric Environment*. 2008;42(39):8967-8978. doi:10.1016/j.atmosenv.2008.05.065

7. Carter WPL, Heo G. Development of revised SAPRC aromatics mechanisms. *Atmospheric Environment*. 2013;77:404-414. doi:10.1016/j.atmosenv.2013.05.021

8. Hu XM, Zhang Y, Jacobson MZ, Chan CK. Coupling and evaluating gas/particle mass transfer treatments for aerosol simulation and forecast. *Journal of Geophysical Research: Atmospheres*. 2008;113(D11). doi:10.1029/2007JD009588

9. Nenes A, Pandis SN, Pilinis C. ISORROPIA: A New Thermodynamic Equilibrium Model for Multiphase Multicomponent Inorganic Aerosols. *Aquatic Geochemistry*. 1998;4(1):123-152. doi:10.1023/A:1009604003981

10. Carlton AG, Bhave PV, Napelenok SL, et al. Model Representation of Secondary Organic Aerosol in CMAQv4.7. *Environ Sci Technol*. 2010;44(22):8553-8560. doi:10.1021/es100636q

11. Hong SY, Noh Y, Dudhia J. A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes. *Monthly Weather Review*. 2006;134(9):2318-2341. doi:10.1175/MWR3199.1

12. Xiu A, Pleim JE. Development of a Land Surface Model. Part I: Application in a Mesoscale Meteorological Model. *Journal of Applied Meteorology and Climatology*. 2001;40(2):192-209. doi:10.1175/1520-0450(2001)040<0192:DOALSM>2.0.CO;2

13. Su L, Fung JCH. Sensitivities of WRF-Chem to dust emission schemes and land surface properties in simulating dust cycles during springtime over East Asia. *Journal of Geophysical Research: Atmospheres*. 2015;120(21):11,215-11,230. doi:10.1002/2015JD023446

14. National Center for Health Statistics, United States Department of Health and Human Services (US DHHS), US Centers for Disease Control (CDC). Compressed Mortality File 1999-2013 with ICD-10 Codes on the CDC WONDER Online Database, in the current release for years 1999 - 2013, is compiled from: CMF 1999-2013, Series 20, No. 2S, 2014. Published 2014. Accessed January 19, 2022. https://www.cdc.gov/nchs/data_access/cmf.htm

15. Hu J, Zhang H, Chen SH, et al. Predicting Primary PM2.5 and PM0.1 Trace Composition for Epidemiological Studies in California. *Environ Sci Technol*. 2014;48(9):4971-4979. doi:10.1021/es404809j

16. McDonald BC, McBride ZC, Martin EW, Harley RA. High-resolution mapping of motor vehicle carbon dioxide emissions. *Journal of Geophysical Research: Atmospheres*. 2014;119:5283-5298. doi:10.1002/2013jd021219

17. Brondfield MN, Hutyra LR, Gately CK, Raciti SM, Peterson SA. Modeling and validation of on-road CO2 emissions inventories at the urban regional scale. *Environmental Pollution*. 2012;170:113-123. doi:10.1016/j.envpol.2012.06.003

18. California Air Resources Board. California Air Resources Board SIP 2016 Emissions Projection Data. Accessed December 7, 2020. https://www.arb.ca.gov/app/emsinv/2017/emssumcat_query.php?F_YR=2012&F_DIV=-4&F_SEASON=A&SP=SIP105ADJ&F_AREA=CA#AREAWIDE

19. Almaraz M, Bai E, Wang C, et al. Agriculture is a major source of NOx pollution in California. *Science Advances*. 2018;4(1). doi:10.1126/sciadv.aao3477

20. Kleeman M, Anikender K, Abhishek D. *Investigative Modeling of PM2.5 Episodes in the San Joaquin Valley Air Basin during Recent Years*. California Air Resources Board; 2019.

21. Guenther AB, Jiang X, Heald CL, et al. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geoscientific Model Development*. 2012;5(6):1471-1492. doi:10.5194/gmd-5-1471-2012

22. Giglio L, Randerson JT, van der Werf GR. Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4). *Journal of Geophysical Research: Biogeosciences*. 2013;118(1):317-328. doi:10.1002/jgrg.20042

23. Paugam R, Wooster M, Freitas S, Val Martin M. A review of approaches to estimate wildfire plume injection height within large-scale atmospheric chemical transport models. *Atmospheric Chemistry and Physics*. 2016;16(2):907-925. doi:10.5194/acp-16-907-2016

24. van der Werf GR, Randerson JT, Giglio L, et al. Global fire emissions estimates during 1997–2016. *Earth System Science Data*. 2017;9(2):697-720. doi:10.5194/essd-9-697-2017

25. Hays MD, Fine PM, Geron CD, Kleeman MJ, Gullett BK. Open burning of agricultural biomass: Physical and chemical properties of particle-phase emissions. *Atmospheric Environment*. 2005;39(36):6747-6764. doi:10.1016/j.atmosenv.2005.07.072