Investigating the underlying structure of particulate matter concentrations: a functional exploratory data analysis study using California monitoring data

Eduardo L. Montoya^{1,a}

^aDepartment of Mathematics, California State University, Bakersfield, USA

Abstract

Functional data analysis continues to attract interest because advances in technology across many fields have increasingly permitted measurements to be made from continuous processes on a discretized scale. Particulate matter is among the most harmful air pollutants affecting public health and the environment, and levels of PM10 (particles less than 10 micrometers in diameter) for regions of California remain among the highest in the United States. The relatively high frequency of particulate matter sampling enables us to regard the data as functional data. In this work, we investigate the dominant modes of variation of PM10 using functional data analysis methodologies. Our analysis provides insight into the underlying data structure of PM10, and it captures the size and temporal variation of this underlying data structure. In addition, our study shows that certain aspects of size and temporal variation of the underlying PM10 structure are associated with changes in large-scale climate indices that quantify variations of sea surface temperature and atmospheric circulation patterns.

Keywords: functional data, particulate matter, principal component analysis

1. Introduction

The effects of air pollution are diverse and numerous, such as increased mortality to increased sensitivity of an ecosystem at high concentrations. Effective air quality standards are designed to mitigate these effects on health and the environment. The pollutants that pose among the highest risk to California are ground level ozone and particulate matter (PM). California continues to mandate ambient air quality standards that tend to be more stringent than national standards. However, reducing air pollution concentrations to acceptable levels remains an on-going challenge in California where seven and eight of its cities rank in the top 10 of the highest levels of ozone and particulate matter pollution, respectively (American Lung Association, 2017).

Air quality is highly associated with weather conditions (Jhun *et al.*, 2015), and climate conditions impact weather patterns (Karl and Knight, 1998; Burroughs, 2007). Thus, air quality is susceptible to changes in climate. For example, climate change has been linked to an extension of the ozone season in the southeastern U.S. (Zhang and Wang, 2016) and to increases in high spring ozone episodes in the western U.S. (Lin *et al.*, 2015). It is important to address the effects of climate change on air quality because its effects may threaten to negate the progress by air quality agencies that develop strategies to reduce levels of air pollution. This challenge will become more complex as climate change continues to be more pronounced over time. While the aforementioned references and many other recent studies have shown evidence of the effects of climate change on ground-level ozone, the connection between

¹ Department of Mathematics, California State University, Bakersfield 93311, USA. E-mail: emontoya2@csu.edu

PM and climate change are less established (Dawson *et al.*, 2014). PM refers to small particles of solid or semi-solid material found in the atmosphere.

PM is one of the six common air pollutants identified by the U.S. Environmental Protection Agency (EPA, 2018), and it varies in size. Some recent results suggest that the Pacific Decadal Oscillation and South Oscillation index influence PM of a certain size (Singh and Palazoglu, 2012). Jerez *et al.* (2013) showed that the positive and negative phases of the North Atlantic Oscillation index are associated with higher and lower levels of PM, respectively, over certain regions. Feng *et al.* (2016) found that the Pacific-North America teleconnection index during positive phases is associated with higher levels of PM. Using California monitoring data, our focus is on studying the variation of PM that is less than 10 micrometers in diameter, commonly referred to as PM10.

Many studies of PM10 utilize daily and monthly summary measures which are assumed to represent daily individual exposure to PM. In this article, we add to the literature by exploring the underlying data structure of the height or size variation and time variation of PM10 and its association with regimes of Pacific climate using functional data (FD) methodologies. FD methodologies allow us to explore the behavior of PM10 at any time t in the day. There have been a few prior studies that apply FD methods to particulate matter data (Ignaccolo *et al.*, 2014; Shaadan *et al.*, 2015; Ranalli *et al.*, 2016), but they do not account for the height and time variation of PM10. Ignoring these variations may lead to biased data summaries or model results. The notable exception is a study by Morlini (2007), which did consider these variations of PM10, but did not explore what drives these processes or how they may be influenced by changes in climate. Better understanding of these variations and their association with climate is important for air pollution studies. Section 2 provides further details about the PM10 sample data, and it describes the large-scale climate indices considered in this study. A curve alignment procedure is applied to separate the time and height variation of the PM10 sample in Section 3. Section 4 applies functional principal component analysis on the resulting variations to extract their dominant modes of variability. We use a generalized additive mixed model to investigate how these dominant modes are associated with spatial location and certain climate indices in Section 5. Overall, we provide a new understanding of the spatial and temporal behavior of daily PM10, as well as its association with certain sea surface temperature and atmospheric circulation patterns. Such awareness is pivotal to better understanding how levels of PM respond to climate change. The data in this study was managed and analyzed with R (R Development Core Team, 2017).

2. Data

Monitoring stations provide hourly measurements of PM10 throughout the day. We processed hourly time series of PM10 concentrations (measured in micrograms per cubic meter) from 117 California monitoring stations among 35 counties for the years 1995–2017. Raw PM10 data is publicly available through the EPA Air Data website (<https://www.epa.gov/outdoor-air-quality-data>). The spatial locations of the stations are shown in the left panel of Figure 1. Of 117 monitoring stations, 37 and 80 were located in a rural and non-rural setting, respectively. A monitoring station in Riverside County, CA, at elevation -50m near the north shore of the Salton Sea in the eastern Coachella Valley, has the lowest elevation. A monitoring station located in Mono County, CA at elevation 2395m in the eastern slopes of California's Sierra Nevada mountain range, has the highest elevation.

We propose to use the relatively high frequency of hourly PM10 sampling which enables us to analyze these hourly PM10 measurements using methods of functional data analysis (FDA). In this framework, we view hourly observations as a continuously evolving function over time at a monitoring station. Since this process is continuous, the hourly measurements are viewed as samples from that

Figure 1: Left panel: Locations of the California PM monitoring stations used in this study. The amount of data available varied by monitoring station. Time periods vary depending on the availability of data, and the data ranges between 1995 and 2017. The symbol $\times$ marks the location of the Kern County monitoring station whose PM₁₀ curves are shown in the right panel. Right panel: Daily PM₁₀ curves for a monitoring station in Kern County, CA from 1995 to 2017. Each PM₁₀ curve corresponds to a specific day. The units of PM₁₀ are in micrograms per cubic meter.

function plus measurement error. Such hourly measurements may be referred to as functional data. Key FDA references include Ramsay and Silverman, (2005) and Kokoszka and Reimherr (2017). Conceptually, PM₁₀ is continuously defined. In practice, PM₁₀ is observed on a discretized hourly scale. We represent the data as functional data using penalized regression B-splines with equidistant knots, where the number of knots are chosen based on the fixed selection method described in Montoya *et al.* (2014), and we use generalized cross validation for smoothing parameter selection. Therefore, we have a sample of PM₁₀ curves, $pm_{ijkl}(t)$, where i , j , k , and l denote the i^{th} monitoring station, j^{th} year, k^{th} month, and l^{th} day. In all, our sample consists of 197,570 PM₁₀ curves. The functional nature of PM₁₀ versus time of day is illustrated in the right panel of Figure 1 for a monitoring station located in Kern County, CA.

Meteorological variables, such as relative humidity, temperature, wind speed, and wind direction, influence PM₁₀ (Csavina *et al.*, 2014). Variability in these local meteorological variables can be attributed to climate variability. Large scale climate indices reflect atmospheric and oceanic variability associated with events such as El Niño (Liu *et al.*, 2018). We study how the dominant modes of variability in PM₁₀ are influenced by two climate indices, the Oceanic Niño Index (ONI) and the Pacific-North America (PNA) index. The ONI characterizes departures from normal sea surface temperature over an east-central region of the Pacific Ocean. The National Oceanic and Atmospheric Administration (<https://www.esrl.noaa.gov/psd/data/correlation/oni.data>) provides monthly ONI data. The PNA index quantifies large-scale atmospheric circulation patterns over North America and the Pacific Ocean. Monthly PNA data are also provided by National Oceanic and Atmospheric Administration (<https://www.esrl.noaa.gov/psd/data/correlation/pna.data>). Negative and Positive values of these climate indices are associated with El Niño and La Niña episodes, respectively (Bridgman and Oliver, 2006; Yu *et al.*, 2011). Figure 2 displays monthly measurements of these climate indices.

Figure 2: Monthly ONI and PNA index values from 1995–2017 shown in the left and right panel, respectively.

Figure 3: The solid lines represent the simulated curves. The dashed line represents the functional average of these two curves. The dotted line represents the functional average of the these two curves after the lateral displacement variation has been removed from this sample.

3. Separating the height and time variation of PM10

Functional data generally consists of two types of variations: height variation and lateral displacement or time variation. The lateral displacement in functional data arises due to randomness of the timing of the features in each curve in the sample, while the height variation represents randomness of the size of the values that these features in the sample assume. These variations are illustrated in Figure 3 using two simulated curves. The solid curves represent two functional observations. Individually, they show two notable features; a morning and afternoon rush hour peak. However, the timing of these features, as well as the size or height of these features, varies by curve. The functional average of these two curves do not represent the typical behavior of these curves. The two curves each have two peaks of varying size, yet the average of these functions would suggest that the general behavior of these two curves is two morning peaks followed by two afternoon peaks. Further, the size of the peaks in the mean curve are much lower than the peaks of the two curves. This simple example illustrates

that the time variation may lead to a distorted summary of the data. Further, the time variation in the sample will also inflate the sample variance (Marron *et al.*, 2015).

To better explore the underlying structure of the data, a curve alignment procedure that separates the time and height variation of the data is applied. Curve alignment methods align a sample of curves to one common form. This alignment is carried out through the estimation of non-linear monotone functions for each curve in the sample that is applied to the time scale. These time scale transformations, often called warping functions, represent the time variation in the sample. The aligned sample represents the amplitude variation in the sample. Marron *et al.* (2015) provides a summary and comparison of different curve alignment methods. Most curve alignment procedures are based on an L_2 distance to quantify discrepancies between curves. For example, consider aligning two functions, $p_1(t)$ and $p_2(t)$. L_2 distance based procedures provide an estimate of a time scale transformation, $h(t)$, to align $p_2(t)$ to $p_1(t)$ by minimizing a criterion that consists of the distance $\|p_1(t) - p_2(h(t))\|$ where $\|p_1(t)\|$ denotes the $\mathbb{L}^2$ norm $(\int |p_1(t)|^2 dt)^{1/2}$. However, as noted in Srivastava and Klassen (2016), problems arise with L_2 distance based alignment methods. Among them, there is a lack of symmetry in that the alignment of $p_1(t)$ to $p_2(t)$ with this criterion would not provide the same estimate of $h(t)$. Furthermore, it lacks isometry because distance is generally not preserved under identical time scale transformations (i.e., $\|p_1(t) - p_2(t)\| \neq \|p_1(h(t)) - p_2(h(t))\|$). Further discussion of these drawbacks and others of L_2 distance based curve alignment procedures may be found in Srivastava and Klassen (2016).

To align the PM10 curves, we utilize the square-root slope function (SRSF) based curve alignment procedure proposed by Srivastava *et al.* (2011). This approach does not have the drawbacks of the L_2 distance based methods. We provide a short summary of this procedure. Assume $p_1(t)$ and $p_2(t)$ are defined on the same intervals as the PM10 curves. Let $\mathcal{F}$ denote the space of all absolutely continuous functions on $[0, 23]$, and $\mathcal{H}$ denote the set of boundary-preserving non-linear strictly monotone transformations on the interval $[0, 23]: \{h : [0, 23] \rightarrow [0, 23] | h(0) = 0, h(23) = 23\}$. To align $p_1(t)$ to $p_2(t)$, this method uses the Fisher-Rao (FR) distance between two curves as the basis for its alignment criterion, but this distance is challenging to compute in $\mathcal{F}$. To circumvent this challenge, functions are transformed using the SRSF defined as $q_i(t) = p'_i(t) / \sqrt{|p'_i(t)|}$ where $p'_i(t)$ denotes the derivative of $p_i(t)$. In the space of the SRSFs, the FR distance becomes an $\mathbb{L}^2$ distance, and the SRSF of $p(h(t))$ is given by $q(h(t)) \sqrt{h'(t)}$. This motivates estimating $h(t)$ in the space of the SRSFs to align $p_2(t)$ to $p_1(t)$ by minimizing $\|q_1(t) - q_2(h(t)) \sqrt{h'(t)}\|$. This criterion is for aligning two curves.

To align multiple curves, the curves are aligned to the Karcher mean under the space of the SRSFs. Srivastava *et al.* (2011) provides a dynamic programming based algorithm to align a sample of curves. This method results in an alignment of the sample of curves such that the functional mean of these aligned curves better reflect the true features and their sizes, while the estimated warping functions indicate their relative positions. In our context, we denote the warping function by $h_{ijkl}(t)$, where i , j , k , and l denote the i^{th} monitoring station, j^{th} year, k^{th} month, and l^{th} day. The aligned PM10 curves are given by $pm_{ijkl}^*(t) = pm_{ijkl}(h_{ijkl}(t))$, for all (i, j, k, l) . The R package *fdasrv* (Tucker, 2017) was used to perform the SRSF based curve alignment procedure on the PM10 curves, and it provides the $pm_{ijkl}^*(t)$'s and the $h_{ijkl}(t)$'s on a discretized scale.

We represent as functions the discretized estimates of the aligned PM10 curves using the B-spline basis expansion method (Ramsay and Silverman, 2005) with 23 basis functions. The discretized estimates of the warping functions are represented as functions in a similar manner but the B-spline basis coefficients are monotonically constrained to ensure that the functional representation of the warping functions are monotonically increasing. One may refer to Montoya and Meiring (2016) and others cited in their paper for details on monotonically constrained B-splines. The left, middle, and

Figure 4: The (a) panel shows the sample of PM10 curves across all monitoring stations, days, months, and years. The estimated warping functions are displayed in the (b) panel. The (c) panel displays the corresponding aligned PM10 curves.

Figure 5: The (a), (b), and (c) panel show functional boxplots of all the PM10 curves, the estimated warping functions, and the corresponding aligned PM10 curves, respectively.

right panels of Figure 4 show the PM10 curves, the estimated warping functions ($h_{ijkl}(t)$'s), and the corresponding registered curves ($pm_{ijkl}^*(t)$'s) for all (i, j, k, l) . Due to the amount of functional observations, it is difficult to see any notable features in these graphs except for a peak in the morning and late afternoon rush hours in the aligned curves.

To better summarize these data, Figure 5 displays the corresponding functional boxplots (Sun and Genton, 2011) of the panels shown in Figure 4. In the functional boxplots, the shaded region corresponds to a 50% central region akin to the box in a traditional boxplot with the functional median represented by a black line. The vertical lines extending from the shaded region to the light gray lines above and below the shaded region represent functional versions of the whiskers in traditional boxplots. The align PM10 curves show some notable characteristics regarding the height variation that are not visible in the unaligned PM10 curves. In particular, the larger features or episodes of PM10 are occurring during the morning and afternoon rush hours, as well as within a few hours of midnight. The timing of these features are quantified by the warping functions. For example, let t_{fc} denote the timing of a particular feature. If $h_{ijkl}(t_{fc}) > t_{fc}$, this implies that the timing of this particular feature for a given PM10 curve is occurring later in the day relative to the timing of this feature for the mean of aligned PM10 curves, whereas if $h_{ijkl}(t_{fc}) < t_{fc}$, then the timing of this particular feature occurs earlier in the day. Overall, the warping functions show that there is substantial variation in the timing of the PM10 curve features.

4. Exploring the underlying structure of the PM10 curves

Functional Principal Component Analysis (FPCA) is a powerful exploratory method that extracts the dominant mode(s) of variability in a functional dataset. Such modes allow us to explore the major sources of variability in the data. We refer the reader to Shang (2014) for a review of different approaches to FPCA. Here, we use regularized FPCA (Silverman, 1996; Ramsay and Silverman, 2005) due its ability to control the smoothness of a functional principal component, and, as a consequence, may make the functional principal components less difficult to interpret.

We use regularized FPCA to explore whether the height and time variation of PM10 could be described by a few underlying components that primarily dominate its behavior. We briefly describe regularized FPCA. Assume the sample functional data are centered about its mean, and denote the centered sample as $x_i(t)$ for $i = 1, \dots, n$. First, define the sample covariance function $\hat{v}(s, t) = 1/(n-1) \sum_{i=1}^n x_i(s)x_i(t)$. Analogous to multivariate PCA, FPCA seeks to find a function, $\kappa(t)$, that maximizes the variance of $\int_0^{23} \kappa(t)x_i(t)dt$ for each i , subject to the constraint $\int_0^{23} \kappa(t)\kappa(t)dt = 1$. The function $\kappa(t)$ is a functional principal component (FPC). The c^{th} FPC, $\kappa_c(t)$, maximizes the variance of $\int_0^{23} \kappa_c(t)x_i(t)dt$ for each i , subject to constraints $\int_0^{23} \kappa_c(t)\kappa_c(t)dt = 1$ and $\int_0^{23} \kappa_c(t)\kappa_b(t)dt = 0$ for $b < c$. The solution to this problem is the eigenfunction that satisfies the equation $\int_0^{23} \hat{v}(s, t)\kappa_c(t)dt = \psi_c\kappa_c(s)$, where ψ_c are the eigenvalues of the eigenfunction $\kappa_c(t)$ (Ramsay and Silverman, 2005). Regularized FPCA can provide a more stable estimate of the FPCs by controlling their smoothness by selecting $\kappa_c(t)$ to maximize

$$\frac{\text{Var}\left(\int_0^{23} \kappa_c(t)x_i(t)dt\right)}{\int_0^{23} (\kappa_c(t))^2 dt + \lambda \int_0^{23} (\kappa_c''(t))^2 dt} \quad (4.1)$$

subject to the constraints $\int_0^{23} \kappa_c(t)\kappa_c(t)dt = 1$ and $\int_0^{23} \kappa_c(t)\kappa_b(t)dt + \int_0^{23} \kappa_c''(t)\kappa_b''(t)dt = 0$ for $b < c$, where $\text{Var}(\int_0^{23} \kappa_c(t)x_i(t)dt)$ denotes the variance of $\int_0^{23} \kappa_c(t)x_i(t)dt$. Regularized FPCA was performed on $h_{ijkl}(t)$ and $pm_{ijkl}^*(t)$ across all (i, j, k, l) using the R package fda (Ramsay *et al.*, 2017). To select the parameter λ , the cross validation procedure given in Ramsay and Silverman (2005) was implemented in R.

Regarding the height variation of the PM10 curves, the first three FPCs shown in the left panel of Figure 6 account for 92.84% (62.67%, 24.40%, 5.77%) of the total variation in the $pm_{ijkl}^*(t)$'s. The first FPC function places the most positive weight in early morning and late afternoon. PM10 curves with above average levels in the rush hours, especially in the morning, would have high positive scores. The second FPC function reflects contrasting behavior between the early morning and late afternoon. PM10 curves with below average levels in the morning rush hour and above average levels in the afternoon rush hour would have high positive scores on this component. The third FPC explains substantially less of the variance.

The right panel of Figure 6 displays the first three FPCs of the time variation of the PM10 curves, and they account for 91.56% (63.16%, 20.98%, 7.42%) of the total variation in the $h_{ijkl}(t)$'s. The first FPC function places positive weight throughout the day. PM10 curves with high positive scores on this component would tend to have features occurring later in the day. The second FPC function shows a strongly negative contribution in the morning rush hour, followed by a strongly positive contribution for the late afternoon rush hour. This reflects a measure of uniformity in regards to timing of PM10 features throughout the day. The third FPC does not explain much of the total variance.

Figure 6: The first three FPC functions of the height and time variation of the PM10 curves are shown in the left and right panels, respectively.

The scores of a given FPC can be used to examine how climate and spatial location may influence the behavior of these components and hence, the behavior of PM10. For brevity, we only focus on the scores of the first FPC for the PM10 height and time variation. We use a generalized additive mixed model (Wood, 2017a) on the scores to examine these potential influences. As covariates, we consider the easting coordinate, northing coordinate, elevation, and location setting of a monitoring station. In addition, we take into account ONI and the PNA index. For each set of scores from the first FPC of the height and time variation of PM10 curves, we fit the following generalized additive mixed model (GAMM),

$$z_{ijkl} = \alpha + q_{ij} + Y_k + V_i \beta_{\text{elev}} + L_i \beta_{\text{ls}} + f_1(E_i, N_i) + f_2(\text{oni}_j, \text{pna}_j) + \varepsilon_{ijkl},$$

$$\text{where } \varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad q_{ij} \sim \mathcal{N}(0, \sigma_q^2), \quad \text{and } Y_j \sim \mathcal{N}(0, \sigma_Y^2), \quad (4.2)$$

where z_{ijkl} denotes a FPC score for the i^{th} monitoring station, j^{th} year, k^{th} month, and l^{th} day. Further, α denotes the intercept, V_i denotes the elevation, L_i denotes the location setting (0 for rural and 1 non-rural), E_i denotes the easting coordinate, and N_i denotes the northing coordinate of the i^{th} monitoring station. PNA_{jk} and ONI_{jk} denote measurements of the PNA and ONI index in month k for year j , respectively. The parameters q_{ij} and Y_j represent a random monitoring station intercept in year j and a random year-specific intercept, respectively, and they quantify variation in the scores that is not explained by the covariates in model (4.2). We fit model (4.2) for each set of scores using the R package mgcv (Wood, 2017b). The non-parametric smooth interactions are represented using a tensor product basis with marginal cubic regression splines (Wood, 2006).

Regarding modeling results of the first FPC scores for the PM10 height variation, only the elevation of a monitoring station was not significant in model (4.2) at significance level 0.05. The model was re-fitted without elevation, and the remaining covariates in model (4.2) were significant at significance level .05. Each model term is interpreted conditional on the other terms in the model being held constant. The estimated values of the intercept and location setting effect are 13.091 and -16.961 , respectively. This suggests that non-rural locations tend to be associated with lower scores on this FPC. The effect of spatial location is given in the top left panel of Figure 7, showing that the southern-most region considered in our study tends to be associated with higher scores, whereas other regions tend

Figure 7: Top panel: Estimated smooth interaction terms for model (4.2) when modeling the first FPC scores of the PM10 height variation. Bottom panel: Analogous to the top panel for the first FPC scores of the PM10 time variation. The latitude and longitude coordinates of a monitoring station were converted to its Universal Transverse Mercator northing and easting coordinates (UTM zone No 11), measured in meters, using the R package PBSmapping (Schnute *et al.*, 2017).

to be associated with lower scores.

The estimated effect of ONI varies with PNA (top right panel of Figure 7), and it shows that the most cold (negative) ONI with cold PNA, and warm (positive) PNA with warm ONI, are associated with smaller scores. Warm ONI with cold PNA is associated with higher scores, and to a lesser extent during warm PNA with cold ONI. This suggests that during the coldest phases of ONI and PNA, as well as the warmest phases of ONI and PNA, features of PM10 tend to be smaller than average, whereas the opposite interactions of those phases are associated with larger than average features. While Feng *et al.* (2016) found that higher monthly averages of particulate matter of a certain size are associated with positive values of PNA, our results suggest similar findings but not necessarily when PNA interacts with warm ONI. Similarly, Singh and Palazoglu (2012) found that over a southern California region, warm SOI was associated with lower daily 8-hour averages of particulate matter of a certain size. Our results support similar findings for the PM10 curves but only during cold PNA.

For the modeling of the first FPC scores for the PM10 time variation, all covariates considered in model (4.2) were significant at significance level 0.05. The estimated values of the intercept, elevation

Figure 8: The left panel displays the predicted random year effect for model (4.2) when modeling the first FPC scores of the PM10 height variation. The linear trend is shown as an orange line. The right panel is analogous to the left panel for first FPC scores of the PM10 time variation.

effect, and location setting effect are 0.121, -0.002 , and 0.567, respectively. These estimates suggest that non-rural locations are associated with higher scores and that monitoring stations at higher elevation have higher scores on this FPC. The effect of spatial location (bottom left panel of Figure 7) shows that the northern-eastern region of our study area is associated with higher scores, but the western regions are mostly associated with lower scores. The bottom right panel illustrates that the estimated effect of ONI varies with PNA, with cold ONI and cold PNA being associated with positive scores, and warm ONI with cold PNA being associated with negative scores. This interaction shows that cold phases of these indices tend to have features of PM10 occur later than average, but during warm ONI with cold PNA, features of PM10 occur earlier than average.

The left and right panel of Figure 8 display the predicted random year effect for model (4.2) on the first FPC scores for the PM10 height and time variation, respectively. Regarding the height variation, the predicted random year effect shows a negative trend. For the time variation, the predicted random year effect shows positive trend. These trends arise due to time related processes not considered in model (4.2).

The left and right panel of Figure 9 display predicted random monitoring station effects when averaged over the years for model (4.2) on the first FPC scores for the PM10 height and time variation, respectively. Both panels exhibit local spatial clustering of similar signed effects. For example, for the height variation, the mid to southern central valley of California (Fresno, CA and Bakersfield, CA area) predominantly show positive predicted random effects. This region has among the highest levels of PM10 in the United States. We also note that regions that typically have better air quality in California, such as the central coast of California (San Luis Obispo, CA to Ventura, CA), mostly show negative predicted random effects. For the time variation, the mid to southern central valley of California mostly reflect positive predicted random effects.

5. Discussion and future direction

This work proposes an exploratory data analysis of PM10 using California monitoring data by first performing curve alignment on PM10 to provide its height and time variation. Ignoring time variation can lead to distorted or biased estimates when modeling PM10 data, and thus, the time variability

Figure 9: Left panel: Predicted random monitoring station effects for model (4.2) when modeling the first FPC scores of the PM10 height variation. Red circles reflect positive scores; the Green circles reflect negative scores. Right panel: Analogous to the left panel for the first FPC scores of the PM10 time variation.

should be taken into account when using FD methodologies on PM10. The results show substantial height and time variation in the features of PM10, and, for brevity, we focused only on the dominating mode of variability for each process. The dominating mode of the underlying structure of the height and time variation processes of PM10 reflected an overall daily size effect primarily during the rush hours and a time shift effect throughout the day, respectively.

The scores of the first FPC allowed us to focus on these two aspects of the behavior of PM10. Using a GAMM, we investigated how both the first FPC scores of the PM10 height and time variation varied with large-scale climate indices and spatial location. The analysis showed that the size and timing of PM10 features varied throughout California. Furthermore, our analysis gives insight into the complicated relationship of the first FPC scores of the PM10 height and time variation with climate indices that measure variations of sea surface temperature and atmospheric circulation patterns. In particular, certain phase interactions between PNA and ONI are associated with above or below average levels of PM10, as well as its features occurring earlier or later than average, throughout the day. The random monitoring station and year effects in the models accounted for additional sources of unexplained variability. The predicted random monitoring station effects showed local spatial clustering of similar signed effects, which potentially may be partly driven by PM10 transport processes such as those discussed in Chow *et al.* (1992). The behavior of the predicted random year effects may be related to other patterns of atmospheric and oceanic variability over the Pacific Ocean not reflected in PNA and ONI.

This work presents an avenue for future research. Specifically, it motivates developing a functional data regression model that quantifies the association between the daily PM10 curves (functional response) with changes in several explanatory variables representing spatial location attributes (scalar covariates) and climate processes (functional covariates). This direction of research is significant to the atmospheric sciences, and would provide a new understanding and knowledge on the spatial and temporal behavior of daily PM10 at any time t . Such results would be important for air quality control districts or agencies as they continue to develop policies to achieve air quality improvements.

Acknowledgement

The reviewers of this manuscript are thanked for their helpful comments. This research was supported in part by NSF grant HRD-1547784.

References

- American Lung Association (2017). American lung association state of the air 2017, Chicago, IL.
- Bridgman HA and Oliver JE (2006). *The Global Climate System* (1st ed), Cambridge University Press, Cambridge.
- Burroughs WJ (2007). *Climate Change: A Multidisciplinary Approach* (2nd ed), Cambridge University Press, Cambridge.
- Chow JC, Watson JG, Lowenthal DH, Solomon PA, Magliano KL, Ziman SD, and Richards LW (1992). PM10 source apportionment in California's San Joaquin Valley, *Atmospheric Environment. Part A. General Topics*, **26**, 3335–3354.
- Csavina J, Field J, Fáliz O, Corral-Avitia AY, Sáez AE, and Betterton EA (2014). Effect of wind speed and relative humidity on atmospheric dust concentrations in semi-arid climates, *Science of The Total Environment*, **487**, 82–90.
- Dawson JP, Bloomer BJ, Winner DA, and Weaver CP (2014). Understanding the meteorological drivers of US particulate matter concentrations in a changing climate, *Bulletin of the American Meteorological Society*, **95**, 521–532.
- EPA (2018). Criteria air pollutants, from: <https://www.epa.gov/criteria-air-pollutants>
- Feng J, Liao H, and Li J (2016). The impact of monthly variation of the Pacific-North America (PNA) teleconnection pattern on wintertime surface-layer aerosol concentrations in the United States, *Atmospheric Chemistry and Physics*, **16**, 4927–4943.
- Ignaccolo R, Mateu J, and Giraldo R (2014). Kriging with external drift for functional data for air quality monitoring, *Stochastic Environmental Research and Risk Assessment*, **28**, 1171–1186.
- Jerez S, Jimenez-Guerrero P, Montávez JP, and Trigo RM (2013). Impact of the North Atlantic Oscillation on European aerosol ground levels through local processes: a seasonal model-based assessment using fixed anthropogenic emissions, *Atmospheric Chemistry and Physics*, **13**, 11195–11207.
- Jhun I, Coull BA, Schwartz J, Hubbell B, and Koutrakis P (2015). The impact of weather changes on air quality and health in the United States in 1994–2012, *Environmental Research Letters*, **10**, 084009.
- Karl TR and Knight RW (1998). Secular trends of precipitation amount, frequency, and intensity in the United States, *Bulletin of the American Meteorological Society*, **79**, 231–241.
- Klumpp A, Ansel W, and Klumpp G (2004). *Urban Air Pollution, Bioindication and Environmental Awareness*, Cuvillier Verlag, Göttingen.
- Kokoszka P and Reimherr M (2017). *Introduction to Functional Data Analysis* (1st ed), Chapman and Hall, Florida.
- Lin M, Fiore AM, Horowitz LW, Langford AO, Oltmans SJ, Tarasick D, and Rieder HE (2015). Climate variability modulates western US ozone air quality in spring via deep stratospheric intrusions, *Nature Communications*, **6**, 7105.
- Liu YC, Di P, Chen SH, and DaMassa J (2018). Relationships of rainy season precipitation and temperature to climate indices in California: long-term variability and extreme events, *Journal of Climate*, **31**, 1921–1942.
- Marron JS, Ramsay JO, Sangalli LM, and Srivastava A (2015). Functional data analysis of amplitude

- and phase variation, *Statistical Science*, **30**, 468–484.
- Montoya EL and Meiring W (2016). An F-type test for detecting departure from monotonicity in a functional linear model, *Journal of Nonparametric Statistics*, **28**, 322–337.
- Montoya EL, Ulloa N, and Miller V (2014). A simulation study comparing knot selection methods with equally spaced knots in a penalized regression spline, *International Journal of Statistics and Probability*, **3**, 96–110.
- Morlini I (2007). Searching for structure in measurements of air pollutant concentration, *Environmetrics*, **18**, 823–840.
- R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Ramsay JO and Silverman BW (2005). *Functional Data Analysis* (2nd ed), Springer, New York.
- Ramsay JO, Wickham H, Graves S, and Hooker G (2017). *fda: Functional Data Analysis*, R package version 2.4.7.
- Ranalli MG, Rocco G, Jona Lasinio G, Moroni B, Castellini S, Crocchianti S, and Cappelletti D (2016). Functional exploratory data analysis for high-resolution measurements of urban particulate matter, *Biometrical Journal*, **58**, 1229–1247.
- Schnute JT, Boers N, Haigh R, Grandin C, Johnson A, Wessel P, and Antonio F (2017). *PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools*, R package version 0.8.0.
- Shaadan N, Jemain AA, Latif MT, and Deni SM (2015). Anomaly detection and assessment of PM₁₀ functional data at several locations in the Klang Valley, Malaysia, *Atmospheric Pollution Research*, **6**, 365–375.
- Shang HL (2014). A survey of functional principal component analysis, *AStA Advances in Statistical Analysis*, **98**, 121–142.
- Silverman BW (1996). Smoothed functional principal components analysis by choice of norm, *The Annals of Statistics*, **24**, 1–24.
- Singh A and Palazoglu A (2012). Climatic variability and its influence on ozone and PM pollution in 6 non-attainment regions in the United States, *Atmospheric Environment*, **51**, 212–224.
- Srivastava A and Klassen EP (2016). *Functional and Shape Data Analysis*, Springer, New York.
- Srivastava A, Wu W, Kurtek S, Klassen E, and Marron JS (2011). Registration of functional data using Fisher-Rao metric, from: arXiv preprint arXiv:1103.3817v2
- Sun Y and Genton MG (2011). Functional boxplots, *Journal of Computational and Graphical Statistics*, **20**, 316–334.
- Tucker JD (2017). *fdasrvf: Elastic Functional Data Analysis*, R package version 1.8.2.
- Wood SN (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models, *Biometrics*, **62**, 1025–1036.
- Wood SN (2017a). *Generalized Additive Models: An Introduction with R*, CRC press, New York.
- Wood SN (2017b). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, R package version 1.8–23.
- Yu JY, Kao HY, Lee T, and Kim ST (2011). Subsurface ocean temperature indices for Central-Pacific and Eastern-Pacific types of El Niño and La Niña events, *Theoretical and Applied Climatology*, **103**, 337–344.
- Zhang Y and Wang Y (2016). Climate-driven ground-level ozone extreme in the fall over the southeast United States. In *Proceedings of the National Academy of Sciences*, **113**, 10025–10030.