Insights Into the Morphology of the East Asia $PM_{2.5}$ Annual Cycle Provided by Machine Learning

Environmental Health Insights
Volume 11: 1–7
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1178630217699611

Daji Wu[1], Gebreab K Zewdie[1], Xun Liu[1], Melanie Anne Kneen[2] and David John Lary[1]

[1]William B. Hanson Center for Space Sciences, The University of Texas at Dallas, Richardson, TX, USA. [2]Department of Environmental Science, Collin College, Plano, TX, USA.

ABSTRACT: The abundance of airborne particulate matter with an aerodynamic equivalent diameter of 2.5 µm or less ($PM_{2.5}$) is a significant environmental and health issue. Many tools have been used to examine the relationship between $PM_{2.5}$ abundance and meteorological variables, but some of the relationships are nonlinear, non-Gaussian, and even unknown. Machine learning provides a broad range of practical algorithms to help examine this issue. In this study, we use machine learning to classify the morphology of $PM_{2.5}$ seasonal cycles in East Asia. Machine learning is able to objectively classify the seasonal cycles and, without a priori assumption, is able to clearly distinguish between urban and rural areas. We show an example of this in the Sichuan Basin of China. Furthermore, machine learning is also able to provide physical insights by identifying the key factors associated with each distinct shape of the seasonal cycle, such as highlighting the key role played by the topography and the built environment.

KEYWORDS: Air pollution, $PM_{2.5}$, annual cycles, SOM, random forests

RECEIVED: September 23, 2016. **ACCEPTED:** December 18, 2016.

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1827 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTEREST: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Daji Wu, William B. Hanson Center for Space Sciences, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA. Email: dxw131230@utdallas.edu

Introduction

Air pollution is a serious environmental issue all over the world, and small airborne particulates (also called aerosols) are among the major causes. In 2012, according to a recent report from the World Health Organization (WHO), an estimated 7 million people died of air pollution–related diseases. In Southeast Asia and the Western Pacific Regions where there is the largest air pollution–related burden, 2.6 million deaths were related to outdoor air pollution.[1] Particulates with an aerodynamic equivalent diameter of 2.5 µm or less ($PM_{2.5}$) are an important component of aerosols with respect to health and environmental impact.

$PM_{2.5}$ can be readily inhaled and has a significant effect on human health (including ischemic heart disease, stroke, chronic obstructive pulmonary disease, and lung cancer).[2] Exposure to $PM_{2.5}$ has been associated with increased morbidity and mortality.[3,4] Many studies have shown that there are a large number of premature deaths, including cardiopulmonary and lung cancer deaths, as a result of exposure to $PM_{2.5}$.[5,6] Also, $PM_{2.5}$ has been shown to have a severe impact on the environment. In the past few years, heavy smog caused by $PM_{2.5}$ covered over 70 cities in Northern China, and the abundance of $PM_{2.5}$ in Beijing, capital of China, has peaked at more than 1000 µg/m^3, 40 times higher than the WHO standard.[7–9] Shijiazhuang, a big city near Beijing, is considered to have the highest $PM_{2.5}$ pollution, it experienced smog for 322 of the 365 days in 2013, and the average abundance of $PM_{2.5}$ in 2013 was 148.5 µg/m^3, with a peak of 676 µg/m^3.[10–13]

There is a strong relationship between climate change and aerosols due to aerosol radiative forcing (RF).[14] The concept of RF was first introduced in studies of climate response to changes in solar insolation and CO_2. It is used to assess and compare the anthropogenic and natural drivers of climate change.[15] RFs have 2 effects on climate change: (1) direct and (2) indirect effects. The direct effect is the mechanism by which aerosols scatter and absorb shortwave and longwave radiations. The indirect effect is the mechanism by which aerosols modify the microphysical and the radiative properties, amount, and lifetime of clouds.[16,17]

In estimating the abundance of $PM_{2.5}$, many studies have sought to overcome the spatial coverage limitation of the $PM_{2.5}$ measurement sites using remote sensing and satellite-derived aerosol optical depth (AOD) data,[18,19] coupled with regression and/or numerical models. These studies showed that the relationships between $PM_{2.5}$ and AOD are not applicable for simple regression models because they are a multivariate function of many parameters, including humidity, temperature, boundary layer height, surface pressure, population density, topography, wind speed, surface type, surface reflectivity, season, land use, normalized variance of rainfall events, size spectrum and phase of cloud particles, cloud cover, and so on.[20–23] Further complications arise from the differences existing between satellite AOD products,[24–27] the difference in spatial scales involved with the in situ $PM_{2.5}$ observational data, the remote sensing data, and, finally, the sharp $PM_{2.5}$ gradients existing across cities and close to sources.

Machine learning provides a broad range of multivariate regression algorithms for empirically estimating $PM_{2.5}$ data when there is no clear and complete theoretical description, but

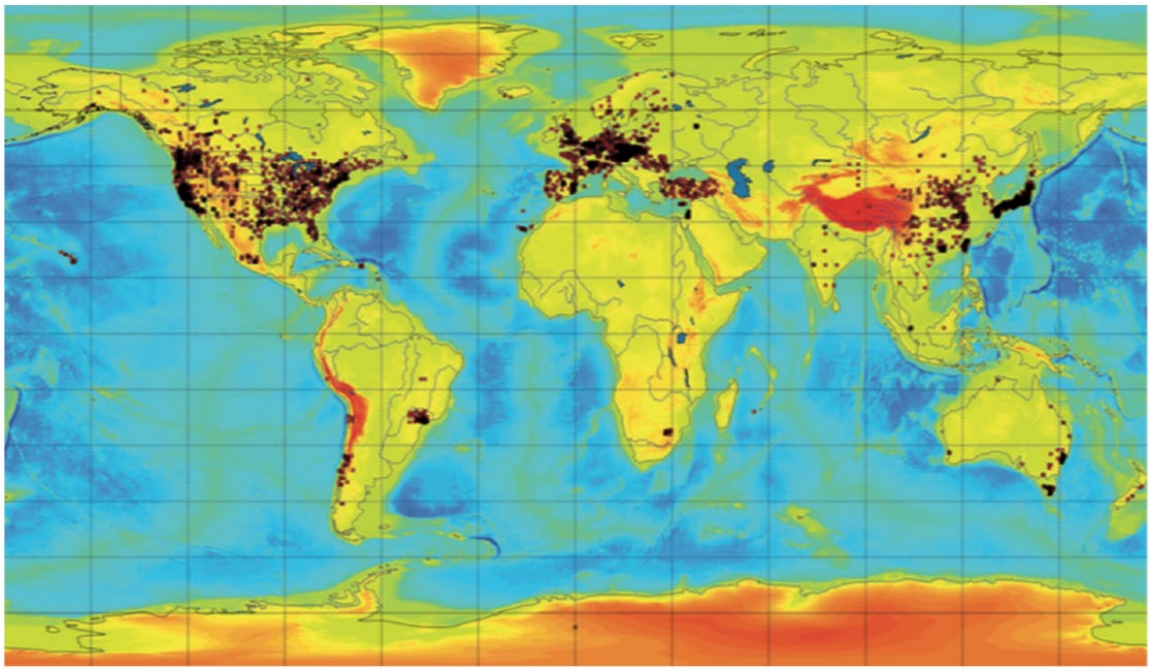

Figure 1. Over 8000 $PM_{2.5}$ measurement sites (red squares) are located in 55 countries over the period 1997-2014. Background color shows the scale of global topography and bathymetry.[28]

a set of useful observational data. In previous studies by Lary et al, machine learning was used to estimate global daily $PM_{2.5}$ data from 1997 to 2014, using in situ hourly $PM_{2.5}$ observations from more than 8000 sites in 55 countries together with comprehensive contextual data on about 100 parameters drawn from satellite data, meteorology, and demographics.

This study is a practical application of the satellite estimate $PM_{2.5}$ data product obtained by Lary et al from 2004 to 2013 in the East Asia region (20°N to 45°N, 100°E to 145°E). In this region, there is a low density of observation sites for long-term monitoring (shown in Figure 1), but heavy air pollution. The data product has a resolution of 10 km × 10 km (approximately 0.1° × 0.1°). Also, quite a new approach which combines self-organizing map and random forests was applied for the purpose of classifying the $PM_{2.5}$ annual cycles of locations over East Asia. This was done by only depending on their morphologies and then obtaining the key meteorological and environmental variables that distinguish annual cycles in a specific class from all other classes.

Datasets and Methods

Global $PM_{2.5}$ data product

The $PM_{2.5}$ observations (ground data) used for machine learning regression were obtained from the ground-based measurement sites shown in Figure 1. North America, Europe, and some parts of Asia recorded the greatest density of measurements, whereas places in the southern hemisphere, such as South America, Australia, New Zealand, and Africa, have measurement sites that are far from uniform with several gaps.

The satellite (remote sensing) data used for estimation were obtained from 3 satellite instruments: the Sea-viewing Wide Field-of-view Sensor (SeaWIFS) launched on August 1, 1997,[29] and 2 moderate-resolution imaging spectroradiometer (MODIS) instruments: MODIS on Terra satellite (EOS AM) launched in 1999 and MODIS on National Aeronautics and Space Administration's (NASA) Aqua (EOS PM) launched in 2002.[30] Detailed development and description of the estimated $PM_{2.5}$ data product are covered by the study by Lary et al.[28,31]

The meteorological and environmental data used in this study were retrieved from the NASA Modern Era Retrospective analysis for Research and Applications (MERRA), including 84 variables that describe the surface meteorology and soil state.[32]

The self-organizing map

When using large data sets to characterize a problem, it is often useful to apply an objective technique to classify the large data sets into subclasses. Self-organizing maps (SOMs) provide a way of performing such an unsupervised classification without any a priori assumption, a way to give the data "a voice."[33]

However, even with the assistance of unsupervised classification, high-dimensional data can still be challenging to visualize, as this study dealt with a 36-dimensional space (the annual cycle was split into 36 ten-day windows). The SOM is a type of artificial neural network for performing unsupervised classification. Also, it is a data visualization and unsupervised classification technique that can reduce the dimensions of

high-dimensional data using self-organizing neural networks.[34] Like other forms of machine learning, an SOM operates in 2 phases: (1) the training phase and (2) the mapping phase. The map is built by training, using examples from the training dataset, and mapping determines the class for a new input vector.[35] An SOM consists of a 2-dimensional regular grid of components called nodes. Each node is associated with a weight vector of the same dimension and a position in the map. The procedure for converting a vector from the input data space to the map is by finding the node with the most similar weight vector to the input data space.[36,37]

Random forest

Random forests were first introduced in 2001 by Leo Breiman.[38] They are a popular and efficient ensemble approach of statistical learning, useful for both classification and regression. A random forest is an ensemble of decision trees (hence the term forest). An ensemble approach allows more robust estimates that are less prone to "over-learning." The size of the "forest" ensemble was estimated by examining the estimated error as a function of the ensemble size.[39,40] In this study, the error rate plateaued at an ensemble size of approximately 30 trees. Thus, an ensemble of 50 decision trees was used in our random forest. Random forests provide an objective way of highlighting the most important predictors and providing a ranking of the relative importance of each predictor. To measure the variable's importance, we first fit a random forest to the training dataset. This provided us with the so-called out-of-bag (OOB) error. If we want the importance of a specific variable X_i (X_i is the ith predictor), the value X_i will be permuted and the OOB error will be computed again for the permuted data, and the importance of X_i will be the average of the difference between the OOB errors before and after all trees.[41,42]

Hybrid approach

In this study, a hybrid approach was used, and it utilized 2 types of machine learning. First, unsupervised classification (SOM) was used to group the $PM_{2.5}$ seasonal cycles into classes. Each class contained geographic locations which have seasonal cycles of a very similar shape. Second, we went through each of 6 typical classes sequentially and used a random forest to objectively rank which variables are most important in distinguishing that class from all the other classes, that is, which environmental factors were the most important for characterizing the shape of the $PM_{2.5}$ seasonal cycle for that class (region).

Methods

This study focused on the annual cycles of $PM_{2.5}$ abundance. Our area of interest is East Asia, and there is a total of 113 201 grid points in this area, as a 10 km × 10 km resolution. To plot an annual cycle of $PM_{2.5}$ abundance for each location, a set of 10-day averages were taken (the first average is the first 10 days of the month, the second average is the second 10 days of the month, and the remaining days of the month constitute the third average), which means that there are 3 average values in a month (36 values in 1 year). Thereafter, the average was taken over 9 years (September 1, 2004, to August 31, 2013) to get the final average annual cycle. However, it is not every location in this area that has readily available data. The Pacific Ocean covers a substantial part of the East Asian region, and ocean locations were not included in our analysis. Some locations over land can also have partial data with gaps due to cloud cover. After removing these locations, 48 186 locations remained with valid annual cycles.

The SOM was used to classify the shapes of annual cycles into 100 classes. The experiment was conducted using different numbers of classes between 90 and 200. Repeated classifications were also tested; each neural network using the same data will not be identical for each repeated training as the gains and weights of the neural network are initialized using random numbers. The classification with 100 classes turned out to be stable over repeated classifications; it assigned all the locations with a similar seasonal cycle to a single class. Classifications using less than 100 classes could not separate visibly different cycles. Using more than 100 classes did not provide any additional clarity.

In the analysis presented here, 6 classes were picked out of 100 classes as illustrative examples: class 19, class 20, and class 21 are 3 continuous classes covering the Sichuan Basin in China; class 66, class 71, and class 94 are 3 other typical classes with very different temporal distributions. Class 66 represents the area around Beijing, class 71 represents the central China area, and class 94 represents South Korea and part of Japan. For each of these 6 classes, machine learning was used to find which of the 84 meteorological and environmental factors are key in distinguishing that class from all other classes. These 6 classes cover a total of 1662 locations. The same method was used with the 84 variables as for the $PM_{2.5}$ abundance: a 10-day average was taken (36 values in 1 year), followed by a 9-year average. Thus, there are 84 annual cycles of 84 variables for each location. By applying random forests, the importance of all 84 variables can be ranked so as to distinguish a given class from all other classes. The variable importance ranking provided by the random forest is useful for gaining insights into the key environmental drivers that go into shaping the seasonal cycle.

Results and Discussion

The outcome of SOM is shown in Figure 2, 100 classes are represented by different colors, locations in the same class are shown in the same color. Figure 3A shows the average $PM_{2.5}$ abundance in the Sichuan basin. The inner basin area (class 19) has the highest average $PM_{2.5}$ abundance, which implies that this region is the most polluted. Pollution in the mid-basin (class 20) is a little less, and the outer basin (class 21) has the least abundance of $PM_{2.5}$. The average regional graph (Figure 3B) showing the probability distribution gives the same result. Class 19 represents the urban area (it includes the 2 big cities in

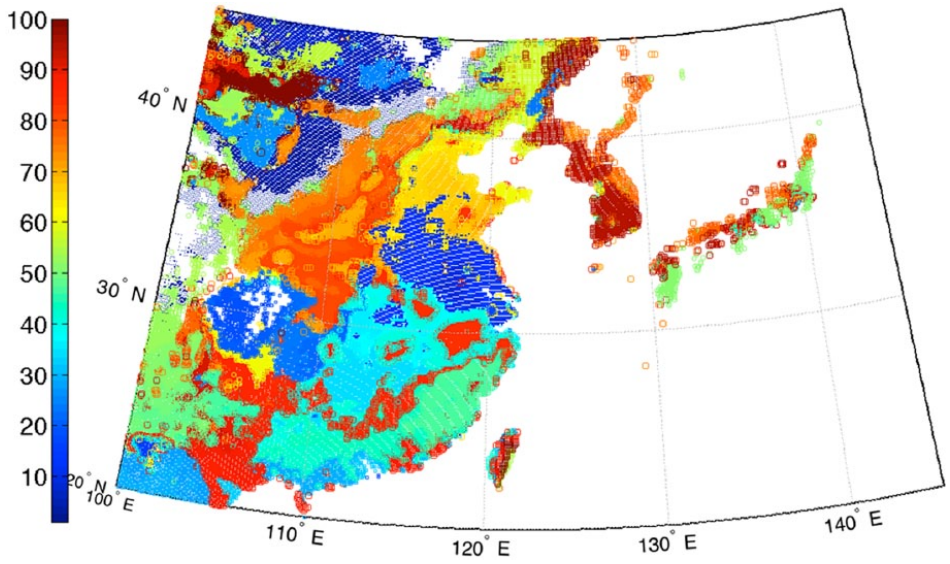

Figure 2. Graphs of the 100 classification class numbers are plotted with the latitude and longitude from source locations. All locations in the same class are shown in the same color.

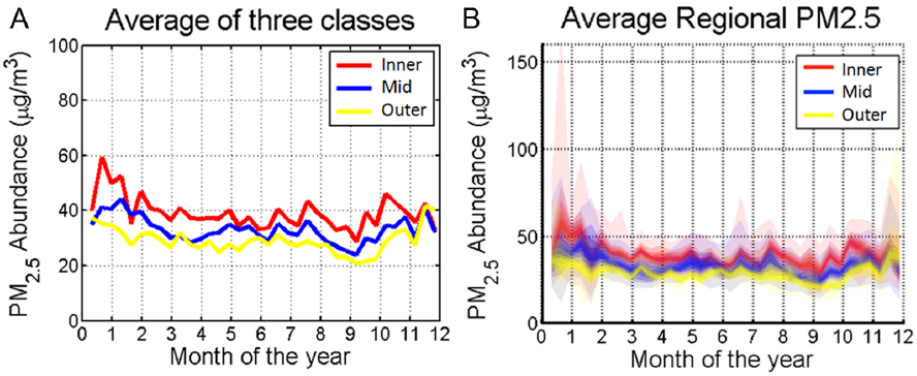

Figure 3. Comparisons among the $PM_{2.5}$ abundance in the 3 continuous classes: class19 (inner basin), class 20 (mid-basin), and class 21 (outer basin). (A) The average abundance in these 3 classes; (B) the probability distribution of $PM_{2.5}$.

Sichuan basin: Chengdu and Chongqing). It is clear that the urban areas have the highest abundance of $PM_{2.5}$ because they have the major sources of $PM_{2.5}$. These high concentrations are transported to the rural areas surrounding the cities. As this transport occurs, the $PM_{2.5}$ abundance decreases due to dilution: as a result of mixing with less polluted air during the transportation. The cities have a higher $PM_{2.5}$ abundance than the surrounding countryside. The annual cycles of these 3 classes are relatively stable throughout the entire year. The topography of the basin shows that it is a rather isolated area, and the interaction of surface air in the basin with the surroundings is weak.[43]

Figure 4 shows the $PM_{2.5}$ annual cycle classification map for East Asia, with a focus on the Sichuan basin. Areas where these annual cycle peaks can be slightly different are noted. For example, the annual cycle for the inner basin area (Figure 4C) records a peak in January. The annual cycle for the outer basin (Figure 4D) peaks at the end of December, whereas in the mid-basin area (Figure 4E), there are peaks in both December (like the outer basin) and January (like the inner basin).

Figure 5 classifies the shapes of the $PM_{2.5}$ annual cycles with a focus on the other 3 classes. In the central China area (Figure 5B), the main peak of the $PM_{2.5}$ annual cycle is in February, with a smaller secondary peak in January. For the Beijing area (Figure 5C), there are 2 peaks, one in February and the other in August. In their study, Xu et al[44] and Chan and Yao[45] pointed out that this area is a heavy industrial base in China, and the amount of emissions increases in winter and spring. These emissions include the combustion of coal for the heating of homes and the increasing emission of vehicles as a result of the low temperature. These factors increase the concentration of $PM_{2.5}$ and its precursors: sulfur dioxide, nitrogen oxides, and volatile organic compounds.[7] In addition, the local weather in winter weakens the diffusion of $PM_{2.5}$. Sandstorms in northern China cause the minor peak in spring.[46] For South Korea (Figure 5D), $PM_{2.5}$ has

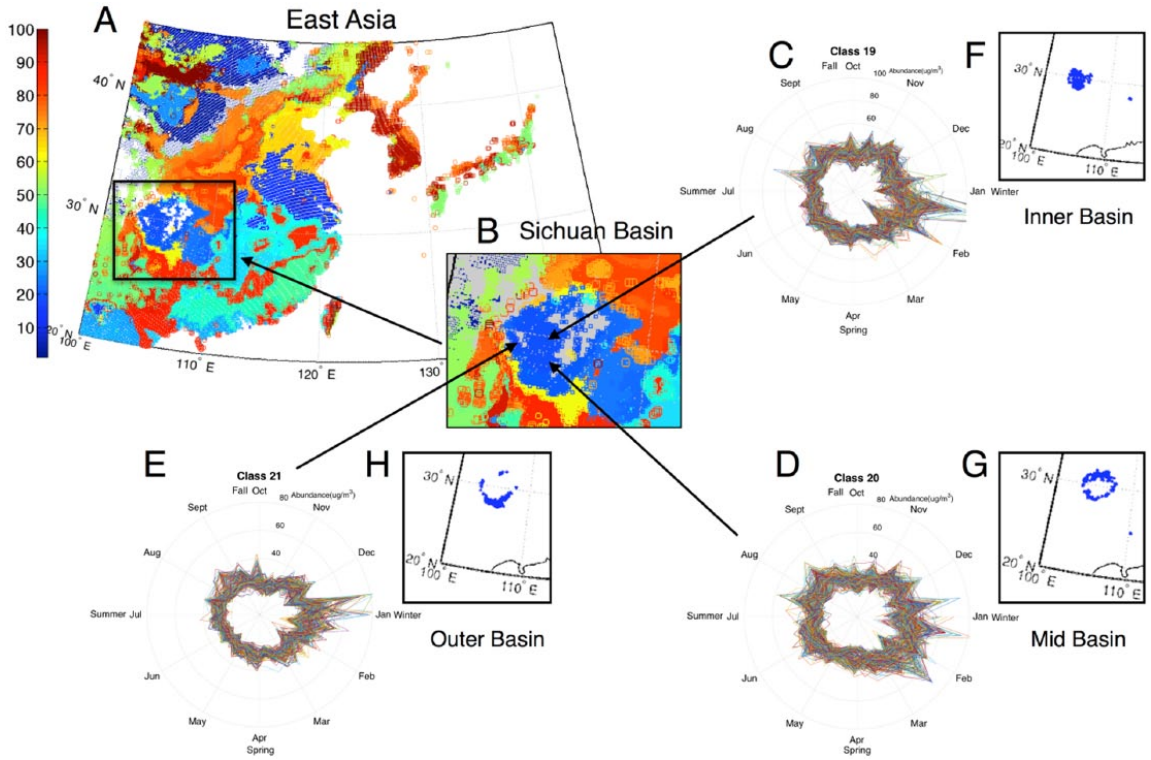

Figure 4. Classification map for the $PM_{2.5}$ annual cycle over East Asia (A), zooms into the Sichuan Basin (B). The shape of the $PM_{2.5}$ annual cycles (shown as polar diagrams) in the Sichuan Basin naturally fall into three regions: inner basin (C,F), mid basin (D, G), and outer basin (E, H).

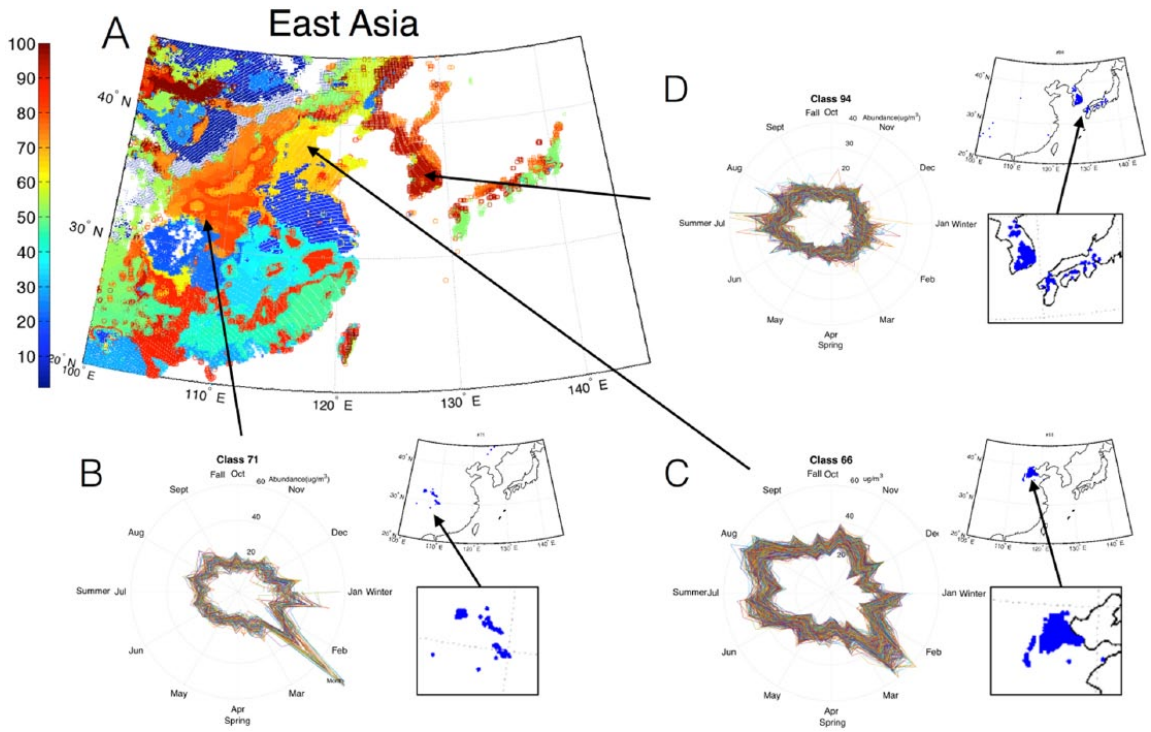

Figure 5. $PM_{2.5}$ annual cycle classification map over East Asia (A) with a focus on the area around Beijing (class 66), central China area (class 71), and South Korea (class 94). (B, C, and D) $PM_{2.5}$ annual cycles as polar diagrams and locations corresponding to classes 66, 71, and 94.

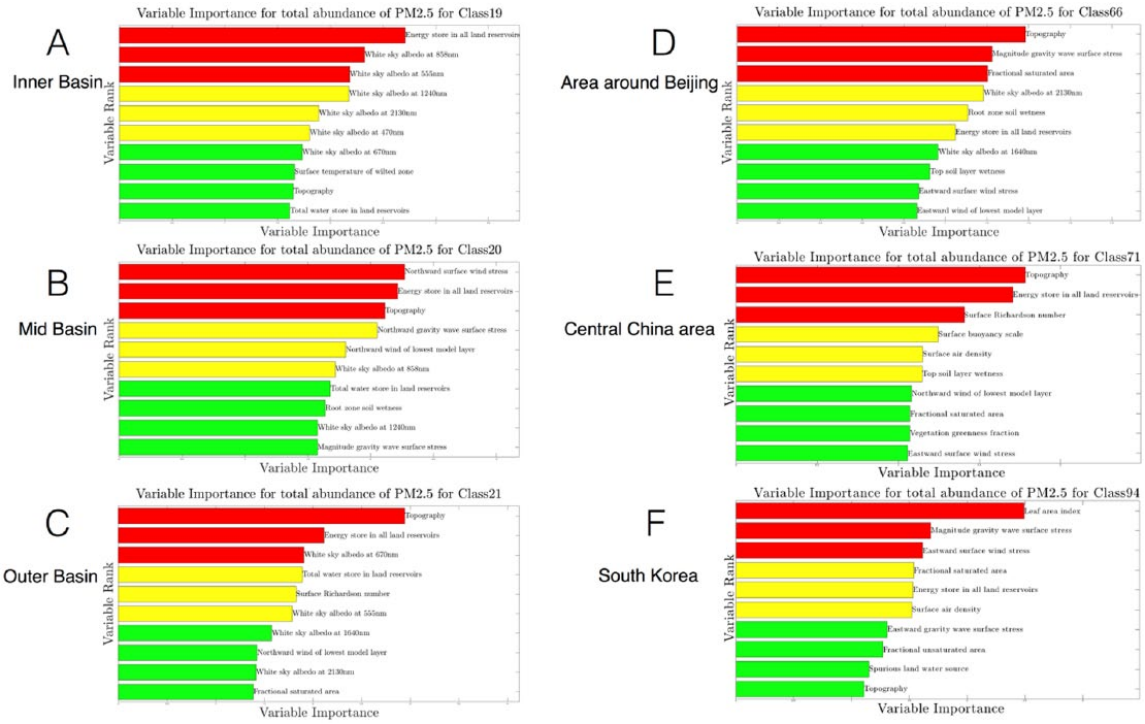

Figure 6. Top 10 most important meteorological variables of these 6 classes are shown. (A, B, and C) Sichuan Basin (A is inner basin, B is mid-basin, and C is outer basin), (D) the area around Beijing, (E) central China area, and (F) South Korea and part of Japan.

lower concentrations throughout the year due to much lower emissions than was recorded for China and due to the ocean breeze transporting and diluting any elevated $PM_{2.5}$.

The rankings of variable importance for each of these regions provided by random forests are shown in Figure 6. For the highly populated inner Sichuan basin (Figure 6A), the top predictive factors for distinguishing the shape of the $PM_{2.5}$ annual cycle from all other classes are as follows: (1) the energy stored in all land reservoirs, which represents the heat content of the land (soil, canopy, and snowpack); (2) white sky albedo at 858 and 555 nm, which represents the reflectance of surface under diffuse solar radiation with wavelengths of 858 and 555 nm; they are related to the surface reflectivity. Other listed factors which play less important roles are related to surface reflectivity at other wavelengths, surface temperature and height, and water content in land. For the mid-basin area (Figure 6B), the top predictive factors are as follows: (1) the northward surface wind stress, (2) the energy stored in all land reservoirs, and (3) the topography. These factors are related to the northward wind, heat content in land, and surface height. Other listed factors are related to surface reflectivity, water content in land, and northward wind. For the outer basin area (Figure 6C), the top predictive factors are as follows: (1) the topography, (2) the energy stored in all the land reservoirs, and (3) the white sky albedo at 670 nm. They are related to the surface height, heat content in land, and surface reflectivity at 670 nm. Other listed factors are related to surface reflectivity, water content in land, and northward wind. It can be seen by comparing Figure 6A to C that topography plays an increasingly important role as one moves from the inner to the outer basin due to the increasing gradient of the surface height.

For the area around Beijing (Figure 6D), the key factors determining the shape of the $PM_{2.5}$ annual cycle are as follows: (1) the topography and (2) the gravity wave surface stress, which means the interface between ocean and land, that is, ocean plays an important role in the coastal area (eg, ocean breeze). Other listed factors are related to surface reflectivity and surface type, eastward wind, and heat content in land. For the central China region (Figure 6E), the key factors are as follows: (1) the topography and (2) the energy stored in all land reservoirs. Other listed factors are related to surface reflectivity, eastward wind, heat content of the lands, and surface type. For South Korea (Figure 6F), the leaf area index plays the most important role and it is a dimensionless quantity that characterizes plant canopies. The gravity wave surface stress and the eastward surface wind stress are also important factors because this area is near ocean, which plays a significant role in determining the morphology of the $PM_{2.5}$ annual cycle. Others listed factors are related to heat content in land, air density above the surface, and surface type.

Conclusions and Future Work

Machine learning has done a remarkable job of both classifying the morphology of the $PM_{2.5}$ annual cycle in East Asia and providing insights into the specific aspects of the physical environment that are associated with the shape of the $PM_{2.5}$ annual cycle and the timing of the $PM_{2.5}$ peaks. This study used a 2-step machine learning methodology where an unsupervised self-organizing map was first used to classify the morphology of the

$PM_{2.5}$ annual cycle into 100 classes. Thereafter, to gain physical insight into the key drivers, a separate supervised random forest for 6 morphology classes out of 100 was used to rank the relative importance of the factors determining the shape of the annual cycle for that class. For the 3 classes in Sichuan Basin, the shapes of the $PM_{2.5}$ seasonal cycle are relatively stable, and the key factors are mostly related to surface type and surface height. The shape of the $PM_{2.5}$ seasonal cycle in the central China area records a peak in winter, and the key factors are related to heat content in land and surface height. However, the $PM_{2.5}$ peak was recorded in summer in South Korea, and the key drivers are related to the ocean effect, plant canopies, and eastward wind. For the $PM_{2.5}$ seasonal cycle in the area around Beijing, the key drivers are related to surface type and height, and ocean effect.

Some natural areas for further work include examining whether there are seasonal population health outcomes that are associated with the peaks of the $PM_{2.5}$ annual cycle identified for various regions.

Author Contributions

DW and DJL conceived and designed the experiments; DW analyzed the data and wrote the first draft of the manuscript; DW, XL, GKZ, and DJL contributed to the writing of the manuscript; DW, DJL, and MAK agree with manuscript results and conclusions; DW, DJL, and MAK jointly developed the structure and arguments for the paper; and DW, XL, GKZ, and DJL made critical revisions and approved the final version. All authors reviewed and approved the final manuscript.

REFERENCES

1. WHO. 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Published 2014. Accessed August 29, 2016.
2. Kampa M, Castanas E. Human health effects of air pollution. *Environ Pollut*. 2008;151:362–367.
3. Pope CA III, Brook RD, Burnett RT, et al. How is cardiovascular disease mortality risk affected by duration and intensity of fine particulate matter exposure? an integration of the epidemiologic evidence. *Air Qual Atmos Health*. 2011;4:5–14.
4. Brook RD, Rajagopalan S, Pope CA, et al. Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the American Heart Association. *Circulation*. 2010;121:2331–2378.
5. Pope CA III, Burnett RT, Thun MJ, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*. 2002;287:1132–1141.
6. Krewski D, Jerrett M, Burnett RT, et al. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. *Res Rep Health Eff Inst*. 2009;140:5–114.
7. Sun Y, Zhuang G, Tang A, et al. Chemical characteristics of $PM_{2.5}$ and PM_{10} in haze-fog episodes in Beijing. *Environ Sci Technol*. 2006;40:3148–3155.
8. Zheng M, Salmon LG, Schauer JJ, et al. Seasonal trends in $PM_{2.5}$ source contributions in Beijing, China. *Atmos Environ*. 2005;39:3967–3976.
9. GREENPEACE. https://www.greenpeace.org.cn/pm25-ranking/. Published 2014. Accessed September 10, 2015.
10. Yu S, Zhang Q, Yan R, et al. Origin of air pollution during a weekly heavy haze episode in Hangzhou, China. *Environ Chem Lett*. 2014;12:543–550.
11. Yu S, Li P, Wang L, et al. Anthropogenic aerosols are a potential cause for migration of the summer monsoon rain belt in China. *Proc Natl Acad Sci U S A*. 2016;113:E2209–E2210.
12. Yan R, Yu S, Zhang Q, et al. A heavy haze episode in Beijing in February of 2014: characteristics, origins and implications. *Atmos Pollut Res*. 2015;6:867–876.
13. Li P, Yan R, Yu S, et al. Reinstate regional transport of $PM_{2.5}$ as a major cause of severe haze in Beijing. *Proc Natl Acad Sci U S A*. 2015;112:E2739–E2740.
14. Stier P, Seinfeld JH, Kinne S, et al. Aerosol absorption and radiative forcing. *Atmos Chem Phys*. 2007;7:5237–5261.
15. Forster P, Ramaswamy V, Artaxo P, et al. Chapter 2. Changes in atmospheric constituents and in radiative forcing. In: Nakajima T, Ramanathan V, eds. *Climate Change 2007. The Physical Science Basis*. Cambridge, UK: Cambridge University Press; 2007;153–180.
16. Ramaswamy V, Boucher O, Haigh J, et al. Radiative forcing of climate. In: Joos F, Srinivasan J, eds. *Climate Change 2001*. Cambridge, UK: Cambridge University Press; 2001:349–416.
17. Hansen J, Sato M, Ruedy R, et al. Radiative forcing and climate response. *J Geophys Res*. 1997;102:6831–6864.
18. Engel-Cox JA, Hoff RM, Haymet A. Recommendations on the use of satellite remote-sensing data for urban air quality. *J Air Waste Manag Assoc*. 2004;54:1360–1371.
19. Hoff RM, Christopher SA. Remote sensing of particulate pollution from space: have we reached the promised land? *J Air Waste Manag Assoc*. 2009;59:645–675.
20. Liu Y, Sarnat JA, Kilaru V, et al. Estimating ground-level $PM_{2.5}$ in the eastern united states using satellite remote sensing. *Environ Sci Technol*. 2005;39:3269–3278.
21. Lyamani H, Olmo F, Alcantara A, et al. Atmospheric aerosols during the 2003 heat wave in southeastern Spain I: spectral optical depth. *Atmos Environ*. 2006;40:6453–6464.
22. Choi YS, Ho CH, Chen D, et al. Spectral analysis of weekly variation in PM_{10} mass concentration and meteorological conditions over China. *Atmos Environ*. 2008;42:655–666.
23. Paciorek CJ, Liu Y. Assessment and statistical modeling of the relationship between remotely sensed aerosol optical depth and $PM_{2.5}$ in the eastern united states. *Res Rep Health Eff Inst*. 2012;167:5–83;discussion 85-91.
24. Hyer E, Reid J, Zhang J. An over-land aerosol optical depth data set for data assimilation by filtering, correction, and aggregation of MODIS collection 5 optical depth retrievals. *Atmos Meas Tech*. 2011;4:379–408.
25. Lary D, Remer L, MacNeill D, et al. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geosci Remote S*. 2009;6:694–698.
26. Reid JS, Hyer EJ, Johnson RS, et al. Observing and understanding the Southeast Asian aerosol system by remote sensing: an initial review and analysis for the Seven Southeast Asian Studies (7SEAS) program. *Atmos Res*. 2013;122:403–468.
27. Shi Y, Zhang J, Reid J, et al. Critical evaluation of the MODIS Deep Blue aerosol optical depth product for data assimilation over North Africa. *Atmos Meas Tech*. 2013;6:949–969.
28. Lary DJ, Faruque FS, Malakar N, et al. Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$). *Geospat Health*. 2014;8:611–630.
29. Melin F, Zibordi G, Carlund T, et al. Validation of SeaWiFS and MODIS Aqua/Terra aerosol products in coastal regions of European marginal seas. *Oceanologia*. 2013;55:27–51.
30. Remer LA, et al. Global aerosol climatology from the MODIS satellite sensors. *J Geophys Res: Atmos*. 2008;113:D14.
31. Lary D, Lary T, Sattler B. Using machine learning to estimate global $PM_{2.5}$ for environmental health studies. *Environ Health Insights*. 2015;9:41–52.
32. Rienecker MM, Suarez MJ, Gelaro R, et al. MERRA: NASA's modern-era retrospective analysis for research and applications. *J Climate*. 2011;24:3624–3648.
33. Hastie T, Tibshirani R, Friedman J. *Unsupervised Learning*. New York, NY: Springer; 2009.
34. Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern*. 1982;43:59–69.
35. Von der Malsburg C. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*. 1973;14:85–100.
36. Kohonen T. The self-organizing map. *Neurocomputing*. 1998;21:1–6.
37. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. *IEEE T Neural Networ*. 2000;11:586–600.
38. Breiman L. Random forests. *Machine Learn*. 2001;45:5–32.
39. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens*. 2005;26:217–222.
40. Strobl C, Boulesteix AL, Zeileis A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8:25.
41. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. *Pattern Recogn Lett*. 2010;31:2225–2236.
42. Liaw A, Wiener M. Classification and regression by randomforest. *R News*. 2002;2:18–22.
43. Tai AP, Mickley LJ, Jacob DJ. Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: implications for the sensitivity of pm 2.5 to climate change. *Atmos Environ*. 2010;44:3976–3984.
44. Xu J, Ding G, Yan P, et al. Componential characteristics and sources identification of $PM_{2.5}$ in Beijing. *J Appl Meteorol Sci*. 2007;18:645–654.
45. Chan CK, Yao X. Air pollution in mega cities in china. *Atmos Environ*. 2008;42:1–42.
46. Zhu X, Zhang Y-H, Zeng L, et al. Source identification of ambient $PM_{2.5}$ in Beijing. *Res Environ Sci*. 2005;18:1–5.