Title
Exploring relationships between outdoor air particulate-associated polycyclic aromatic hydrocarbon and PM2.5: A case study of benzo(a)pyrene in California metropolitan regions

Permalink
https://escholarship.org/uc/item/6z31d6qm

Journal
Atmospheric Environment, 41(27)

Authors
Lobscheid, Agnes B.
McKone, Thomas E.
Vallero, Daniel A.

Publication Date
2007-09-01
Exploring relationships between outdoor air particulate-associated polycyclic aromatic hydrocarbon and PM2.5: A case study of benzo(a)pyrene in California metropolitan regions

Agnes B. Lobscheid∗a, Thomas E. McKonea,b, and Daniel A. Vallero c

a Indoor Environment Department, Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory MS90R3058, Berkeley, CA 94720, USA
b Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720, USA
c U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, North Carolina 27711, USA

*Corresponding author. Tel: +1-510-486-6945; Fax: +1-510-486-6658. Email address: ablobscheid@lbl.gov

Abstract
Polycyclic aromatic hydrocarbons (PAHs) and particulate matter (PM) are co-pollutants emitted as by-products of combustion processes. Convincing evidence exists for PAHs as a primary toxic component of fine PM (PM_{2.5}). Because PM_{2.5} is listed by the US EPA as a “Criteria Pollutant”, it is monitored regularly at sites nationwide. In contrast, very limited data is available on measured ambient air concentrations of PAHs. However, between 1999-2001, ambient air concentrations of PM_{2.5} and benzo(a)pyrene (BaP) are available for California locations. We use multivariate linear regression models (MLRM s) to predict ambient air levels of BaP in four air basins based on reported PM_{2.5} concentrations and spatial, temporal and meteorological variables as variates. We obtain an $R^2$ ranging from 0.57-0.72 among these basins. Significant variables (p<0.05) include the average daily PM_{2.5} concentration, wind speed, temperature and relative humidity, and the coastal distance as well as season, and holiday or weekend. Combining the data from all sites and using only these variables to estimate ambient BaP levels, we obtain an $R^2$ of 0.55. These $R^2$-values, combined with analysis of the residual error and cross validation using the PRESS-statistic, demonstrate the potential of our method to estimate reported outdoor air PAH exposure levels in metropolitan regions. These MLRMs provide a first step towards relating outdoor ambient PM_{2.5} and PAH concentrations for epidemiological studies when PAH measurements are unavailable, or limited in spatial coverage, based on publicly available meteorological and PM_{2.5} data.

Keywords: toxic air pollutants, particulate matter, regression models, combustion sources
1. Introduction

Environmental decisions must frequently be made without sufficient data on exposure. In particular, data for toxic air pollutant exposures are usually less reliable than data for criteria pollutants such as ozone and particulate matter (PM). For example, the spatial coverage, intensity, and confidence about PM exposures are much greater than those for polycyclic aromatic hydrocarbons (PAHs). Since it is unlikely that toxic air pollutant data in general, and PAH data specifically, will substantially improve in the next decade, indirect means of estimating exposures to air toxics are needed. Our research demonstrates one such approach.

PAHs and PM are co-pollutants emitted as by-products of combustion processes (Fraser et al. 1998; Guo et al. 2003; Kaupp and McLachlan 2000; Kiss et al. 1998; Leotz-Gartziandia et al. 2000; Vardar and Noll 2003; Hou et al., 2006; Spencer et al., 2006). Convincing evidence exists that PAHs, such as benzo(a)pyrene (BaP), are significant toxic components of the fine fraction of PM with aerodynamic diameters less than 2.5 µm (PM$_{2.5}$) leading to adverse human health risks (Adonis and Gil 2000; Dejmek et al. 2000; Deng et al. 2006; Massolo et al. 2002; Pohjola et al. 2003; Strandell et al. 1994). Major sources of atmospheric emissions of PM$_{2.5}$ and particulate-bound PAHs are gasoline and diesel powered motor vehicles and residential heating, primarily wood combustion (Bjorseth and Ramdahl 1985; Schauer and Cass 2000; Zheng et al. 2002). Around 7 million metric tons of PM$_{2.5}$ (US EPA 2000) and between 10,000-30,000 metric tons of associated particle-phase PAHs (Baek et al. 1991; US EPA 1998a) are emitted to the atmosphere annually in the US.

The important chemical reaction steps leading to the formation of PAHs and PM$_{2.5}$ involve the oxidation of benzene, the formation of cyclopentadienyl radicals, and reaction with C$_3$H$_3$ radicals (Richter and Howard 2000). PAHs can subsequently form soot via hydrogen abstraction and acetylene-addition (Richter and Howard 2000). Larger molecular-weight PAHs, such as BaP, are semivolatile organic compounds, with vapor pressures between $10^{-2}$ and $10^{-5}$ kilopascals. They are primarily particle-associated, and have been measured on archived PM2.5 filters (Pleil et al. 2004). BaP and other PAHs partition by adsorbing onto particle surfaces or absorbing into an organic material on the particle. This partitioning has been studied and characterized extensively (Harner and Bidleman 1998; Lohman et al. 2000; Mader and Pankow 2002; Odum et al. 1994; Offenberg and Baker 2002).
Because both PM$_{10}$ and PM$_{2.5}$ are classified as criteria pollutants by the US EPA (1997), they are monitored regularly at sites nationwide. Outdoor air PM$_{2.5}$ concentrations are available through the US EPA’s Air Quality Subsystem (AQS; formerly the Aerometric Information Retrieval System or AIRS (US EPA 2004). Since 1999, over a thousand US EPA and affiliated monitoring stations have collected data on ambient air PM$_{2.5}$. Of these, 93 stations are in California. In contrast however, measured outdoor air concentrations of PAHs are extremely limited. Relating outdoor air concentrations of primarily particulate-associated PAHs, such as BaP, to available PM$_{2.5}$ concentrations offers an important opportunity to characterize levels of human exposure where measured outdoor PAH concentrations are lacking. Of the limited studies reported in the literature relating particulate-associated PAHs with PM, the majority report poor ($r < 0.6$) Pearson product-moment correlation coefficients ($r$) between BaP and coarse PM, or PM$_{10}$ (Lewis et al. 1995; Sakai et al. 2002).

The current literature identifies key temporal and meteorological factors that influence both PAH and PM$_{2.5}$ outdoor air concentrations. This suggests the need to further explore these factors for predictive relationships associating PAH concentrations and PM$_{2.5}$ levels. The photochemical degradation of PAHs is a function of particle composition and meteorological factors, such as humidity (Kamens et al. 1998) and temperature (Dimashki et al. 2001). Furthermore, temporal parameters, such as day of week (weekday/weekend/holiday), can impact PM and PAH levels, because of the prevalence of certain activities, such as residential wood combustion and motor vehicle activity.

Our objectives in this paper are to investigate and evaluate the reliability of daily (24-hour) average ambient air concentrations of PM$_{2.5}$, adjusted for spatial, temporal and daily averaged meteorological parameters, to characterize ambient air particulate-associated PAH, specifically BaP, concentrations.

2. Materials and Methods

Our analysis is based on co-located measurements of outdoor air BaP concentrations and PM$_{2.5}$ concentrations measured throughout California by the California Air Resources Board (CARB), and reported through the AQS. These reported BaP concentrations are measured by HPLC techniques from a portion of PM$_{10}$ filters (CARB 1998), and will henceforth be referred to as $BaP_{10}^{\text{pt}}$. PM$_{2.5}$ concentrations are determined by gravimetric analysis (CARB 2002). The AQS data set of concentrations reported between January 1999-December 2001 is available for
four air basins in California—the San Francisco Bay (SF Bay), San Joaquin Valley (SJ Valley), South Coast, and San Diego air basins. After 2001, PAHs are not reported for California. Although $BaP_{10 \text{upt}}$ concentrations are reported as early as January 1988 for some of these air basins, corresponding PM$_{2.5}$ levels are not recorded before 1999.

We obtain daily (24-hour average) meteorological data corresponding to the AQS sites in each basin from the National Climate Data Center (NCDC/NOAA 2003). Because NCDC meteorological data are not available at a sufficiently close distance for two of the four SF Bay AQS sites: Concord (37.94 latitude and -122.02 longitude), and Fremont (37.54 latitude and –121.96 longitude), we use meteorological data from the Bay Area Air Quality Management District (BAAQMD). Figure 1 displays the co-located AQS $BaP_{10 \text{upt}}$ and PM$_{2.5}$ sites with corresponding publicly available meteorological data.

We first develop initial descriptive statistics of $BaP_{10 \text{upt}}$ and PM$_{2.5}$ concentrations in each of the four air basins individually at each site, and for all sites within a basin. These include the percent non-detects (NDs) and simple Pearson product-moment correlation coefficients (r) between $BaP_{10 \text{upt}}$ and PM$_{2.5}$, and the relative mass concentrations of $BaP_{10 \text{upt}}$ to PM$_{2.5}$.

Next, we run multivariate linear regression models (MLRMs) in a stepwise fashion, using both Microsoft Excel and S-PLUS (Mathsoft Inc. 2000), to find the “best-fit” MLRM that adjusts the relationship between the BaP concentrations ($BaP_{\text{est}}$) and PM$_{2.5}$ concentrations for spatial, temporal and surface meteorological variables. The specific meteorological and temporal variables we adjust PM$_{2.5}$ concentrations in our MLRM are:

1) average daily (24-hour average) temperature ($T$) [$^\circ$C and Kelvin, K];
2) average daily (24-hour average) wind speed ($WS$) [m/s];
3) relative humidity (24-hour average) ($RH$) [%];
4) weekend or holiday (binary);
5) season (winter is December-February; spring is March-May; summer is June-August; and autumn is September-November);
6) rain (binary).

We run separate analysis to test for a season binary variable for each of four seasons, and another set using a binary variable for ‘winter or fall’ only. Additionally, because RH is not reported in the NCDC unedited climate data, we calculate it (Eq.S-1 in Supplementary Material).
Because population densities tend to increase around coastal areas, PM emissions in California tend to rise sharply at the coastline where prevailing wind patterns bring clean ocean air against dense traffic. Thus, we expect some ‘aging’ with respect to the accumulation of BaP on particles as one moves east in a California air basin (Miguel et al. 2004). We therefore include the westward distance [km] to the coast to account for this potential effect. Lastly, we explore whether there are any site-specific effects on particle composition or other meteorological phenomenon that are not characterized by our other explanatory variables by including a dummy variable for each site within a basin to explain the variation in $BaP_{10}^{rp}$ levels.

We investigate the relationship between $BaP_{10}^{rp}$ and PM$_{2.5}$ adjusted for the meteorological and temporal variables based on the following MLRM procedure:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \text{error}$$

(Eq. 1)

where $y_i$ are the observed $BaP_{10}^{rp}$ [ng/m$^3$]; $\beta_0$ is the intercept; $\beta_1$ to $\beta_n$ are the coefficients associated with the continuous and any/all of the binary and categorical variables $x_1$, $x_2$, $\ldots$, $x_n$. Our “best-fit” MLRM has:

1) Normal distribution of residuals with constant variance; and

2) residual errors are independent of the explanatory variables.

To assess the precision of the final “best-fit” MLRM to estimate BaP concentrations in each of the four air basins, we compare the standard error of the estimate (SEE) and validate our models using the Predictive Residual Sum of Squares (PRESS). These statistics allow us to evaluate the reliability and the validity of our MLRMs. The SEE is computed as (Hamburg 1970):

$$\text{SEE of } \hat{y}_i = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

(Eq. 2)

where $y_i$ are the observed ($BaP_{10}^{rp}$) and $\hat{y}_i$ are the predicted outcomes ($BaP_{est}^{est}$ [ng/m$^3$]) respectively, based on our “best-fit” MLRM; and $n$ is the sample size of corresponding co-located PM$_{2.5}$ and $BaP_{10}^{rp}$ measurements.

Although a small SEE and an R$^2$ near 1 are necessary conditions for a valid MLRM, these statistics sometimes fail to fully account for chance correlations. Therefore, we apply the PRESS statistic to assess the validity of the regression models. With cross validation we can simulate a validation set by creating multiple modified observations from the existing data set (taking away each observation only once) and developing a model for each reduced data set. The response
values, $BaP_{10}^{rpt}$, of the deleted observations are then predicted from the model and the squared differences between predicted and ‘actual’ values are summed into the PRESS statistic, calculated as (Wold, 1991):

$$PRESS = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1 - h_{ii})^2} \quad \text{(Eq. 3)}$$

where $y_i$ and $\hat{y}_i$ are defined above (Eq. 2) and $h_{ii}$ are the diagonal elements of the “hat” matrix, calculated as:

$$H = X(X'X)^{-1} X' \quad \text{(Eq. 4)}$$

Where $X$ is a two-column and n-row matrix in which the first column is filled with 1’s corresponding to the intercept ($b$) and the second column are the n predictors, corresponding to values of the sum of the coefficients on the $x_n$ variables; $X'$ refers to the transpose of $X$.

According to Wold (1991), if $PRESS<SSY$ (the sum of the squared observed responses, $y_i$, or $BaP_{10}^{rpt}$), the model predicts better than chance.

3. Results and Discussion

3.1 $BaP_{10}^{rpt}$ Concentrations

Overall, more than 60% of the $BaP_{10}^{rpt}$ concentrations reported in each basin are below the LOD (0.05 ng/m$^3$), resulting in a bimodal distribution of $BaP_{10}^{rpt}$. As shown in Table 1, the majority of these non-detects (NDs) are in spring and summer. Therefore, $BaP_{est}$ values from our MLRM analyses best characterize autumn and winter activities, such as residential wood combustion. $BaP_{10}^{rpt}$ values $\geq$LOD in the four air basins are similarly distributed (see Supplementary Material), with peak concentrations between 0.1-0.5 ng/m$^3$.

In autumn, ANOVA of the $BaP_{10}^{rpt}\geq$LOD indicates that the differences in mean $BaP_{10}^{rpt}$ among sites within each basin are not significant (p>0.05). When NDs are included, there is a significant (p<0.05) difference between the mean $BaP_{10}^{rpt}$ at sites in the South Coast Basin in the autumn. This is most likely due to the wider range in the proportion of NDs (i.e., 33% at Burbank, and 73% at Simi Valley). However, in the winter, $BaP_{10}^{rpt}$ are significantly (p<0.05) different in each basin, regardless of whether or not NDs are included. This difference in the mean concentrations within each basin is most likely due to varying source-activity levels, e.g., residential wood combustion. In the Supplementary Material we present the $BaP_{10}^{rpt}$ distributions (boxplots) for each site in each basin for autumn and winter seasons.

3.2 PM$_{2.5}$ Concentrations
All reported PM$_{2.5}$ concentrations are $>$ LOD. We present distributions (boxplots) of autumn and winter PM$_{2.5}$ concentrations in the Supplementary Material. Similar to winter $BaP_{10}^{rpt}$, ANOVA indicates that the mean PM$_{2.5}$ concentrations are significantly different ($p<0.05$) between sites within each respective basin for the winter. However, unlike $BaP_{10}^{rpt}$ values, the autumn mean PM$_{2.5}$ concentrations are also significantly different ($p<0.05$) within each basin, except in San Diego. The distributions (histograms) of PM$_{2.5}$ concentrations within a basin are presented in the Supplementary Material.

### 3.3 Relative mass of $BaP_{10}^{rpt}$ to PM$_{2.5}$

The logarithm of the relative mass ratios of $BaP_{10}^{rpt}$ ($\geq$ LOD) to PM$_{2.5}$, or log($BaP_{10}^{rpt}$/PM$_{2.5}$), in the four basins are similarly distributed (Figure 2), indicating similar relationships between $BaP_{10}^{rpt}$ and PM$_{2.5}$ levels. Further, the log($BaP_{10}^{rpt}$/PM$_{2.5}$) at sites within the SF Bay and SJ Valley basins are similarly distributed (Figure 2). At each site within these two basins, the mean and medians of the logarithm of the log($BaP_{10}^{rpt}$/PM$_{2.5}$) are on the order of -4.9. However, the means and medians of the log($BaP_{10}^{rpt}$/PM$_{2.5}$) at each site within the South Coast basin, are slightly lower. Figure 2 shows that there are significant differences in the site-specific log($BaP_{10}^{rpt}$/PM$_{2.5}$) in the South Coast basin. For the four westernmost sites in the South Coast, the mean and median are close to -5.0. Additionally, the mean and median log($BaP_{10}^{rpt}$/PM$_{2.5}$) at the two easternmost sites in the South Coast, i.e., Fontana and Rubideux, are the lowest of all the sites in California (mean and median of -5.4 at Fontana and -5.2 at Rubideux). Based on these concentrations, the lower log($BaP_{10}^{rpt}$/PM$_{2.5}$) at these two sites appear to be due to the relatively higher PM$_{2.5}$ concentrations in the autumn, instead of lower $BaP_{10}^{rpt}$ concentrations (see Supplementary Material). In contrast, the lower log($BaP_{10}^{rpt}$/PM$_{2.5}$) in the winter at these sites is due to lower $BaP_{10}^{rpt}$ values relative to other South Coast sites (except Simi Valley) and not higher PM$_{2.5}$ concentrations.

At the two San Diego sites, the median (and mean) log($BaP_{10}^{rpt}$/PM$_{2.5}$) mass ratios are the highest of all the sites in California, i.e., -4.8 (-4.9) at Chula Vista and -4.5 (-4.6) at El Cajon. When compared to other sites, this appears due to relatively higher $BaP_{10}^{rpt}$ values in the autumn and winter and correspondingly lower PM$_{2.5}$ concentrations.

### 3.4 Unadjusted correlations ($r$) and $R^2$ between $BaP_{10}^{rpt}$ and PM$_{2.5}$

Overall, we obtain the highest correlation coefficients ($r$) using natural-logarithmic (ln) transformed $BaP_{10}^{rpt}$ and PM$_{2.5}$ concentrations, as opposed to the non-transformed
concentrations. This relationship generally applies on a per-site basis and for all sites combined within a basin (see Supplementary Material). Only four sites: San Jose (SF Bay basin), Simi Valley (South Coast basin), and Bakersfield and Fresno (SJ Valley basin) have \( r > 0.7 \), indicating a relatively strong linear correlation between the two pollutant concentrations. Small negative correlations at Fontana (South Coast basin; in both the ln-transformed and non-transformed concentrations), and Chula Vista (San Diego basin; in only the non-transformed concentrations) suggests unique features in either PM\(_{2.5}\) composition and/or BaP and/or PM\(_{2.5}\) sources in these areas.

3.5 The “best-fit” MLRMS

Figure 3(a-d) displays our “best-fit” MLRM for each of the four air basins. We exclude \( BaP_{10}^{\text{rpt}} < \text{LOD} \) because if they are included, our MLRM violates the assumptions of (i) linearity, (ii) constant variance of the residuals, and (iii) errors that are independent of the explanatory variables. We also found that a natural-logarithmic transformation of the continuous variables did not violate these assumptions and produced a higher \( R^2 \) than leaving the continuous variables untransformed. Further, we report model results based on absolute temperature [K], because this consistently provides higher \( R^2 \)-values and a significant intercept (\( p < 0.01 \)), than [°C]. Additionally, because most basins have few, if any, spring measurements, and because autumn is not significant (\( p > 0.05 \)) in all basins for explaining BaP concentrations (but winter and summer are), we rely on the binary ‘fall or winter’ season variable. Compared with the MLRMs including dummy variables for each season with measurements available, this “fall or winter” variable does not decrease the \( R^2 \) nor increase the SEE appreciably.

As shown in Table 2, in each of the four basins we obtain an \( R^2 \) of at least 0.57, or explain at least 57% of the variation in \( \ln(BaP_{10}^{\text{rpt}}) \) concentrations by the MLRMs. The linear expressions of the independent variables in our “best-fit” MLRMs for the four air basins are listed in Table 2. The \( \ln(\text{PM}_{2.5}) \) is positively related to \( \ln(BaP_{10}^{\text{rpt}}) \) and is significant at \( p < 0.01 \) in each basin, except in San Diego (\( p < 0.05 \)). \( \ln(T) \) [K] is also significant (\( p < 0.01 \)) and inversely related to \( \ln(BaP_{10}^{\text{rpt}}) \) concentrations in all four basins. This effect of temperature on particle associated PAH levels has been reported elsewhere and is likely due to increased partitioning or sorption with decreasing temperature (Miguel et al. 2004; Motelay Massei et al. 2003). The natural-logarithm of the other continuous meteorological variables, RH and WS are also inversely related to \( \ln(BaP_{10}^{\text{rpt}}) \) concentration.
Interestingly, the binary variable indicating a weekend or weekday measurement is significant in our “best-fit” MLRMs at p<0.01 for the South Coast and SJ Valley air basin, but is not significant (p>> 0.10) for the SF Bay and the San Diego basins. This may be due partly to the higher number of weekend measurements reported in the South Coast (47 out of 140, or 34%), and the SJ Valley (22 out of 77, or 29%) basins, as compared to the SF Bay basin (only 9 out of 50, or 18%). However, because 43% of the 28 measurements were taken on the weekend in San Diego, another reason is needed to explain why weekend is not significant in the San Diego basin. Perhaps seasonal activities, such as residential wood combustion, are less prevalent there than in the other basins.

Further evaluation is needed to determine why WS is a significant variable (p<0.05) only in the MLRM for the South Coast basin. This might be due to the fact that ln(WS) is also significantly (p<0.01) correlated with the ‘fall or winter’ binary variable (r= -0.4) in the South Coast Basin. Correlation analysis reveals that WS is significantly correlated (p<0.01) with $BaP_{10}^{rpt}$ levels and PM2.5 in all basins, with the exception of San Diego basin for WS and $BaP_{10}^{rpt}$. We can rule out a correlation between wind speed and temperature because we found they are not correlated (R²= 0.007) using daily meteorological data from the NCDC. In our “best-fit” MLRMs for each basin, ln(WS) is significant at p=0.05 only in the South Coast model and ”fall or winter” is significant at p<0.01 in the SF Bay and at p<0.10 in the South Coast and San Diego basins.

Because this is the first attempt to characterize outdoor BaP levels based on our set of explanatory variables we retain all variables, with the exception of the site-specific dummy variable, that are significant (at p≤0.05) in at least one of the air basin MLRMs. Rain was the only variable that did not fulfill this requirement and this is corroborated by previously published research by Panther et al. (1999) who report no significant correlation between rainfall levels and TSP or PAH concentrations.

Although the site-specific dummy variable for the southernmost site (Bakersfield) in the SJ Valley basin is a borderline significant (p=0.05) estimator of $BaP_{10}^{rpt}$ concentrations in the final MLRM, we exclude it from the final model because our objective is to ideally formulate general relationships that can be applied throughout a particular region or air basin, not at particular sites. By doing so, the multiple R²'s decreases from 0.60 to 0.57. Despite the large distance between the three sites in the SJ Valley, relative to distances between sites in the South
Coast and SF Bay region, we obtain a multiple R² in SJ Valley comparable to those obtained for South Coast and San Diego basins.

Additionally, the dummy variable for the Burbank site in the South Coast Basin is significant (p<0.05) in the MLRM and explains an additional 5% of the variation in ln(BaP₁₀ᵣᵣ) when added to our “best-fit” MLRM, suggesting there are unique source-activity or meteorological events in sub-regions of the South Coast basin. We recommend further research, outside the scope of this exploratory study, to determine any site-specific sources influencing this finding.

When we combine all the corresponding BaP₁₀ᵣᵣ and PM₂.₅ concentrations from the four basins to evaluate a MLRM to estimate BaP levels in metropolitan regions of California, we obtain an R² slightly lower (0.55) than the R² value for any of the individual basins. In the rightmost column of Table 2 we present the linear expression for the combined data set. Figure 4 provides the corresponding graphical presentation of the “best-fit” MLRM based on the seven significant variables from our basin MLRM analysis applied to the data from all sites in the four basins.

3.6 The reliability of our ‘best-fit’ MLRMs

The SEE of ln(BaPᵣᵣ) and the geometric spread implied, that is exp[SEE(ln(BaPᵣᵣ))], allows us to characterize the accuracy of our BaPᵣᵣ relative to BaP₁₀ᵣᵣ. The SEE of ln(BaPᵣᵣ) is fairly consistent among the four air basins (Table 3), ranging from 0.48 to 0.68. When we interpret exp(SEE) as the geometric error, we estimate the BaP₁₀ᵣᵣ are within a factor less than two among the four metropolitan air basins. That is, a factor of two captures one standard deviation of the normal distribution of the residuals in ln(BaPᵣᵣ). Among the specific locations, the one geometric standard deviation of the residuals is 1.6 for SF Bay and the South Coast, 1.7 for San Diego, and 2.0 for the SJ Valley basin. Additionally, the exp(SEE) for our MLRM combining data from all basins, is 1.9. As shown by Figures 3(a-d) and Figure 4, all the BaPᵣᵣ values are within the 95th percentile confidence interval of each of our MLRMs.

3.7 Cross Validation of our MLRMs

Because an independent set of data that has undergone strict quality assurance measures is not available at this time, we perform a cross validation of our MLRMs using a subset of our data and the PRESS statistic. First, we apply the MLRM based on combined data from all the sites, to each respective basin. This results in R²-values ranging from 0.41 (San Diego and SF Bay) to
0.56 (SJ Valley), as summarized in Table 3. Next, we apply the MLRM based on the highest $R^2$ (0.72) and lowest SEE (0.48), using data from the SF Bay basin. Based on this application, we obtain $R^2$-values ranging from 0.47 (San Diego) to 0.72 (SF Bay). These results, summarized in Table 3, suggest that the MLRMs based on the coefficients of the seven variables summarized in Table 2 from all sites are slightly better (lower SEE with comparable $R^2$) in explaining the variation in $\ln(BaP_{10}^{pp})$ levels, than the SF Bay MLRM. However, when the MLRM based on data from all sites is applied to SF Bay it performs rather poorly compared with the best-fit model for that region. This may be due to particular meteorology or source characteristics that affect the SF Bay region, and requires further research.

When we randomly select a subset of data (n= 64) from all of the sites (N=296), the resulting ratio of the PRESS statistic to the SSY is 0.11. According to Wold (1991), this indicates a nearly ”excellent model” (PRESS/SSY<0.1), leading us to believe the reliability of the “all-sites” MLRM is quite high. Presently, we consider the all-sites MLRM as a useful surrogate measure to estimate outdoor exposure levels of BaP in metropolitan regions of California where measurements are not available. Whether this MLRM is representative of metropolitan regions outside of California, requires further investigation.

4. Conclusions

Our research demonstrates that MLRMs can be useful when measurements are not available or are limited in spatial coverage, particularly in epidemiological studies relating outdoor ambient PM$_{2.5}$ and PAH concentrations to disease surveillance. Our models provide a first step in developing reliable proxy measures of PAH concentration using widely available PM$_{2.5}$ measurements and available meteorological data. There are uncertainties, however, both in the reliability of this method for regions outside of California and whether this approach is sufficiently robust for exposure studies. In the following paragraphs we describe and evaluate sources of uncertainty in our MLRM, including: model formulation, precision and accuracy of inputs and interpretation of results. We also summarize the relevance of our findings in the context of these uncertainties.

4.1 Model Formulation

We select a parsimonious set of variables to take into account both temporal, meteorological, and spatial factors that we believe do not “over-fit” the model nor result in variables that are randomly correlated with BaP concentrations. However, additional variables may increase the
precision of our MLRMs. Examples of such factors include solar radiation \([\text{W/m}^2]\) or hours of sunlight at a certain site to capture the rapid reaction and degradation of particle-associated BaP. However, these variables are not available for all sites and are likely explained by the \(\ln(T)\) variable. Similarly, atmospheric stability (reported by the BAAQMD, but not in the NCDC data) may better explain the variation in the \(\text{BaP}_{10\text{pt}}\) levels. This may be particularly relevant in the South Coast basin due to the impact of its low and varying inversion layer. The extent to which WS may serve as a proxy for atmospheric stability needs further examination, especially because WS is significant (\(p<0.05\)) only in the South Coast basin.

Including other chemical constituents to adjust ln(PM\(_{2.5}\)) concentrations may also increase the model precision. Hughes et al. (2000) demonstrated how particulate matter composition changes in the South Coast basin as one travels from Long Beach (coast) to Riverside (inland) due to the addition/accumulation of secondary ammonium nitrate and organics to the original sea-salt and primary carbon constituents. Kleeman et al. (1999) found that this phenomenon was the “single largest contribution to ambient PM2.5 concentrations” at the Riverside site. Unfortunately, sea-salt and background sulfate concentrations are not reported at the AQS sites and therefore cannot be tested in our model framework at this time.

### 4.2 Accuracy and quality assurance issues related to input data

Undoubtedly, model reliability increases considerably when input data pass quality assurance measures. The AQS \(\text{BaP}_{10\text{pt}}\) and PM\(_{2.5}\) concentrations were determined based on standard operating procedures intended to minimize the effects of sampling artifacts due to:

1) loss of PAH from particles on filters by volatilization;

2) sorption of PAHs onto filter; and

3) reaction of compounds with reagents during sampling.

In addition, AQS sites are part of the US EPA administered State and Local Air Monitoring Station (SLAMS) network. As such, the EPA’s Office of Air Quality Planning and Standards requires annual monitoring summaries of each SLAMS monitor, and additional details upon request. These quality assurance criteria strengthen the reliability of our MLRMs.

In contrast however, the NCDC unedited local climatological data were not yet quality assured. These data are simply collected by automated weather stations operated by the US government. In three instances, we had to limit the calculated RH to 100%, because we calculated over 100% based on \textbf{Eq.S-1} (Supplementary Material) and the reported station
pressure and wet bulb temperatures. However, because we use lognormally-transformed meteorological variables, the degree to which the lack of knowledge about the ‘true’ value of the meteorological data decreases the reliability of our “best-fit” MLRMs, is likely minimal.

4.3 Alternative modeling efforts to characterize outdoor air concentrations

To assess levels of HAPs in outdoor air, the US EPA has applied process-based transport models such as the Assessment System for Population Exposure Nationwide (ASPEN) in exposure studies (e.g., the Cumulative Exposure Project (US EPA 1998b; Woodruff et al. 1998) and the National Air Toxic Assessment for the Integrated Urban Air Toxics Strategy (US EPA 1999). Although ASPEN has a tendency to underestimate monitored concentrations (Pratt et al. 2000; US EPA 1998b), our “best-fit” per-basin, and “all sites” MLRMs have $R^2$-values (0.57-0.72, and 0.55, respectively) are comparable with the ASPEN. $R^2$-values of 0.59, 0.53, and 0.23 have been reported for ASPEN when it is applied to the Northeast, West and Central regions of the US (Rosenbaum et al. 1999).

Although our $R^2$-values are comparable, ASPEN and other process-based models tend to incorporate additional uncertainties associated with problem specification and the algorithms (such as reactive decay, deposition, etc) used to formulate the model. Process-based models also rely on estimated emissions, which have considerable uncertainty due to a lack of knowledge arising from missing, incomplete, or erroneous data. Therefore, we believe that the additional uncertainties generally associated with process-based models result in less reliable outdoor air concentration estimates than those we achieve with our “best-fit” MLRMs to estimate outdoor air BaP concentrations.

4.4 Relevance of our findings

Our results are encouraging in light of the scarcity of monitoring data available to characterize human exposure levels to toxic air pollutants, such as PAHs. Our “best-fit” MLRMs explain approximately the same amount of variation in $\ln(BaP_{est})$ concentrations (multiple $R^2 > 55\%$), in each of the four air basins. This is a considerable increase from the $R^2$-values of $\ln(BaP_{10,prts})$ against $\ln(\text{PM}_{2.5})$, which explain approximately 35% of the variation in both the SJ Valley and SF Bay basin, only 7% in the South Coast, and only 3% in the San Diego basin (see Supplementary Material).

There are limits to the representativeness of our results. For example, certain activities, such as coal combustion, are not carried out in California, and therefore our results may not
represent metropolitan regions with these types of emissions. Additionally, generalizing our MLRM results is uncertain due to unknown contributions from other combustion sources, such as automobiles without catalytic converters. Because the data we rely on for \( BaP_{10}^{\text{rpt}} \) is concentrated in the autumn and winter months, our MLRM derived estimates may tend to reflect ambient air concentrations due to residential wood combustion.

Our MLRM methods to estimate outdoor-air \( BaP \) levels based on surrogate PM\(_{2.5}\) concentrations can likely be applied toward characterizing levels of other particulate-associated PAHs, such as benzo(k)fluoranthene, benzo(ghi)perylene, indeno(1,2,3 cd)pyrene, and dibenz(a,h)anthracene, as well as other semivolatile organic compounds, such as many pesticides, PCBs, and dioxins. Our approach may therefore complement the research and exposure studies proposed as part of the human health tracking surveillance system in California (CPRC 2004). Because of the relatively similar features of the MLRMs of the respective metropolitan regions (\( R^2 \), significant variables) the authors lay the foundation for future studies and validation.

**Acknowledgements**

We thank Arpad Horvath, Robert Spear, and S. Katherine Hammond of University of California, Berkeley (UC-Berkeley), for their early reviews and Coe Owen of the US EPA, for air concentration data and Kurt Malone of the Bay Area Air Quality Management District for meteorological data. This research was supported in part by the US EPA National Exposure Research Laboratory through Interagency Agreement No. DW-988-38190-01-0, with the Lawrence Berkeley National Laboratory operated for the US Department of Energy (DOE) under Contract Grant No. DE-AC02-05CH11231 and by Cooperative Agreement Number U50/CCU922409-01 from the US Centers for Disease Control and Prevention (CDC). Additional funding assistance was provided by the University of California Toxic Substance and Research Teaching Program and the Russell M. Grossman Endowment from the School of Public Health, UC-Berkeley. Although the research described in this article has been funded and reviewed by the US EPA, it does not necessarily reflect the views of the federal government and no official endorsement should be inferred.
Disclaimer
This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

References
CARB (California Air Resources Board), 1998. Standard operating procedure for the determination of selected polycyclic aromatic hydrocarbons (PAH) in ambient air (SOP MLD 028).
CARB (California Air Resources Board), 2002. Standard operating procedure for the
determination of PM2.5 mass in ambient air by gravimetric analysis (SOP MLD 055): 2002
hydrocarbons and fine particles on pregnancy outcome. Environmental Health Perspectives
108(12):1159-64.
levels and cytotoxicity of PAHs and heavy metals in TSP and PM_{2.5} at an electronic waste
dependence, and relative reactivity of atmospheric polycyclic aromatic hydrocarbons.
data for organics. 5. C6-C22 nonpolar and semipolar aromatic compounds. Environmental
Science and Technology 32(12):1760-70.
hydrocarbons in urban air of Hong Kong. Atmospheric Environment 37:5307-5317.
Jovanovich, Inc.
partitioning of aromatic compounds in urban air. Environmental Science and Technology
aromatic hydrocarbons and fatty acids in PM_{2.5} aerosols in dust season in China. Atmospheric
Hughes L.S., Allen J.O., Bhave P., Kleeman M.J., Cass G.R., Liu D.Y., Fergenson D.P.,
crossing the Los Angeles Basin. Environmental Science and Technology 34(15):3058-68.

16

Kaupp H., McLachlan M.S., 2000. Distribution of polychlorinated dibenzo-p-dioxins and dibenzofurans (PCDD/Fs) and polycyclic aromatic hydrocarbons (PAHs) within the full size range of atmospheric particles. Atmospheric Environment 34:73-83.


Appendix

List of Nomenclature

AIRS: Aerometric Information Retrieval System
AQS: US EPA’s Air Quality Subsystem
ASPEN: Assessment System for Population Exposure Nationwide
BAAQMD: Bay Area Air Quality Management District
BaP: benzo(a)pyrene
BaP_{10}^{\text{rp}}: reported particle-associated (PM_{10}) BaP levels
BaP_{10}^{\text{est}}: estimated BaP levels
CARB: California Air Resources Board
EPA: US Environmental Protection Agency
ln: natural logarithm
LOD: limit of detection
MLD: Monitoring and Laboratory Division
MLRM: multivariate linear regression model
NCDC: National Climate Data Center
ND: non-detect
PAHs: polycyclic aromatic hydrocarbons
PM: particulate matter
PM_{2.5}: fine PM with aerodynamic diameter less than 2.5 µm
PM_{10}: fine PM with aerodynamic diameter less than 10 µm
PRESS: Predictive Residual Sum of Squares
r: Pearson product-moment correlation coefficient
R^2: multiple correlation coefficient
RH: relative humidity
SEE: standard error of the estimate
SF Bay: San Francisco Bay Area Air Basin
SJ Valley: San Joaquin Valley Air Basin
SLAMS: State and Local Air Monitoring Station
SOP: Standard Operating Procedure
SSY: sum of square of responses
T: average daily (24-hour) temperature
WS: wind speed
Figure Captions

Figure 1. Location of US EPA AQS sites with corresponding meteorological sites (BAAQMD, 2003 and NCDC, 2003) used in our MLRM analysis. Co-located PM$_{2.5}$ and $BaP_{10}^{rpt}$ concentrations and meteorological data are available for four AQS sites in the SF Bay air basin, all six sites in the South Coast basin, and all three sites in the SJ Valley basin. The air basin boundaries are roughly defined, but otherwise the figure is drawn to scale.

Figure 2. The logarithm of the relative mass of $BaP_{10}^{rpt}$ to PM$_{2.5}$ in the four air basins. Outliers are presented as “○”.

Figure 3. The “best-fit” MLRMs and 95th% confidence intervals shown for the a) SF Bay, b) SJ Valley, c) South Coast, and d) San Diego air basins. The form for the MLRMs is: ln(PM$_{2.5}$ [µg/m$^3$]) + ln(WS [m/s]) + ln(T [K]) + ln(RH [%]) + weekend/holiday + season + ln(coast distance [km]). Coefficients are provided in Table 2. These intervals are calculated by adding and subtracting the following from the $BaP^{est}$: $2 \times \exp\left(\frac{\text{stdev}(\text{ln}(BaP^{est}))}{\sqrt{N}}\right)$ where stdev is the standard deviation and $N$ is the size of the MLRM data set, given in Table 2.

Figure 4. “Best-fit” MLRM combining data from all four air basins in California, with 95th% confidence intervals. The form for the MLRMs is: ln(PM$_{2.5}$ [µg/m$^3$]) + ln(WS [m/s]) + ln(T [K]) + ln(RH [%]) + weekend/holiday + season + ln(coast distance [km]). Coefficients are provided in Table 2.
Table 1: The total sample size, n, from each air basin (not including the BaP_{10}\text{ng}<\text{LOD}) and the percent of BaP_{10}\text{ng}<\text{LOD}.

<table>
<thead>
<tr>
<th>Basin</th>
<th>Site</th>
<th>Site</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
<th>Summer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td></td>
<td>46-60%</td>
<td>8-16%</td>
<td>50-83%</td>
<td>89-100%</td>
</tr>
<tr>
<td>SF Bay</td>
<td>Fremont</td>
<td>17</td>
<td>60%</td>
<td>11%</td>
<td>75%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td>Concord</td>
<td>10</td>
<td>50%</td>
<td>8%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>San Francisco</td>
<td>13</td>
<td>46%</td>
<td>16%</td>
<td>83%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>San Jose</td>
<td>11</td>
<td>53%</td>
<td>11%</td>
<td>75%</td>
<td>100%</td>
</tr>
<tr>
<td>SJ Valley</td>
<td>Fresno</td>
<td>28</td>
<td>40%</td>
<td>0%</td>
<td>63%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Bakersfield</td>
<td>36</td>
<td>53%</td>
<td>5.6%</td>
<td>79%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Modesto</td>
<td>16</td>
<td>57%</td>
<td>17%</td>
<td>63%</td>
<td>88%</td>
</tr>
<tr>
<td>South Coast</td>
<td>Burbank</td>
<td>26</td>
<td>33%</td>
<td>5%</td>
<td>79%</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>Los Angeles</td>
<td>27</td>
<td>53%</td>
<td>5%</td>
<td>67%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>Long Beach</td>
<td>27</td>
<td>46%</td>
<td>0%</td>
<td>79%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>Rubidoux</td>
<td>25</td>
<td>38%</td>
<td>28%</td>
<td>89%</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>Fontana</td>
<td>29</td>
<td>40%</td>
<td>40%</td>
<td>58%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>Simi Valley</td>
<td>6</td>
<td>73%</td>
<td>33%</td>
<td>96%</td>
<td>100%</td>
</tr>
<tr>
<td>San Diego</td>
<td>Chula Vista</td>
<td>23</td>
<td>46%</td>
<td>5.6%</td>
<td>89%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>El Cajon</td>
<td>5</td>
<td>40%</td>
<td>0%</td>
<td>88%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 2
Outdoor air ln-BaP\textsuperscript{est} [ng/m\textsuperscript{3}] stepwise regression results (95\textsuperscript{th} percentile confidence limits). Unless otherwise indicated, coefficients ($\beta$s) are significant at $p<0.01$.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variable</th>
<th>SF Bay</th>
<th>SJ Valley</th>
<th>South Coast</th>
<th>San Diego</th>
<th>All Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>Intercept</td>
<td>195</td>
<td>184</td>
<td>137</td>
<td>229</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(123, 268)</td>
<td>(111, 255)</td>
<td>(98, 175)</td>
<td>(81, 377)</td>
<td>(151, 212)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>ln(PM\textsubscript{2.5} [\mu g/m\textsuperscript{3}])</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.1, 0.8)</td>
<td>(0.2, 0.7)</td>
<td>(0.2, 0.6)</td>
<td>(0.1, 1.5)</td>
<td>(0.3, 0.6)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>ln(T [K])</td>
<td>-34.1</td>
<td>-32.5</td>
<td>-23.9</td>
<td>-40.9</td>
<td>-32.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-46.5, -21.7)</td>
<td>(-44.9, -20.2)</td>
<td>(-30.6, -17.2)</td>
<td>(-14.8, -66.9)</td>
<td>(-37.4, -26.8)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>ln(WS [m/s])</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.3</td>
<td>1.0</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.6, 0.5)</td>
<td>(-0.7, 0.2)</td>
<td>(-0.6, -0.1)</td>
<td>(-0.2, 2.1)</td>
<td>(-0.4, -0.06)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>ln(RH [%])</td>
<td>-1.4</td>
<td>-0.7</td>
<td>-1.0</td>
<td>-1.1</td>
<td>-0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-2.4, -0.5)</td>
<td>(-1.6, 0.1)</td>
<td>(-1.2, -0.7)</td>
<td>(-1.8, -0.4)</td>
<td>(-1.0, -0.5)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>Weekend or holiday</td>
<td>-0.02</td>
<td>-0.4</td>
<td>-0.3</td>
<td>0.1</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.4, 0.4)</td>
<td>(-0.8, -0.01)</td>
<td>(-0.5, -0.1)</td>
<td>(-0.4, 0.6)</td>
<td>(-0.5, -0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>Fall or winter</td>
<td>0.8</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.4, 1.2)</td>
<td>(-0.3, 0.7)</td>
<td>(-0.01, 0.5)</td>
<td>(-0.2,1.2)</td>
<td>(-0.2, 0.6)</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>ln(coast distance [km])</td>
<td>0.1</td>
<td>0.2</td>
<td>-0.2</td>
<td>0.5</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.1, 0.2)</td>
<td>(-0.6, 0.9)</td>
<td>(-0.3, -0.04)</td>
<td>(0.05, 0.9)</td>
<td>(-0.04, 0.1)</td>
</tr>
</tbody>
</table>

$N = \ 50$ $^p$ $\quad 78$ $^p$ $\quad 140$ $\quad 28$ $\quad 296$

$R^2 = 0.72 \quad 0.57 \quad 0.57 \quad 0.65 \quad 0.55$

$\text{SEE} = 0.48 \quad 0.68 \quad 0.50 \quad 0.51 \quad 0.62$

The regression model is: $\beta_0 + \beta_1 \times \ln \ (\text{PM}_{2.5} \ [\mu g/m^3]) + \beta_2 \times \ln \ (T \ [K]) + \beta_3 \times \ln(WS \ [m/s]) + \beta_4 \times \ln \ (\text{RH} \ [%]) + \beta_5 \times \text{(Weekend or holiday)} + \beta_6 \times \text{season} + \beta_7 \times \ln(\text{coast distance} \ [\text{km}]) + \text{error}$.

\text{a-n p-value: } ^a \ 0.74; \ \ ^b \ 0.94; \ \ ^c \ 0.27; \ \ ^d \ 0.29; \ \ ^e \ p<0.10; \ \ ^f \ 0.15; \ \ ^g \ 0.14; \ \ ^h \ p<0.05; \ \ ^i \ 0.61; \ \ ^j \ 0.15; \ \ ^k \ 0.48$

\text{p: the final sample size has been decreased from that shown in Table 1 because of missing meteorological data.}
Table 3
R² and SEE-values resulting from applying the “best-fit” MLRM from three regions (with highest R², lowest SEE and from all sites combined MLRM; based on coefficients listed in Table 1) to other regions.

<table>
<thead>
<tr>
<th>Region MLRM</th>
<th>All Sites</th>
<th>SF Bay</th>
<th>SJ Valley</th>
<th>South Coast</th>
<th>San Diego</th>
<th>All Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R²</td>
<td>0.41</td>
<td>0.56</td>
<td>0.47</td>
<td>0.41</td>
<td>0.55</td>
</tr>
<tr>
<td>SF Bay</td>
<td>SEE</td>
<td>0.90</td>
<td>1.1</td>
<td>0.71</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>(highest R²)</td>
<td>R²</td>
<td>0.72</td>
<td>0.50</td>
<td>0.47</td>
<td>0.47</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>SEE</td>
<td>0.48</td>
<td>0.91</td>
<td>1.2</td>
<td>0.87</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Figure 1

North Central Data Center (NCDC) meteorological sites
× Environmental Protection Agency (EPA) Air Quality System (AQS) sites

San Francisco Bay Area Air Basin
San Joaquin Valley Air Basin
South Coast Air Basin
San Diego Air Basin

0 300 600 km
Figure 2
Figure 3
Figure 4