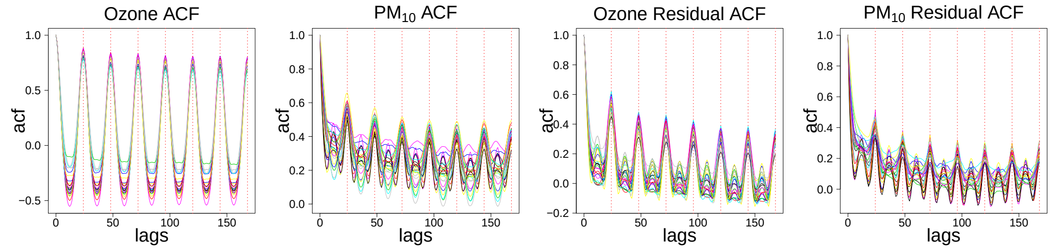
ONLINE SUPPLEMENT

Pollution State Modeling for Mexico City

Philip White, Alan E. Gelfand, Eliane R. Rodrigues, Guadalupe Tzintzun

1 | TIME-SERIES AND MEAN-VARIANCE DATA EXPLORATION

For preliminary examination of the temporal pattern of model residuals, we fit a model with site-specific regression coefficients effects for relative humidity and temperature on hourly ozone and PM$_{10}$ concentrations. Although phase alerts depend on averaged PM$_{10}$ levels and Mexican ambient air quality standards depend on averaged O_3 and PM$_{10}$ levels, we carry out all modeling and exploratory analyses on the hourly pollutant levels. In Figure 1, we supply the empirical autocorrelation function (ACF) for both pollutants and their model residuals for each site. It is evident that daily seasonality is very strong for both pollutants. However, the ACF has somewhat similar behavior across sites but also varies significantly across sites, suggesting the need for a hierarchical time-series specification. It should be noted that ACF plots serve as exploratory tools. However, they are not as useful for selecting specific lags (e.g. seasonality). To capture seasonality in the data, we consider models that use autoregressive (AR) terms of one day (24 hours) and one week (168 hours). These seasonal AR terms, in conjunction with AR terms of lower order, account for overall changes over the year in Figures 2b and 3b.



(a) O$_3$ ACF

(b) PM$_{10}$ ACF

(c) O$_3$ residual ACF

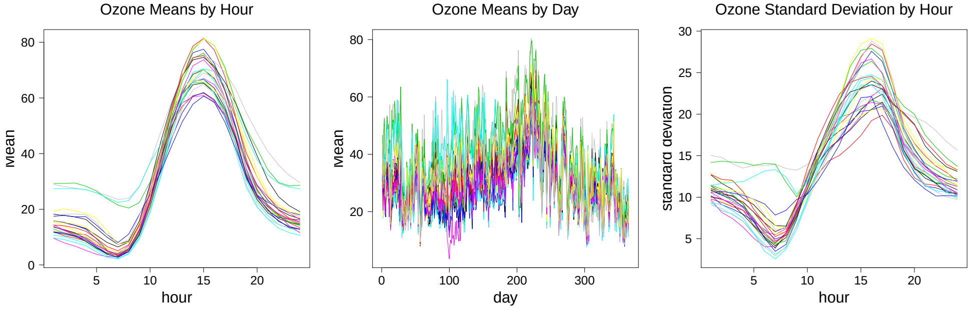
(d) PM$_{10}$ residual ACF

FIGURE 1 Site-specific autocorrelation functions for ozone and PM$_{10}$ for one week of lags.

We also examined partial ACF plots to gain insight into which autoregressive lags may be necessary in the model; however, we found them to be uninformative given the strong seasonality of the data. Ultimately, we use out-of-sample predictive performance to select the autoregressive structure for ozone and PM$_{10}$.

We plot site-specific means for both pollutants in two ways. We consider hourly means (Figures 2a and 3a), aggregated over the year, and daily means over the course of the year (Figures 2b and 3b). Figure 2a show that ozone concentrations generally peak around 4 pm (the warmest time of the day). In general, the highest ozone levels in

Mexico City occur during spring months and June. In 2017, May, June, and July have the highest average ozone levels (see Figure 2b). Figure 3a shows two daily peaks in PM_{10} concentrations corresponding to commuting hours; however, annual trends for PM_{10} are less clear in Figure 3b. We also plot standard deviation of ozone and PM_{10} concentrations as a function of hour of the day for all stations in Figures 2c and 3c. Note that there is strong correlation between the standard deviation and the mean for both pollutants (compare Figures 2a and 2c and compare Figures 3a and 3c). This correlation could be addressed through modeling in a variety of ways. First, and most simply, one could use a variance stabilizing transformation (VST) to address the correlation between the mean and variance (e.g. log, square-root, Box-Cox). Alternatively, we could use heteroscedastic models that specify variance directly as a function of hour or month.

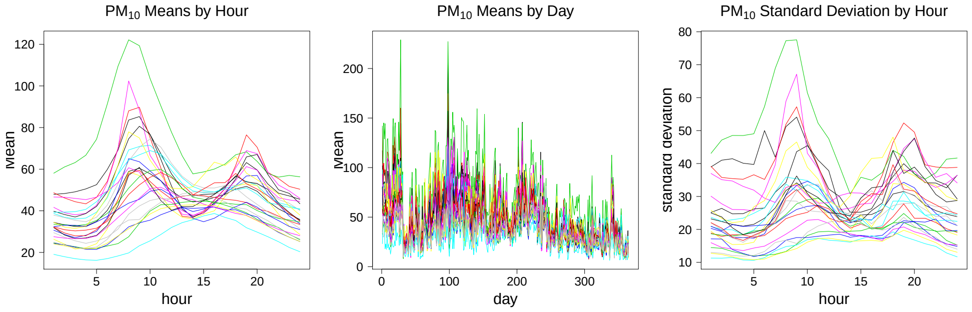


(a) Hourly ozone means

(b) Daily ozone means

(c) Hourly ozone std. dev.

FIGURE 2 Site-specific means by hour averaged over the year



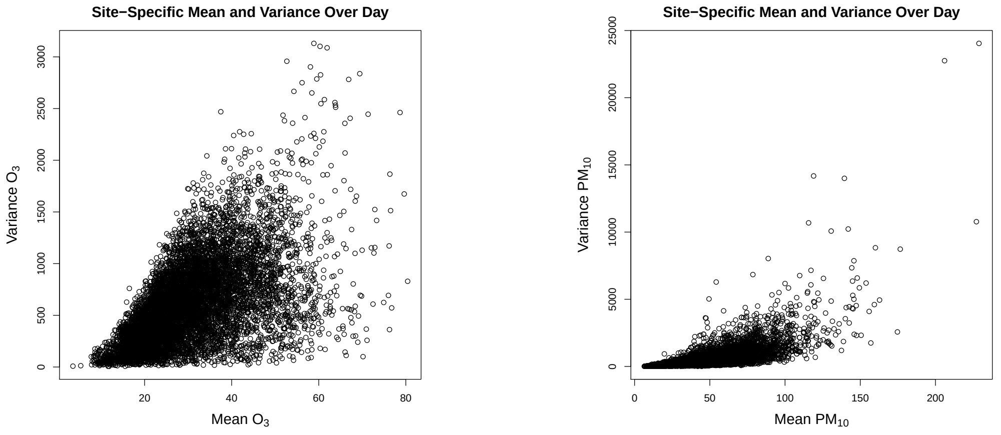
(a) Hourly PM_{10} means

(b) Daily PM_{10} means

(c) Hourly PM_{10} std. dev.

FIGURE 3 Site-specific hourly means, means averaged over every day, and standard deviations over hour of the day

To further investigate the relationship between the mean and variance, we fit a simple AR model with site-specific regression and AR coefficients and explore the residuals as a function of the mean. We include lags 1, 2, 24, and 168. After fitting the model, we bin observations according to their mean and calculate the variance of the associated residuals within that bin. The results for ozone and PM_{10} concentration are in Figure 4. These plots suggest that using VST's may effectively address the correlation between the mean and variance of model residuals. Specifically, the mean-variance relationship for ozone is strong (and approximately linear for values less than 50 ppb), and the mean-variance relationship for PM_{10} appears to be approximately quadratic. Linear mean-variance relationships are stabilized by a square-root transformation, and quadratic mean-variance associations are roughly removed using log transformations (see, for example, Section 3.3. in Hocking, 2013).



(a) O_3 mean-variance

(b) PM_{10} mean-variance

FIGURE 4 Binned residual variance for ozone and PM_{10} plotted against mean.

2 | MODEL SELECTION AND PREDICTION OF HELD-OUT DATA

This model selection centers around how many and which lagged terms should be used in our spatiotemporal model. For model comparison, we hold out 10% of the data and impute or update these held-out values each step of the Gibbs sampler which is described below.

$$\mu_{1it} = \mathbf{x}_{i(t-1)}^T \boldsymbol{\beta}_{1i} + \mathbf{L}_{it}^{O\,T} \boldsymbol{\gamma}_{1i} + \psi_{1i},$$
$$\mu_{2it} = \mathbf{x}_{i(t-1)}^T \boldsymbol{\beta}_{2i} + \mathbf{L}_{it}^{PM\,T} \boldsymbol{\gamma}_{2i} + \psi_{2i},$$

and let $Y_{it}^O \mid \cdots$ and $Y_{it}^{PM} \mid \cdots$ denote the full conditional distributions of missing observations. For the heteroscedastic model, the full conditional distributions for the missing data are

$$Y_{it}^O \mid \cdots \sim N(\tau_{1it}^* \mu_{1it}^*, \tau_{1it}^*)$$
$$Y_{it}^{PM} \mid \cdots \sim N(\tau_{2it}^* \mu_{2it}^*, \tau_{2it}^*)$$

with

$$\tau_{1it}^* = \left(\frac{1}{\sigma_{1t}^2} + \sum_{j=1}^{n_{1l}} \frac{\gamma_{1j}^2}{\sigma_{1(t+l_{1j})}^2} \right)^{-1}$$
$$\mu_{1it}^* = \mu_{1it} + \sum_{j=1}^{n_{1l}} \frac{\gamma_{1ij}(Y_{i(t+l_{1j})}^O - m_{1i(t+l_{1j})} + \gamma_{1ij} Y_{it}^O)}{\sigma_{1(t+l_{1j})}^2}$$
$$\tau_{2it}^* = \left(\frac{1}{\sigma_{2t}^2} + \sum_{j=1}^{n_{2l}} \frac{\gamma_{2j}^2}{\sigma_{2(t+l_{2j})}^2} \right)^{-1}$$

$$\mu^*_{2it} = \mu_{2it} + \sum_{j=1}^{n_{2l}} \frac{\gamma_{2ij}(Y^{PM}_{i(t+l_{2j})} - m_{2i(t+l_{2j})} + \gamma_{2ij} Y^{PM}_{it})}{\sigma^2_{2(t+l_{2j})}},$$

where l_{1j} is the j^{th} lag for ozone with coefficient γ_{1ij} and l_{2j} is the j^{th} lag for PM$_{10}$ with coefficient γ_{1ij}. The imputation method for the homoscedastic model is a special case of the heteroscedastic model.

The heteroscedastic models with variance that varies over hour of the day and month of the year performed uniformly worse than homoscedastic counterparts that used VST's (results not shown). Examining various combinations of square-root transformations, log transformations, and truncated distributions, we found that models using the square-root transformation for ozone and the log transformation for PM$_{10}$ gave the best predictive performance. So, for model selection, we only provide the results for six models which use the square-root transformation for ozone and the log-transformation for PM$_{10}$ but differ in terms of the lagged terms are included in the model. The results of this comparison are given in Table 1.

We further note that in preliminary model comparison, we found that models which included a lag-3 and other higher order lags or that excluded lag-2 saw no improvement in terms of prediction; thus, we arrived at the models included in Table 1. Given these results, though the differences are modest, we adopt the model for ozone and PM$_{10}$ uses lags 1, 2, 24, and 168 and the ensuing results are presented for this model.

Turning to which terms of our spatiotemporal model are used. We address whether ozone and PM$_{10}$ random effects should be modeled jointly or whether they should be included in the model altogether. To do this, we compare out-of-sample predictive performance of various reduced models to the full model where spatial random effects for ozone and PM$_{10}$ are modeled jointly. The performance of these models is presented in Table 2. The table reveals essentially no differences among the models, particularly allowing for Monte Carlo error in the values for the criteria. However, since the dependence spatial model is always preferred to the independence spatial model and since the dependence coefficient, $a^{(\phi)}_{12}$ is significantly positive (see Section 3 below), we present results for the former in the manuscript.

			O_3	O_3	O_3	O_3	PM$_{10}$	PM$_{10}$	PM$_{10}$	PM$_{10}$
	Lags	ES	CRPS	RMSE	MAE	Cov	CRPS	RMSE	MAE	Cov
1	(1,2)	0.2552	2.5392	5.0448	3.3409	0.8867	6.7263	14.5427	8.7725	0.9228
2	(1,2,24)	0.2513	2.5158	4.9709	3.3035	0.8925	6.6176	14.1469	8.6189	**0.9217**
3	(1,2,24,168)	**0.2505**	**2.5140**	**4.9614**	**3.2982**	**0.8941**	**6.5947**	**14.0922**	**8.6285**	0.9229
4	(1,2,12)	0.2540	2.5298	5.0274	3.3244	0.8887	6.6959	14.4314	8.7864	0.9230
5	(1,2,12,24)	0.2509	2.5168	4.9726	3.3115	0.8917	6.6035	14.0981	8.6296	0.9220
6	(1,2,12,24,168)	0.2507	2.5154	4.9651	3.2993	**0.8941**	6.5976	14.0972	8.6378	0.9225

TABLE 1 Predictive model comparison. The "Lags" label indicates which lags are used for both outcomes. "ES," "CRPS," "MSE," "MAE," and "Cov" head columns giving ES, CRPS, MSE, MAE, and 90% prediction interval coverage. Best performances are indicated with bold text.

			O_3	O_3	O_3	O_3	PM_{10}	PM_{10}	PM_{10}	PM_{10}
	Model Constraints	ES	CRPS	RMSE	MAE	Cov	CRPS	RMSE	MAE	Cov
1	$\mathbf{V}_1 = 0$	0.2507	2.5141	4.9622	3.2982	0.8938	6.5986	14.1055	8.6298	0.9225
2	$\mathbf{V}_2 = 0$	0.2505	2.5150	4.9637	3.2986	0.8936	6.5915	14.0853	8.6135	0.9229
3	$\mathbf{V}_1 = \mathbf{V}_2 = 0$	0.2505	2.5149	4.9635	3.2991	0.8941	6.5897	14.0837	8.6085	0.9234
4	$a_{12}^{(\psi)} = 0$	0.2509	2.5156	4.9635	3.3004	0.8941	6.6032	14.1193	8.6447	0.9237
5	Full Model	0.2505	2.5140	4.9614	3.2982	0.8941	6.5947	14.0922	8.6285	0.9229

TABLE 2 Predictive model comparison. The “Lags” label indicates which lags are used for both outcomes. “ES,” “CRPS,” “MSE,” “MAE,” and “Cov” head columns giving ES, CRPS, MSE, MAE, and 90% prediction interval coverage. Best performances are indicated with bold text.

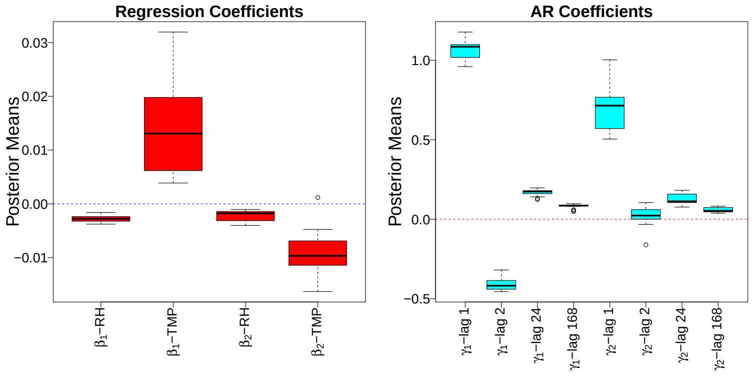
3 | POSTERIOR SUMMARIES AND DISCUSSION

Posterior summaries for covariance parameters (σ_1^2, σ_2^2, $a_{11}^{(\psi)}$, $a_{12}^{(\psi)}$, and $a_{22}^{(\psi)}$) and the overall means for hierarchical coefficient parameters ($\boldsymbol{\beta}_{01}$,$\boldsymbol{\beta}_{02}$,$\boldsymbol{\gamma}_{01}$,$\boldsymbol{\gamma}_{02}$) are given in Table 3.

	Mean	Std Dev	2.5%	97.5%
σ_1^2	0.5180	0.0016	0.5148	0.5211
σ_2^2	0.1490	0.0005	0.1481	0.1499
$a_{11}^{(\psi)}$	0.5102	0.0857	0.3633	0.6988
$a_{22}^{(\psi)}$	0.5203	0.1093	0.3176	0.7438
$a_{12}^{(\psi)}$	0.4723	0.2622	0.0224	1.0090
$\boldsymbol{\beta}_{01}$ (Intercept)	0.4503	0.0087	0.4330	0.4668
$\boldsymbol{\beta}_{01}$ (RH)	-0.0027	0.0002	-0.0031	-0.0024
$\boldsymbol{\beta}_{01}$ (TMP)	0.0138	0.0019	0.0100	0.0176
$\boldsymbol{\beta}_{02}$ (Intercept)	0.3573	0.0124	0.3300	0.3836
$\boldsymbol{\beta}_{02}$ (RH)	-0.0022	0.0002	-0.0026	-0.0018
$\boldsymbol{\beta}_{02}$ (TMP)	-0.0093	0.0009	-0.0110	-0.0076
$\boldsymbol{\gamma}_{01}$ (lag 1)	1.0649	0.0113	1.0432	1.0878
$\boldsymbol{\gamma}_{01}$ (lag 2)	-0.4079	0.0080	-0.4238	-0.3921
$\boldsymbol{\gamma}_{01}$ (lag 24)	0.1683	0.0041	0.1603	0.1763
$\boldsymbol{\gamma}_{01}$ (lag 168)	0.0835	0.0028	0.0780	0.0889
$\boldsymbol{\gamma}_{02}$ (lag 1)	0.6952	0.0250	0.6446	0.7442
$\boldsymbol{\gamma}_{02}$ (lag 2)	0.0227	0.0112	0.0011	0.0448
$\boldsymbol{\gamma}_{02}$ (lag 24)	0.1261	0.0065	0.1130	0.1391
$\boldsymbol{\gamma}_{02}$ (lag 168)	0.0572	0.0033	0.0507	0.0638

TABLE 3 Posterior summaries for covariance parameters and overall or common means for the hierarchical regression and autoregression coefficients. $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\gamma}_{01}$ are interpreted with respect to the square-root ozone scale. $\boldsymbol{\beta}_{02}$ and $\boldsymbol{\gamma}_{02}$ are interpreted as effects on PM_{10} on the log-scale.

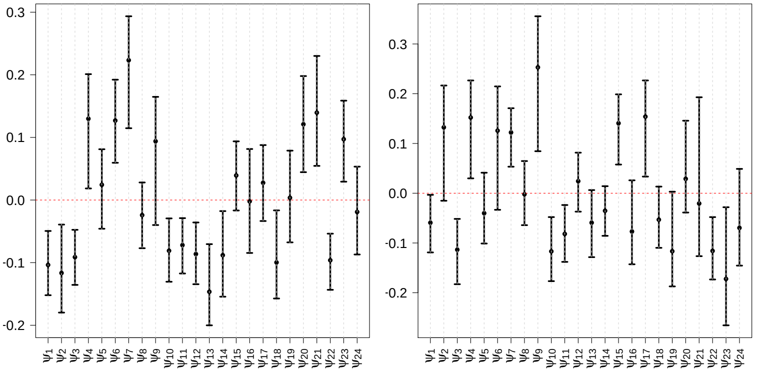
Because $a_{12}^{(\psi)}$ is significantly positive, this indicates that regions with higher spatial random effects for ozone generally have higher random effects for PM_{10}. The inference given by $\boldsymbol{\beta}_{01}$ confirms that ozone is negatively related to RH and positively related to TMP, again as we noted in our exploratory analyses. For PM$_{10}$, we see a negative relationship with RH and TMP via $\boldsymbol{\beta}_{02}$, while the only relationship that was evident in our exploration was the negative relationship with RH. The autoregressive terms for PM$_{10}$ are positive, and the lag-1 and lag-24 terms are largest. For ozone, the autoregressive terms for the lags 1, 24, and 168 are positive, but the lag-2 coefficient is negative which tempers the effect of the lag-1 coefficient. While $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$ represent the average relationships between covariates and ozone and PM$_{10}$, each site has unique covariate effects. We provide box plots for the posterior means of site-specific regression and AR coefficients in Figure 5 (each box displays the 24 site-specific posterior means for that coefficient). In general, the site-specific coefficients are in the same direction as the overall effect, as we would expect. Interestingly, the effect of temperature varies significantly between locations. The site-specific AR coefficients generally are tightly clustered except the lag 1 coefficient for PM$_{10}$. We plot the posterior means and credible intervals for ψ_{1i} and ψ_{2i} in Figure 6. For most sites, the 95% credible intervals for ozone's spatial random effects exclude 0. By contrast, the credible intervals for PM$_{10}$'s spatial random effects include 0 for 10 of the 24 stations.



(a) Site-specific regression coefficients

(b) Site-specific AR coefficients

FIGURE 5 Posterior means for site-specific regression and AR coefficients for ozone and PM$_{10}$. Each box displays the 24 site-specific posterior means for that coefficient.



(a) Site-Specific CAR Effects for Ozone

(b) Site-Specific CAR Effects for PM$_{10}$

FIGURE 6 Posterior means and credible intervals for site-specific CAR random effects for ozone and PM$_{10}$.

4 | ANALYSIS OF MEXICAN AMBIENT AIR QUALITY STANDARDS FOR ILLUSTRATIVE MONTHS

For our purposes, we group either type of ozone exceedance, one or eight-hour, together. In the online supplement, we focus on three months, April, August, and December, to illustrate how exceedance probabilities change over the course of the year. April and August are warm months, and ozone creation needs heat. August is the wettest month of the year in Mexico City, on average. Rainfall tends to clear out PM, so PM_{10} levels are expected to be low in August. April, on the other hand, precedes the rainy season and is normally dry. December is cold and dry. These months naturally contrast each other by illustrating how yearly climate affects pollution levels and the probability of exceeding Mexican ambient air quality standards. We plot regional daily exceedance probabilities for ozone ($P(W_{jd}^{O} > 95\text{ ppb} \cup W_{jd}^{\overline{O}} > 70\text{ ppb})$) and PM_{10} ($P(W_{jd}^{PM} > 75\,\mu g/m^3)$) over April (Figure 7), August (Figure 8), and December (Figure 9).

Recall that April had low phase probabilities except for April 6th when phase I probabilities spiked in all regions due to high ozone levels. Unsurprisingly, the daily exceedance probabilities for ozone are high for all regions in April. The northern regions (NE and NW) have the lowest ozone exceedance probabilities but are still above 1/2 most of the month. Daily PM_{10} exceedance probabilities vary over the month, with probabilities near one before the 13th, zero from the 13th of April to April 19th, and again high toward the end of the month.

In our phase analysis, we showed that August had uniformly low probabilities of predicted pollution emergencies. Because August is in the rainy season, we expected that it would have low PM_{10}. This is confirmed by our analysis with August having low daily PM_{10} exceedance probabilities, with the exception of a three days (8/11, 8/15, 8/16). Daily ozone exceedance, on the other hand, is high over most of the month for three regions (CE, SE, SW). Like in April, the northern regions have lower probability of ozone exceedance.

In December, we showed that phase probabilities were predicted to be low with the exception of a single peak in phase I probabilities in the northeast region due to high PM_{10} concentrations. Because the phase I PM_{10} threshold is nearly three times Mexican ambient air quality standards, it is unsurprising to observe high exceedance probabilities for PM_{10} in the northeast region. It is, however, interesting that four of the five regions have predicted daily exceedance probabilities of one (or very close to one) for 20 or more days. Ozone has many periods of low exceedance probabilities but does exhibit high daily exceedance probabilities overall.

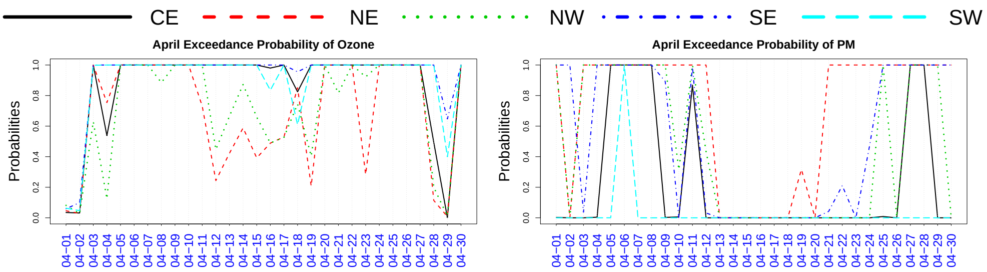


FIGURE 7 Exceedance probabilities of (Left) ozone and (Right) PM_{10} for April for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

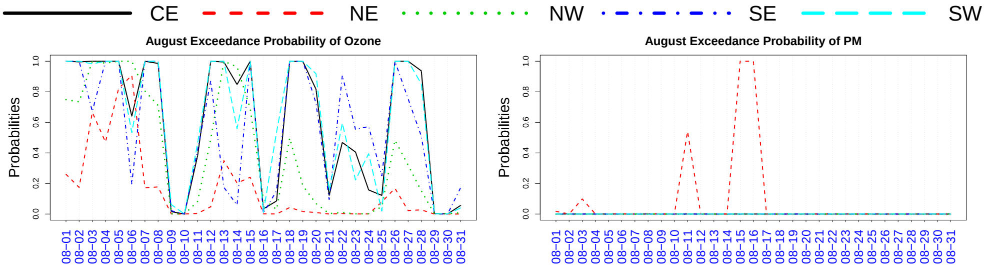


FIGURE 8 Exceedance probabilities of (Left) ozone and (Right) PM_{10} for August for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

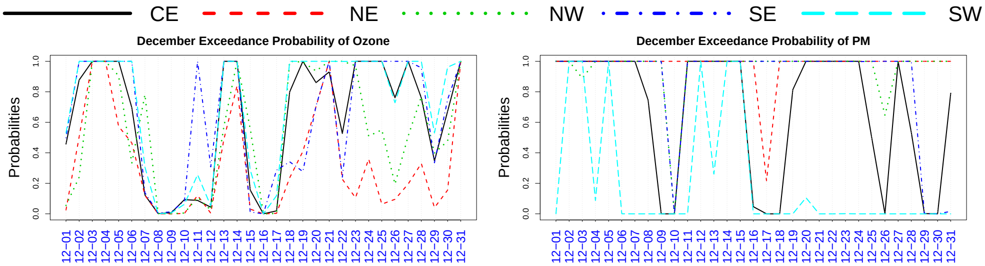


FIGURE 9 Exceedance probabilities of (Left) ozone and (Right) PM_{10} for December for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

references

Hocking, R. R. (2013) *Methods and applications of linear models: regression and the analysis of variance*. John Wiley & Sons.