A new approach for determining optimal placement of PM$_{2.5}$ air quality sensors: case study for the contiguous United States

Makoto M Kelp$^{1,*}$, Samuel Lin$^2$, J Nathan Kutz$^3$ and Loretta J Mickley$^4$

1 Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, United States of America
2 Department of Computer Science, Harvard University, Cambridge, MA 02138, United States of America
3 Department of Applied Mathematics, University of Washington, Seattle, WA 98195, United States of America
4 John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 01238, United States of America

* Author to whom any correspondence should be addressed.

E-mail: mkelp@g.harvard.edu

Keywords: fine particulate matter (PM$_{2.5}$), sensor placement, multiscale dynamics

Abstract

Considerable financial resources are allocated for measuring ambient air pollution in the United States, yet the locations for these monitoring sites may not be optimized to capture the full extent of current pollution variability. Prior research on best sensor placement for monitoring fine particulate matter (PM$_{2.5}$) pollution is scarce: most studies do not span areas larger than a medium-sized city or examine timescales longer than 1 week. Here we present a pilot study using multiresolution dynamic mode decomposition (mrDMD) to identify the optimal placement of PM$_{2.5}$ sensors from 2000 to 2016 over the contiguous United States. This novel approach incorporates the variation of PM$_{2.5}$ on timescales ranging from 1 d to over a decade to capture air pollution variability. We find that the mrDMD algorithm identifies more high-priority sensor locations in the western United States than those expected along the eastern coast, where a large number of Environmental Protection Agency (EPA) PM$_{2.5}$ monitors currently reside. Specifically, 53% of mrDMD optimized sensor locations are west of the 100th meridian, compared to only 32% in the current EPA network. The mrDMD sensor locations can capture PM$_{2.5}$ from wildfires and high pollution events, with particularly high skill in the west. These results suggest significant gaps in the current EPA monitoring network in the San Joaquin Valley in California, northern California, and in the Pacific Northwest (Idaho, and Eastern Washington and Oregon). Our framework diagnoses where to place air quality sensors so that they can best monitor smoke from wildfires. Our framework may also be applied to urban areas for equitable placement of PM$_{2.5}$ monitors.

1. Introduction

Air pollution consisting of fine particulate matter (PM$_{2.5}$) presents a major environmental risk to public health, causing 4.1 million premature deaths worldwide in 2019 (Murray et al 2020). Given the financial cost and limited resources allocated for measuring ambient air pollution, however, optimizing sensor placement is a central mathematical challenge. Monitoring PM$_{2.5}$ is especially difficult as there are many outdoor sources including direct emissions from wildland and agricultural fires (Johnston et al 2012, Casworth et al 2018), incomplete fuel combustion from vehicles and industrial processes (Bond et al 2007), and secondary chemical formation from gas-phase precursors of anthropogenic (McDuffie et al 2021) or biogenic (Goldstein et al 2009) origin. Furthermore, the formation and transport of PM$_{2.5}$ are both sensitive to meteorological conditions from the local to synoptic scales (Tai et al 2012), including, for example, convective events and horizontal winds that can disperse pollution as well as precipitation that can immediately remove particles. Here we provide an optimization framework that takes into account the
multiscale variability of PM$_{2.5}$ to determine the placement of a minimum number of air pollution sensors for maximal spatiotemporal coverage.

In the United States, monitoring of air pollution has historically relied on the Environmental Protection Agency (EPA) network of sampling sites (US EPA 2013). These monitoring locations are used to assess local and regional attainment of the National Ambient Air Quality Standard (NAAQS), to analyze air pollution impacts on public health, and to validate satellite measurements and air quality models. However, these sites are not equally distributed across the United States and may not adequately sample the full range of concentrations (Di et al 2019). A greater number of sites are in the eastern United States, along the western coast, and in urban areas, with relatively fewer sites in mountainous regions and rural areas. Due in part to the high cost of maintenance, new monitoring sites are added infrequently (US EPA 2020).

Improvements in low-cost sensor technologies coupled with the rise of citizen science air quality monitoring offers potential for greater spatial coverage of pollution observations, but even so these sensor locations may not provide adequate coverage. Many low-cost sensors can report measurements publicly in real-time (Snyder et al 2013, Kumar et al 2015, Parmar et al 2017). These sensors, deployed by citizen science and crowdsourced organizations such as Purple Air (Barkjohn et al 2021), represent a core component of an internet of things network for monitoring air pollution in cities (Dhingra et al 2019, Toma et al 2019). However, this promising democratization of air pollution monitoring may inadequately capture the spatial distribution and variability of PM$_{2.5}$. Crowdsourced datasets often rely on volunteers who are responsible for installation and upkeep of each sensor, resulting in deployment in predominantly white areas characterized by higher incomes and levels of education relative to the US census tracts with EPA monitors (deSouza and Kinney 2021). Furthermore, areas with a greater density of low-cost sensors report lower annual-average PM$_{2.5}$ concentrations than the EPA monitors in all states except California (deSouza and Kinney 2021). To be sure, such citizen science efforts are well-intentioned, but they exacerbate disparities in the spatial coverage of PM$_{2.5}$ monitors and limit the pursuit of environmental justice (Sorensen et al 2019).

Few studies have considered both spatial features and temporal dynamics of air pollution in determining the optimal placement of outdoor PM$_{2.5}$ sensors. While strategies based on computational fluid dynamics can locate optimal placement of indoor sensors (Löhner and Camelli 2005, Waeytens and Sadr 2018), such approaches are infeasible over large atmospheric domains. Mukherjee et al (2020) developed an optimal sensor placement algorithm for mobile monitoring of PM$_{2.5}$ and SO$_2$ in Atlanta, Georgia, but such an algorithm can be applied for just a short time period (hours to a day). For a case study in Cambridge, UK, Sun et al (2019) created an algorithm that considers areas with PM$_{2.5}$ hotspots such as high-traffic roads and locations such as hospitals and elderly care homes where more vulnerable populations reside. However, this framework does not guarantee capturing representative PM$_{2.5}$ concentrations across a broader region. Here we build on prior studies that determined the optimal placement of sensors using modal decomposition techniques in the domains of ocean monitoring (Yildirim et al 2009), fluid dynamics (Bai et al 2017), and neuroscience (Brunton et al 2016). Application of such techniques is yet nascent for air pollution, which is characterized by its own unique multiscale behavior. To our knowledge, there exist no studies of optimal air pollution sensor placement in the United States that span regions larger than an urban area or timescales longer than 1 week.

In this pilot study, we demonstrate a data-driven approach that determines the optimal placement of sensors to capture PM$_{2.5}$ concentrations and variability across the contiguous United States on timescales ranging from days to over a decade. We employ multisiresolution dynamic mode decomposition (mrDMD), which recursively decomposes a dataset into low-rank spatial modes and their temporal Fourier dynamics (Kutz et al 2016, Manohar et al 2019). This algorithm allows us to create a library of modes that not only captures PM$_{2.5}$ concentrations spatially and temporally on short (daily) and long-term (years to decade) timescales, but also incorporates information from significant transient phenomena, such as wildfires and temperature inversions, that would otherwise be discarded or averaged out using similar data reduction techniques. We emphasize that the current distribution of over 2000 EPA monitors is needed to keep track of adherence to the NAAQS. However, the modeling approach and results presented here may be used to locate gaps in the EPA sensor network and inform locations of future PM$_{2.5}$ monitors.

2. Data and methods

2.1. PM$_{2.5}$ dataset

We use a dataset consisting of daily PM$_{2.5}$ concentrations for the contiguous United States for the period January 2000 to December 2016. The dataset was produced through a data fusion method involving ensemble machine learning that combines surface monitoring measurements, satellite aerosol optical depth, land-use data, and chemical transport model results, among other variables (Di et al 2019, 2021). The high spatial and temporal resolution of this PM$_{2.5}$ dataset reveals spatial variability across the United States, including a stark east–west gradient, with the eastern United States experiencing relatively higher
PM$_{2.5}$ concentration than the west. Hotspots in the dataset reveal the wide diversity of PM$_{2.5}$ sources, such as emissions from power plants in the Midwest, organic aerosol produced from biogenic volatile organic compounds in the southeast, and wildfires and associated smoke in the west. We coarsen the dataset’s 1 km × 1 km resolution to 10 km × 10 km to yield a spatial grid of 871 × 413 cells. Missing values and erroneous observations are removed from the dataset by applying a mask to the spatial grid.

2.2. Dynamic mode decomposition (DMD)

To isolate the underlying multiscale variability of PM$_{2.5}$, we use a DMD modeling approach. The DMD method provides a spatiotemporal decomposition of data into a set of dynamic modes derived from snapshots or measurements of a given system in time. DMD is similar to principal components analysis (PCA) as they are both dimensionality reduction algorithms that take advantage of underlying low-rank features found in large datasets. The resulting modes of these methods reveal the dominant spatial patterns that account for variation in the dataset. However, unlike PCA, which does not model temporal dynamics of time-series data explicitly, DMD computes a set of modes with an exponential temporal dependence whose imaginary components model an oscillating time frequency characterizing the modal dynamics. Importantly, DMD can accommodate heterogeneous spatial and temporal sampling. For noisy datasets, DMD can be assembled in order to produce uncertainty metrics (Sashidhar and Kutz 2021). Many variants of DMD exist (Tissot et al 2014, Noack et al 2015, Askham and Kutz 2017, Li et al 2017, Sashidhar and Kutz 2021), but here we follow Manohar et al (2018), as described in the supplement section 1 (available online at stacks.iop.org/ERL/17/034034/mmedia). We apply the DMD method to the PM$_{2.5}$ dataset to construct a library of modes that reveals the dominant spatiotemporal patterns in the dataset.

2.3. Multiresolution dynamic mode decomposition (mrDMD)

Recently, a multiresolution extension to the DMD algorithm has been proposed (Kutz et al 2016, Manohar et al 2019), which allows for the incorporation of significant transient phenomenon into the DMD library. Much like a multiresolution analysis using wavelets (Kutz 2013), the mrDMD algorithm recursively removes low frequency, or slowly varying, signals to allow separation of background (i.e. average PM$_{2.5}$ concentrations) from foreground (i.e. high pollution episodes) in the PM$_{2.5}$ dataset on varying timescales specified by the user. The mrDMD algorithm is more precise in capturing spatiotemporal variability than methods based on singular value decomposition such as DMD and PCA (Manohar et al 2019).

The number of snapshots $M$ from our dataset are chosen so that the DMD modes provide a full rank approximation of the dynamics observed. Thus, $M$ should be representative of all high- and low-frequency content of the dataset. The DMD algorithm is then applied recursively on the dataset. The slowest $m1$ modes are removed in the initial pass, and remaining snapshots are divided into $M/2$ segments and are fed into the DMD algorithm again. The slowest $m2$ modes are removed, and this process is continued until a termination condition is reached. The mrDMD separates the DMD approximate solution in the first pass as follows:

$$X_{\text{mrDMD}}(t) = \sum_{k=1}^{M} b_k(0) \psi_k^{(1)}(\xi) \exp(\omega_k t)$$

$$= \sum_{k=1}^{m1} b_k(0) \psi_k^{(1)}(\xi) \exp(\omega_k t) \quad \text{(slow modes)}$$

$$+ \sum_{k=m1+1}^{M} b_k(0) \psi_k^{(1)}(\xi) \exp(\omega_k t) \quad \text{(fast modes)}$$

where $b_k(0)$ is the initial amplitude, $\xi$ is the spatial coordinates of each mode, $\psi_k^{(1)}$ represents the DMD modes (eigenvectors) obtained from the complete snapshot matrix, and $\exp(\omega_k t)$ are the corresponding eigenvalues. After the first level of decomposition, the fast modes are again decomposed into two matrices:

$$X_{M/2} = X_{M/2}^{(1)} + X_{M/2}^{(2)}.$$  

At this level of decomposition, the first matrix $X_{M/2}^{(1)}$ represents slower modes and $X_{M/2}^{(2)}$ represents faster modes, which undergo decomposition again. This process is repeated recursively by removing slower modes in each iteration. At each level of decomposition, the separation of the foreground mode from the background is deemed significant if its eigenvalues exceed a user-set tolerance. Here, we set this tolerance to a standard value of $1 \times 10^{-2}$.

A formal expansion of the mrDMD theory and modeling approach may be found elsewhere (Kutz et al 2016, Manohar et al 2019).

The mrDMD algorithm operates by decreasing the time domain by a factor of two at each successive decomposition level. We apply mrDMD to training windows starting at 11.4 years ($M = 4096$ d), followed by 13 decomposition levels so that the shortest frequency is daily. This approach thus yields a long-term mode characterizing the average PM$_{2.5}$ concentrations over 11.4 years, with potential identification of
transient pollution events spanning timescales from 5.7 years (2048 d) to 1 d. To analyze all PM$_{2.5}$ data from 2000 to 2016 employing these decomposition levels, we divide the dataset into two time-windows: January 2000 to March 2011 (first 4096 d) and September 2005 to December 2016 (last 4096 d).

2.4. Optimal sensor placement

We use the matrix libraries containing all DMD and mrDMD modes as tailored basis sets $\psi_r \in \mathbb{R}^{m\times r}$ to optimize for sensor placement. We identify the optimal sensor locations by employing QR pivoting to our DMD and mrDMD basis sets (Heck et al 1998, Manohar et al 2018). QR pivoting is a ‘greedy’ selection algorithm that is computationally efficient for finding near-optimal sensor locations. Greedy approaches are often favored over other optimization techniques as the true optimal solution often involves a combinatorially intractable optimization.

QR column pivoting identifies rows in the modal library $\psi_r$ with the highest two-norm, which corresponds to locations with the largest PM$_{2.5}$ modal frequencies and therefore greatest variability. The reduced matrix QR factorization with column pivoting decomposes a matrix $A \in \mathbb{R}^{m\times n}$ into a unitary matrix $Q$, an upper-triangular matrix $R$, and a column permutation matrix $C^T$ such that $AC^T = QR$. Thus, the QR factorization with column pivoting yields $r$ point sensors ( pivots) that best sample the $r$ tailored basis modes $\psi_r$:

$$\psi_r^T C^T = QR. \quad (3)$$

That is, each QR pivot identifies those spatial locations in the modal library that exhibit the most variability and where sensor placement captures significant pollution episodes above background concentrations.

A formal expansion of the sparse sensor placement approach may be found elsewhere (Kutz et al 2016, Manohar et al 2018, 2019, Brunton and Kutz 2019).

In the supplement, we reconstruct the full field of PM$_{2.5}$ concentrations across the contiguous United States using measurements from only those sensor locations diagnosed as optimal in the step above. These reconstructions serve as validation of the DMD and mrDMD libraries and the ability of these algorithms to select sensor locations with high information content.

3. Results and discussion

Here we present the mrDMD results and analysis. The modal library of DMD is less interpretable than that of mrDMD, and we present the DMD results and analysis in the supplement.

3.1. mrDMD modes describing PM$_{2.5}$ variability

The mrDMD modal library reveals that most significant air pollution episodes occur on timescales between 1 week and 1 month, with many of these events arising from wildfires. We diagnose these wildfire modes by visual inspection based on regions with known fire activity and minimal industrial PM$_{2.5}$ sources. Figure 1 presents the resulting mrDMD modal maps in the time-frequency domain. Figure 1(a) corresponds to the first 11.4 years of the time period (January 2000 to March 2011) and figure 1(b) corresponds to the last 11.4 years (September 2005 to December 2016). The background mode in both training windows corresponds to the mean PM$_{2.5}$ concentration distribution for that window. Both background modes display a sharp east–west gradient with most of the eastern United States, except the Appalachian Mountains and some remote areas of Maine, having relatively higher PM$_{2.5}$ concentration than the west. This pattern agrees with the observed annual mean spatial distribution of PM$_{2.5}$ (Tai et al 2010, Di et al 2019). The background mode for the second half of the training period exhibits smaller modal amplitudes in hotspot regions than the first half. This trend is consistent with Di et al who reported that PM$_{2.5}$ concentrations decreased noticeably after 2008 due to a combination of the economic recession and emission controls on coal-fired power plants. The 2000–2008 annual average PM$_{2.5}$ concentration across the contiguous United States was 7.8 $\mu$g m$^{-3}$; the 2009–2016 annual average was 6.2 $\mu$g m$^{-3}$.

The mrDMD method also identifies relatively short-term air pollution episodes on timescales of less than a year. Most episodes occur on timescales between 1 week and 1 month. However, we find an especially long-lived mode over Los Angeles and the San Joaquin Valley, California, lasting for 256 d from September 2007 to May 2008. This mode may correspond to intense winter PM$_{2.5}$ due to temperature inversions trapping emissions in the southern California basin. From 2007 to 2008, the San Joaquin Valley reported significantly higher PM$_{2.5}$ concentrations than are typical, with 66 d above the 24 h 35 $\mu$g m$^{-3}$ standard in 2008 (US EPA 2021). The 2007 Murphy Complex Fire and other wildfires in Idaho are captured as a significant mode occurring on a timescale of 64 d from June to August. Although the lifetime of PM$_{2.5}$ in the atmosphere is ~1 week and is sensitive to meteorology such as rainout, no pollution modes occur on timescales less than a week. We attribute the absence of modes on this timescale to the eigenvalue tolerance used in the mrDMD algorithm. If we tighten the tolerance, then more significant modes would be found at daily timescales, but such a tolerance would result in higher-frequency (i.e. noisy) and less-interpretable modal maps, making it more difficult to discern prolonged pollution events. Overall, we detect 72 significant pollution modes during the
first time window and 77 during the second time window. We find 27 wildfire modes (38% of all modes) in the first window and 46 wildfire modes (60% of all modes) in the second window.

Based on figure 1, we conclude that the mrDMD framework can distinguish high pollution episodes from the mean PM$_{2.5}$ concentrations through time. The mrDMD framework provides a background mode that agrees with Di et al (2019), but also contains modal snapshots of events that stand out from this background.

3.2. Optimal sensor locations

By applying QR pivots to our mrDMD modes, we can identify a set of 1369 optimal sensor locations across the contiguous United States, a number of roughly the same magnitude as the number of EPA sensors. Our method assumes that air pollution spatiotemporal patterns and concentration variability from 2000 to 2016 are representative of future conditions. We argue that this assumption is reasonable, as discussed below. Figure 2 shows that the mrDMD sensor distribution is more clustered in California and across the west,
along the Gulf Coast, and in the Industrial Midwest and Northeast, relative to that of the EPA. The mrDMD sensors thus reflect those regions characterized by higher PM$_{2.5}$ variability, with few sensors in the central United States. These results suggest that the west is a driver of PM$_{2.5}$ variability in the United States due to periodic wildfires, distributed urban sources isolated by rugged terrain, and reduced meteorological mixing. For example, the mrDMD algorithm yields a cluster of sensors in Los Angeles and Salt Lake City as these urban areas are basins surrounded by mountains where the meteorological conditions favor the production and trapping of PM$_{2.5}$ pollution. Additionally, we present a sparse distribution of sensors that can capture the overall spatiotemporal variability of PM$_{2.5}$ in the supplement.

The mrDMD method identifies significantly more sensors in the San Joaquin Valley and in northern California, and the Pacific Northwest than the current EPA sensor network (figure 3). In particular, the San Joaquin Valley dominates the modal signatures of the mrDMD library (figure 1) and is
characteristic of high PM$_{2.5}$ variability due to large-scale agricultural activities, smoke from wildfires, and frequent temperature inversions (Wolyn and McKee 1989, Wei et al 2013, Rémy et al 2017). The greater density of mrDMD sensors in northern California and the Pacific Northwest (especially in Idaho and the eastern parts of Oregon and Washington) can be attributed to their sensitivity to smoke PM$_{2.5}$ from wildfires. Indeed, we find that wildfires constitute 60% (46 out of 77) of significant pollution modes in the mrDMD library from September 2005 to December 2016. Also, mrDMD identifies locations along the Gulf Coast (i.e. the eastern shoreline of Texas) that are not covered by the EPA network (figure S1).

Our results suggest that the EPA monitoring network does not optimally capture the spatial patterns and variability of PM$_{2.5}$ pollution in the contiguous United States today. We posit that the current locations of EPA sensors were informed by the legacy of air pollution from the 1980s to 1990s when the eastern United States had a higher regional PM$_{2.5}$ background with large variability in PM$_{2.5}$ concentrations arising from interactions between meteorology and emissions from coal-fired power plants, diesel emissions, and industrial activities (Bachmann 2007). However, with the passage of the Clean Air Act (and subsequent amendments), stricter air quality regulations have reduced these emissions (Bachmann 2007). Urban air pollution levels in turn declined nation-ally while at the same time the frequency and magnitude of wildfires in the west have increased significantly (Abatzoglou and Williams 2016). Our mrDMD sensor locations underscores this shift to the west as 53% of sensors (726 out of 1369) are found west of the 100th meridian (Seager et al 2017) compared to 32% (694 out of 2156) for the current network of EPA monitors. We emphasize that the current distribution EPA monitors is needed to keep track of adherence to the NAAQS, especially in densely populated regions. We do not advocate for the removal of existing PM$_{2.5}$ sensors, but rather for the deployment of more sensors in the west (especially in the San Joaquin Valley and Pacific Northwest) to capture high-pollution variability.

4. Conclusions

We present a pilot study using a data-driven approach to identify the optimal placement of PM$_{2.5}$ sensors to capture air pollution over the contiguous United States at 10 km $\times$ 10 km resolution from 2000 to 2016. While earlier studies are limited in spatial and temporal scope, this is the first national-scale study that diagnoses the optimal placement of PM$_{2.5}$ sensors. Our method employs multiresolution dynamic mode decomposition (mrDMD), an approach that incorporates the variability of PM$_{2.5}$ on timescales ranging from 1 d to over a decade. We identify 72 significant pollution modes from January 2000 to March 2011 and 77 pollution modes from September 2005 to December 2016, ranging in duration from 8 to 256 d. Smoke from wildfire events comprise 38% and 60% of these modes during these respective time frames. Our results underscore the large impact that wildfires
have on PM$_{2.5}$ concentrations and variability, especially in the west.

The mrDMD modal library identifies 1369 optimal sensor locations in the western United States, with significantly lower density of sensors in the eastern coast. We show that 53% of mrDMD sensor locations are found west of the 100th meridian compared to only 32% for the current network of EPA monitors. We hypothesize that the current network of EPA monitors was designed to capture PM$_{2.5}$ spatial patterns during the 1980s–1990s when urban air pollution and high background concentrations in the East, coupled with meteorological factors, represented a significant driver of PM$_{2.5}$ variability. However, with reduced emissions from increased air quality regulations, our results suggest a shift to the west as the main driver of PM$_{2.5}$ variability across the contiguous United States. This shift can be traced to periodic wildfires and to distributed urban sources isolated by rugged terrain. We suggest that the mrDMD sensors identified here may help capture high PM$_{2.5}$ episodes not adequately monitored by EPA sensors. We acknowledge, however, that our method assumes that air pollution spatiotemporal patterns and concentration variability from 2000 to 2016 are representative of future conditions. We argue that this assumption is reasonable, as we expect wildfires to continue to be a significant source of PM$_{2.5}$ and that our optimal sensor locations capture that variability.

More specifically, our mrDMD results recommend adding PM$_{2.5}$ sensors in the San Joaquin Valley in California, northern California, and in the Pacific Northwest (Idaho and the eastern parts of Oregon and Washington). Agricultural activities, temperature inversions, and fires drive pollution variability in the San Joaquin Valley and wildfire smoke drives variability in northern California and the Pacific Northwest. All three regions have large populations of non-white Hispanic agricultural workers who are disproportionately affected by an underserved PM$_{2.5}$ monitoring network (Garcia 2007, Chandrasekaran 2021, Marlier et al 2022). The disparity in sensor locations is especially apparent in the San Joaquin Valley where the EPA network of monitors is relatively sparse (figure 3).

Smoke from wildfires represents a significant human health and ecosystem hazard. With the frequency and magnitude of such fires predicted to increase in the future, identifying where to place sensors to track smoke plumes is becoming especially important (McClure and Jaffe 2018, David et al 2021). The mrDMD framework skillfully diagnoses where to place air quality sensors that can best monitor smoke plumes, especially those from large fires. We posit that such a sensor distribution may help stakeholders weigh the benefits and hazards of prescribed fires.

Although we identify optimal PM$_{2.5}$ sensor locations on a national scale, our framework may also be impactful for regional and local domains. With the proliferation of low-cost sensors and citizen science efforts, the techniques provided here could help inform high utility areas to place sensors. For example, the original 1 km $\times$ 1 km PM$_{2.5}$ dataset by Di et al (2021) may be applied to investigate sensor placement at the regional and urban scale. Extensions to the DMD framework could incorporate sensor measurements with different signal-to-noise ratios—i.e. high-cost and low-cost sensors (Clark et al 2020)—or could consider cost-constraining functions that optimize sensor placement based on topography, population density, or metrics related to environmental justice (Clark et al 2019).

**Data availability statement**

The DMD algorithm, mrDMD algorithm, and 10 km $\times$ 10 km PM$_{2.5}$ dataset from 2000 to 2016 that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.5809461.

**Acknowledgments**

This work was partly funded by the US Environmental Protection Agency (EPA) Grant 83587201. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the EPA. We thank Yaguang Wei for technical assistance and John Bachmann for valuable discussion.

**ORCID iDs**

Makoto M Kelp @ https://orcid.org/0000-0003-0451-3439
J Nathan Kutz @ https://orcid.org/0000-0002-6004-2275
Loretta J Mickley @ https://orcid.org/0000-0002-7859-3470

**References**

Abatzoglou J T and Williams A P 2016 Impact of anthropogenic climate change on wildfire across western US forests Proc. Natl Acad. Sci. 113 11770–5
Ashkham T and Kutz J N 2017 Variable projection methods for an optimized dynamic mode decomposition (arXiv:1704.02343 [math.NA])
Bachmann J 2007 Will the circle be unbroken: a history of the US National Ambient Air Quality Standards J. Air Waste Manage. Assoc. 57 652–97
Bai Z, Brunton S L, Brunton B W, Kutz J N, Kaiser E, Spohn A and Noack B R 2017 Data-driven methods in fluid dynamics: sparse classification from experimental data Whither Turbulence and Big Data in the 21st Century? ed A Pollard, I Castillo, I Danaila and M Glauser (Cham: Springer International Publishing) pp 323–42
Barkjohn K K, Gantt B and Clements A L 2021 Development and application of a United States-wide correction for PM$_{2.5}$ data collected with the PurpleAir sensor Atmos. Meas. Tech. 14 4617–37
Bond T C, Bhardwaj E, Dong R, Jogiari R, Jung S, Roden C, Streets D G and Trautmann N M 2007 Historical emissions of black and organic carbon aerosol from energy-related combustion, 1850–2000 Glob. Biogeochem. Cycles 21 GB2018
Brunton B W, Johnson I A, Ojemann J G and Kurtz J N 2016 Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition J. Neurosci. Methods 258 1–15
Brunton S L and Kurtz J N 2019 Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control (Cambridge: Cambridge University Press)
Chandrasekaran P R 2021 Remaking ‘the people’: immigrant farmworkers, environmental justice and the rise of environmental populism in California’s San Joaquin Valley J. Rural Stud. 82 393–405
Clark E, Ashkan T, Brunton S L and Nathan Kurtz J 2019 Greedy sensor placement with cost constraints IEEE Sens. J. 19 2642–56
Clark E, Brunton S L and Kurtz J N 2020 Multi-fidelity sensor selection: greedy algorithms to place cheap and expensive sensors with cost constraints (arXiv:2005.03650) [eess.SP]
Caswell D H, Mickley L J, Sulprizio M P, Liu T, Marlier M E, DeFries R S, Guttikunda S K and Gupta P 2018 Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India Environ. Res. Lett. 13 044018
David L M, Ravishankara A R, Brey S J, Fischer E V, Volckens J and Kreidenweis S 2021 Could the exception become the rule? ‘Uncontrollable’ air pollution events in the US due to wildland fires Environ. Res. Lett. 16 034029
deSouza P and Kinney P L 2021 On the distribution of low-cost PM2.5 sensors in the US: demographic and air quality associations J. Exposure Sci. Environ. Epidemiol. 31 514–24
Dhingra S, Madda R B, Gandomi A H, Patan R and Dhingra S 2019 Internet of things mobile-air pollution monitoring system (IoT-Mobair) IEEE Internet Things J. 6 5577–84
Di Q et al 2019 An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution Environ. Int. 130 104999
Di Q et al 2021 Daily and annual PM2.5 concentrations for the contiguous United States, 1-km grids, v1 (2000–2016) (https://doi.org/10.7927/RVR-4538)
Garcia J 2007 Mexicans in North Central Washington (San Francisco, CA: Arcadia Publishing)
Goldstein A H, Koven C D, Heald C L and Fung I Y 2009 Biogenic carbon and anthropogenic pollutants combine to form a cooling haze over the southeastern United States Proc. Natl. Acad. Sci. 106 8833–40
Heck L P, Olkin J A and Naghshinleh K 1998 Transducer placement for broadband active vibration control using a novel multidimensional QR factorization J. Vib. Acoust. 120 663–70
Johnston F H, Henderson S B, Chen Y, Randerson J T, Marlier M, DeFries R S, Kinney P, Bowman D M J S and Brauer M 2012 Estimated global mortality attributable to smoke from landscape fires Environ. Health Perspect. 130 695–701
Kumar P, Morawiska L, Martani C, Biskos G, Neophytou M, di Sabatino S, Bell M, Norford L and Britter R 2015 The rise of low-cost sensing for managing air pollution in cities Environ. Int. 75 199–205
Kutz J N 2015 Data-driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data (Oxford: Oxford University Press)
Kutz J N, Fu X and Brunton S L 2016 Multiresolution dynamic mode decomposition SIAM J. Appl. Dyn. Syst. 15 713–35
Li Q, Dietrich F, Bollt E M and Kevrekidis I G 2017 Extended dynamic mode decomposition with dictionary learning: a data-driven adaptive spectral decomposition of the Koopman operator Chaos 27 103111
Löhner R and Camelli F 2005 Optimal placement of sensors for contaminant detection based on detailed 3D CFD simulations Eng. Comput. 22 260–73
Manohar K, Brunton B W, Kurtz J N and Brunton S L 2018 Data-driven sparse sensor placement for reconstruction: demonstrating the benefits of exploiting known patterns IEEE Control Syst. Mag. 38 63–86
Manohar K, Kaiser E, Brunton S L and Kurtz J N 2019 Optimized sampling for multiscale dynamics Multiscale Model. Simul. 17 117–36
Marlier M E, Brenner K I, Liu J C, Mickley L J, Raby S, James E, Ahmadov R and Riden H 2022 Exposure of agricultural workers in California to wildfire smoke under past and future climate conditions Environ. Res. Lett. (in review)
McCuller C D and Jaife D A 2018 US particulate matter air quality improves except in wildfire-prone areas Proc. Natl. Acad. Sci. 115 7901–6
McDuffie E E et al 2021 Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales Nat. Commun. 12 3594
Mukherjee R, Diwekar U M and Kumar N 2020 Real-time optimal spatiotemporal sensor placement for monitoring air pollutants Clean Technol. Environ. Policy 22 1091–105
Murray C J L et al 2020 Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019 Lancet 396 1223–49
Noack B R, Stankiewicz W, Morzynski M and Schmid P J 2015 Recursive dynamic mode decomposition of a transient cylinder wake (arXiv:1511.06876 [physics.flu-dyn])
Parmar G, Lakhani S and Chattopadhyay M K 2017 An IoT based low cost air pollution monitoring system 2017 Int. Conf. on Recent Innovations in Signal Processing and Embedded Systems (RISE) pp 524–8
Rémy S et al 2017 Two global data sets of daily fire emission injection heights since 2003 Atmos. Chem. Phys. 17 2921–42
Sashidhar D and Kurtz J N 2021 Bagging, optimized dynamic mode decomposition (BOP-DMD) for robust, stable forecasting with spatial and temporal uncertainty-quantification (arXiv:2107.10878 [cs.ILG])
Seager R, Feldman J, Liu N, Ting M, Williams A P, Nakamura J, Liu H and Henderson N 2017 Whither the 100thth meridian? The one and future physical and human geography of America’s arid–humid divide. Part II: the meridian moves east Earth Interact. 22 1–24
Snyder E G, Watkins T H, Solomon P A, Thoma E D, Williams R W, Hagler G S W, Shelow D, Hindin D A, Kilaru V J and Preus W P 2013 The changing paradigm of air pollution monitoring Environ. Sci. Technol. 47 11 369–77
Sorensen A E, Jordan R C, LeDau S L, Biehler D, Wilson S, Pitas J H and Leisnham P T 2019 Optimal sensor placement for air quality monitoring: a case study of city of Cambridge, the United Kingdom IEEE Access 7 13623–37
Tai A P K, Mickley L I and Jacob D J 2010 Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: implications for the sensitivity of PM2.5 to climate change Atmos. Environ. 44 3976–84
Tai A P K, Mickley L I, Jacob D J, Leibensperger E M, Zhang L, Fisher J A and Pye H O T 2012 Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: implications for PM2.5 sensitivity to climate change Atmos. Chem. Phys. 12 3131–45
Tissot G, Cordier L, Benard N and Noack B R 2014 Model reduction using dynamic mode decomposition C. R. Mecanique 342 410–6
Toma C, Alexandru A, Popa M and Zamfiroua I 2019 IoT solution for smart cities’ pollution monitoring and the security challenges Sensors 19 3401
US EPA 2013 Air quality system (AQS) (available at: www.epa.gov/aqs) (Accessed 21 September 2021)

US EPA 2020 Ambient air monitoring quality assurance guidance documents (available at: www.epa.gov/amtic/ambient-air-monitoring-quality-assurance-guidance-documents) (Accessed 21 September 2021)

US EPA 2021 Clean air plans; California; San Joaquin Valley moderate area plan and reclassification as serious nonattainment for the 2012 PM$_{2.5}$ (available at: www.federalregister.gov/documents/2021/09/01/2021-18764/clean-air-plans-california-san-joaquin-valley-moderate-area-plan-and-reclassification-as-serious) (Accessed 15 November 2021)

Waeytens J and Sadr S 2018 Computer-aided placement of air quality sensors using adjoint framework and sensor features to localize indoor source emission Build. Environ. 144 184–93

Wei L, Pu Z and Wang S 2013 Numerical simulation of the life cycle of a persistent wintertime inversion over Salt Lake City Bound.-Layer Meteorol. 148 399–418

Wolyn P G and McKee T B 1989 Deep stable layers in the intermountain western United States Mon. Weather Rev. 117 461–72

Yildirim B, Chryssostomidis C and Karmiadakis G E 2009 Efficient sensor placement for ocean measurements using low-dimensional concepts Ocean Modell. 27 160–73