AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

RESEARCH ARTICLE
10.1029/2021EA001849

Special Section:
Applications of Machine Learning Algorithms in Modeling Atmospheric Aerosols, Clouds, and Radiation

Detecting and Evaluating Dust-Events in North China With Ground Air Quality Data

Pei Feng Tong[1], Song Xi Chen[1,2,3], and Cheng Yong Tang[4]

[1]Guanghua School of Management, Peking University, Beijing, China, [2]School of Mathematical Science, Peking University, Beijing, China, [3]Center for Statistical Science, Peking University, Beijing, China, [4]Department of Statistical Science, Temple University, Philadelphia, PA, USA

Key Points:

- Develop a dust event detection procedure based on air quality data with finer spatial-temporal resolution and better weather adaptation
- Construct dust transport networks and find altering dependence between PM_{10} and CO, SO_2 and NO_2 at the start of dust events
- Dust-events contributed to 23.3%–34.6% for PM_{10} and 18.2–33.2% for $PM_{2.5}$ in the source regions and 2.0%–7.3% and 0.8%–4.0%, respectively, in the down-stream provinces in spring in North China

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. X. Chen,
songxichen@pku.edu.cn

Citation:

Tong, P. F., Chen, S. X., & Tang, C. Y. (2022). Detecting and evaluating dust-events in North China with ground air quality data. *Earth and Space Science*, *9*, e2021EA001849. https://doi.org/10.1029/2021EA001849

Received 21 JUL 2021
Accepted 10 NOV 2021

Author Contributions:

Conceptualization: Song Xi Chen, Cheng Yong Tang

Abstract We propose a dust-event detection and tracking procedure based on air quality data from the ground monitoring network by detecting temporal and spatial change-points in PM_{10} concentration. It supplements the existing remote sensing based approach with high temporal resolution and better weather adaptivity. Applications of the procedure on the labeled data showed its having high discriminating power for dust events, pollution events, and clean periods. Our study finds changing correlation patterns between PM_{10} and other air pollutants at the start of the dust events, which are utilized to enhance the discriminating power of the dust-event detection procedure. The detection and tracking procedure allows the construction of transport networks of the dust-events as well as the identification of the source regions and the transportation pattern, and assess the intensity and severity of the dust-events in North China. Our analysis find the dust-events contributed to 23.3%–34.6% for PM_{10} and 18.2–33.2% for $PM_{2.5}$ in the source regions and 2.0%–7.3% and 0.8%–4.0%, respectively, in the down-stream provinces in the spring season from 2015 to 2020.

Plain Language Summary Dust storm can be tracked via the air quality monitoring network via a statistical semi-supervised learning algorithm.

1. Introduction

Mineral dust is a major air pollutant in North China with a great influence on human health and geophysical-geochemical circulation. Dust aerosol is a major components of airborne aerosols. Dust storms or events can affect radiation balance directly through scattering and absorbing solar radiation (Z. Han et al., 2012; Sokolik & Toon, 1996), and indirectly by affecting cloud formation process and cloud properties (Huang et al., 2014; Rosenfeld et al., 2001). Dust aerosol exposure has adverse health effects (Peters et al., 2001). Heavy metals, soil-dwelling fungus, and other organic matters lodged on the dust aerosols can cause heritable DNA mutations, allergic reactions, and infections (Bener et al., 1996; Somers et al., 2002; Sprigg et al., 2014). Hundreds of millions of people living in North China are adversely affected by the Asian dust storms, which is the most prevalent in the spring.

In meteorology, dust events are usually defined in terms of visibility and wind speed. The China Meteorological Administration (CMA) has classified dust events into four categories as floating dust, blowing dust, dust storm and severe dust storm (CMA, 1979; Yang et al., 2008), which correspond to horizontal visibility of fewer than 10 km, 1–10 km with blowing wind, 500 m–1 km and fewer than 500 m, respectively. Remote sensing techniques have been the most popular method to determine dust related Aerosol Optical Depth (AOD) and near surface particulate matter (PM) concentrations based on satellite images (Sowden et al., 2018; Tian et al., 2020; van Donkelaar et al., 2015; You et al., 2016). Radiation and the spectra absorption differences between dust and other substances are the main basis for distinguishing dust events and other processes. There is an array of dust indices based on the spectrum of the MODIS products such as the thermal-infrared dust index (TDI), the normalized difference dust index (NDDI), and the enhanced dust index (EDI) (L. Han et al., 2013). Among these, there is a particular concern toward those dust events happened in East Asia. A recent investigation by Guo et al. (2019) found a trend reversal of dust AOD in 2009 using reanalysis data; Nobakht et al. (2021) built a dust emission source inventory using MODIS imagery data; Zhou et al. (2019) reanalyzed those dust sources in Central/East Asia and evaluated their impacts on dust emissions and concentrations by a simulation study. The remote sensing approach has the advantages of large scale spatiotemporal coverage, and can be integrated with the Chemical Transport Models (CTM) to improve simulation accuracy. For example, Wong et al. (2015) used a neural network

Data curation: Pei Feng Tong
Formal analysis: Pei Feng Tong
Funding acquisition: Song Xi Chen
Methodology: Pei Feng Tong, Song Xi Chen, Cheng Yong Tang
Software: Pei Feng Tong
Supervision: Song Xi Chen
Validation: Pei Feng Tong
Visualization: Pei Feng Tong
Writing – original draft: Pei Feng Tong, Song Xi Chen, Cheng Yong Tang

retrieval model for dust storm detection in East Asia with integrated remote sensing and data generated by CTM. However, the remote sensing approach has limitations for dust-events detection as high temporal and spatial resolution cannot be achieved simultaneously due to the satellite orbital constraints. The data availability is also affected by the cloud and other weather conditions. Ma et al. (2016) and Chen et al. (2019) reported 80%–90% missing rates of the MODIS AOD products in China related studies, which basically cause the approach not applicable for general trend analysis on the dust events. Another issue with the remote sensing approach is that the missingness of the AOD may not be completely random and may be subject to selection bias, which implies that simply ignoring the missing values may result in biased analysis. In contrast, the ground air quality monitoring network data may offer a complementary and more direct approach for detecting and tracking dust-events with a high temporal resolution, which can be used to assess the severity and trend of dust events in a large scale region equipped with the air quality monitoring sites like North China. To the best of our knowledge, this is the first study for detecting and tracking dust events using ground air quality monitoring network data.

China has established a nation-wide air quality monitoring network since 2013, which now covers all prefecture-level cities with nearly 1,500 centrally administrated sites, which is supplemented by the provincially administrated monitoring sites. Each monitoring site records hourly concentrations of six air pollutants: PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3. The PM_{10} (aerodynamic diameter less than 10 μm) is a major tracer of dust-events as it constitutes the main content of dust episodes and can be transported thousands kilometers from the source region. Previous studies showed that high PM_{10} is highly positively correlated with dust events (Krasnov et al., 2016; Querol et al., 2009). Although PM_{10} is a good indication of dust events, it is also highly associated with severe air pollution episodes, especially when our study region encompasses not only the arid regions but also the highly polluted industrial regions in North China. In this study, ground-based air quality data and the associated hourly wind field data are used to detect North China's dust events and track their movement. The proposed method is based on a combination of change-point detection in PM_{10} and dynamic clustering algorithms which cannot only detect the start of the dust events but also track their trajectories over the study region. The detection and tracking aspects of the algorithms provides better discrimination between dust and pollution events since those events originated in urban or non-arid areas can be ruled out based on the prior knowledge on the dust source regions. The correlation structures between PM_{10} and other air pollutants are utilized in the algorithm in addition to the wind profile information to enhance the detection accuracy. Our goal is to establish a dynamic event detection method using spatial-temporal observation data for identifying and tracking dust storms' movement over time.

Spatial-temporal event detection is a topic that has been studied extensively in statistics and machine learning, where the focus has been in finding target events with unlabeled or limited labeled spatial-temporal data. See Yu et al. (2020) and Unkel et al. (2012) for summaries. These methods have a wide variety of applications, for example, in finding infectious disease outbreaks, locating deforestation regions and spatial pollution segmentation.

There are three major tasks in dust-event detection: (a) detecting multiple clusters over a large geographical region, (b) simultaneously tracking each event, and (c) efficient computation for practical feasibility. These tasks arise from the specific features of the problem. First, due to the interference of the meteorological conditions, the shapes of the dust events are highly dynamic such that any predefined shapes (elliptical or arc-shaped) of the affected area may not be appropriate. This means that the shapes of the two dust affected areas may differ at two time points, even they may belong to the same event. This also implies in-applicability of the existing methods which search with predefined shapes of the affected area. Second, two dust events may occur at the same time in a study region, bringing in the task to identify and track simultaneously different events. Third, the conventional scanning approaches (Duczmal & Assunção, 2004; Neill et al., 2005), which attempted to exhaustive search all possible sub-areas for event detection, would quickly become a daunting task computationally in our setting. In our moderate size data sets, there are 688 monitoring sites at each time step and a total of 2,208 hourly records at each monitoring site each spring. The all directional scanning with unconstrained shapes and a time varying window has a computing complexity at the order of $2^{688 \times 2,208}$.

In this study, we propose a novel two-stage dust-event detection and tracking procedure which provides tangible solutions to the three aforementioned major tasks based on ground monitoring data at hourly frequency. The stage one algorithm searches the site-wise data for multivariate change point detection to locate potential target events and their spatial and temporal locations. In stage two, the spatial-temporal dependence of dust events is utilized with a bottom-up majority voting component to aggregate abnormal areas and locate the event clusters. In particular, it not only maps the affected regions of the dust-events, but also their dynamic properties to

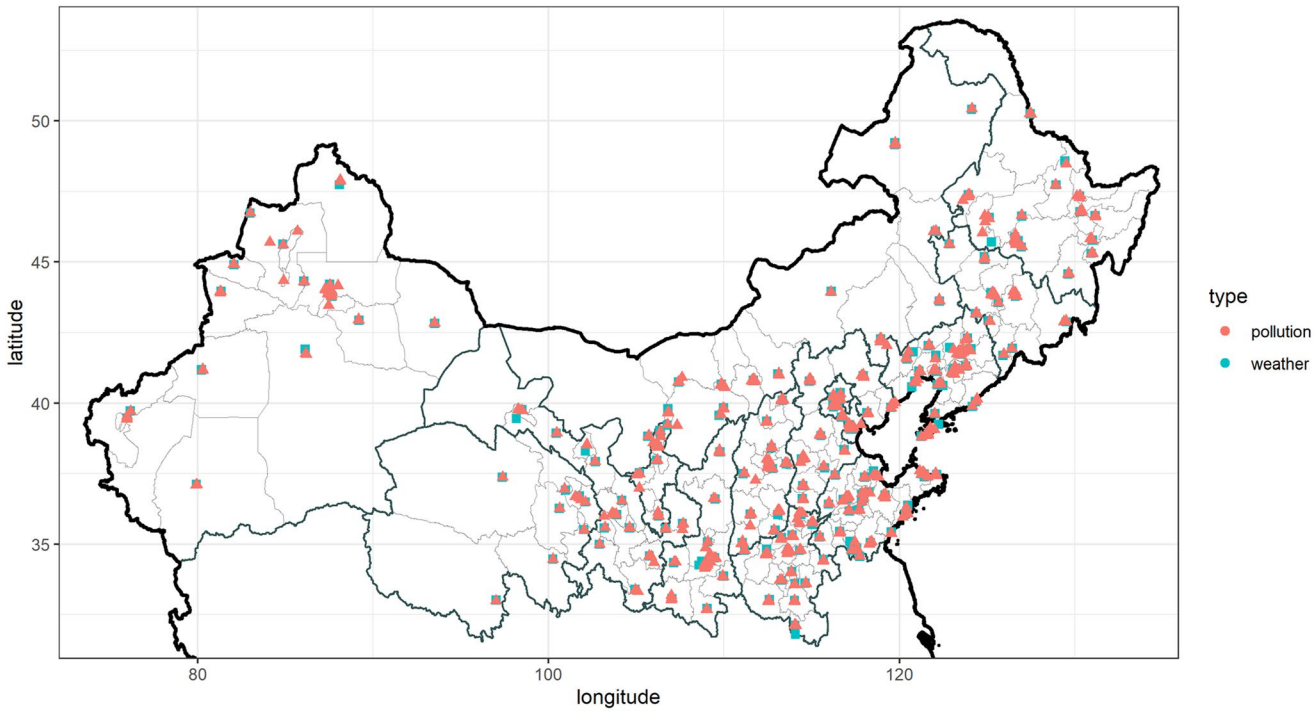

Figure 1. Study region with the locations of the 688 air quality monitoring sites (red) and the 258 matched CMA weather observing stations (blue) in North China.

identify trajectory and source regions. The proposed spatial-temporal event detection method can identify arbitrary shaped and varying sized dust clusters without a preknowledge on the number of clusters. The method is operational with limited label data for tuning parameter selection, which can track the movement of a dynamic event automatically and is able to re-create the transportation networks of the dust events.

By analyzing the detected episodes of dust events and their dynamic transport networks, we assess the frequencies, within-episode severity, and the size of the affected area of the dust-events in North China in the springtime from 2015 to 2020. It was found that there was no significant changes in the occurrence frequencies, within-event average PM_{10} concentration and the average affected area per dust event in the six springs from 2015 to 2020. The contribution of the dust events to the average spring season PM_{10} and $PM_{2.5}$ were enumerated which showed a clear regional heterogeneity. The dust-events contributed to 59.2% of ground PM_{10} concentration in Xinjiang and 17.0% near Alashan desert, while the effects were typically less than 5% in the down stream provinces. Our procedure detected three major dust source regions in North China. The constructed transportation networks recognized that dust events originated in Xinjiang were less likely to travel out of the province. The detected events also reveal changing correlation pattern between PM_{10} and three air pollutants SO_2, NO_2, and CO while maintaining a steady correlation with $PM_{2.5}$ at the start of the events.

The rest of the paper is organized as follows. Section 2 describes the air quality and meteorological data used in the study. The two-stage detection and tracking procedure is outlined in Section 3, followed by tuning parameter selection and empirical validation of the procedure with classified data. Section 4 assess the dust situations in North China over the six springs from 2015 to 2020, which includes the intensity, severity and the affected areas as well as dynamic transport of the dust event. The correlation patterns between PM_{10} and the other air pollutants are investigated in Section 5. Section 6 evaluates the contribution of the dust events to PM_{10} and $PM_{2.5}$. Extra details on the methods and results are reported in Supporting Information S1.

2. Study Region and Data

The study region is the 15 provinces located entirely above 32° latitudes north in China, which encompasses the arid and semi-arid regions such as Xinjiang, Gansu, Ningxia, Qinghai, Shaanxi, Shanxi provinces and Inner Mongolia, and additional nine provinces in North China as shown in Figure 1. For better interpretation, we divide

Inner Mongolia into east and west parts by the 111° longitudes east. Our study concentrates on the spring season from March to May when the dust is the most prevalent.

China has established a national air quality monitoring network since 2013. Considering the data quality and availability, air quality data before 2015 are not used due to high missing rates and the fact that many monitoring sites have not yet been established near the sand storm prone area. These led us to choose the spring season from 2015 to 2020 as the study periods. The ground air quality monitoring data contained hourly records of six pollutants (PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3) from 688 sites which are directly administrated by the Ministry of Ecology and Environment.

Hourly meteorological data were obtained from China's Central Meteorological Administration (CMA), containing air temperature and pressure, the relative humidity, the dew point temperature, wind speed, and direction observed at two meters above the surface. The wind directions are grouped into 16 categories (N, NNE, NE, ENE, etc.).

We also employ the assimilated data in ERA5 of Global Reanalysis from the European Center for Medium-Range Weather Forecast (ECMWF) (Hersbach et al., 2020) over the study region, which is hourly grid data with a 0.25° by 0.25° spatial resolution. U and V-winds at 500 hPa pressure level were used to complement the air quality and CMA data. The meteorological variables from both CMA and ECMWF were matched to the nearest ground air quality sites according to the geographical distance. The portion of missing data of the air quality data was the highest at 4.8% for SO_2, while the highest missing rates of CMA data were 6.1% for wind directions followed by 2.3% for wind speed and lower than 2% for other variables. Linear interpolation was applied to impute the missing values. The spatial locations of the pollution monitoring sites and weather stations are shown in Figure 1.

To validate the proposed approach, we manually identified and labeled (classified) 12 dust events as reported by the CMA and confirmed via the media reports, along with four pollution events and two clean periods to service as contrasts. The time intervals of the 12 labeled dust events and four pollution events are listed in Table S1 of Supporting Information S1. The labeled periods constituted 23.3% of the total time range of the study, with 4.2% of the labeled records as dust, 1.5% as pollution and the other 94.2% as clean periods.

3. Dust Event Detection and Tracking Method

Spatial-temporal clustering methods have been extensively investigated in statistics and machine learning literature. A commonly applied approach is by exhaustive searching and processing all potential clusters defined by spatial-temporal closeness; such methods are computationally prohibitive for practical applications of moderate or large scale; see, for example (Duczmal & Assunção, 2004; Lin, 2014).

We are to develop a two-stage computationally efficient method, delicately incorporating the features associated with the dust events. The first stage detects multiple site-wise change-points associated with dust in the time series measurements of the air pollutants. The second stage performs a spatial-temporal clustering to build the complete trajectories of the dust events. Compared with the exhaustive searching methods for spatial-temporal clustering, our approach only requires sequential examinations by going through the entire spatial-temporal data set only once, enabling its appealing applications in solving large-scale problems.

Our detection algorithm first focus on PM_{10}, a proven species strongly associated with dust. Later we include the correlation between PM_{10} and other air pollutants as our analysis finds that correlation with some pollutants carries dust signals.

Let $S = \{s_i\}_{i=1}^n$ denote the study region containing n air quality monitoring sites. We are to find a collection of spatial-temporal dust clusters $\{C_1, C_2, \ldots, C_M\}$. Here each cluster C_k is associated with the occurrence of an event over a time interval $T_k = [B_k, E_k]$ with the beginning-time B_k and ending-time E_k. For a $t \in T_k$, let $R_t^k \subset S$ be the collection of the dust affected sites at time t. Then, the kth dust event is fully represented by a time series of spatial sets $C_k = \left\{R_t^k,\ \ t \in T_k\right\}$. The proposed two-stage procedure is to identify and track each spatial-temporal set C_k over space and time.

3.1. Stage One—Detecting Dust Episodes at a Site

The goal of the stage one procedure is detecting the site-wise change-points associated with the sudden changes in the measurements of the pollutants. A drastic increment in the PM_{10} is a prominent indication of a dust storm. For each site s_i, let $x(s_i, t) = (x_1(s_i, t), \ldots, x_6(s_i, t))' \in \mathbb{R}^6$ be the concentrations of the six pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3) at time $t \in \{1, \ldots, T\}$. As a practice for handling positive-valued data, we perform a log-transformation and normalize them as $y_{it} = (y_{it,1}, \ldots, y_{it,6})'$ with

$$y_{it,k} = \frac{\log x_k(s_i, t) - \hat{\mu}_{i,k}}{\hat{\sigma}_{i,k}}, k = 1, 2, \ldots, 6,$$

where $\hat{\mu}_{i,k} = T^{-1} \sum_{t=1}^{T} \log x_k(s_i, t)$, $\hat{\sigma}^2_{i,k} = (T-1)^{-1} \sum_{t=1}^{T} (\log x_k(s_i, t) - \hat{\mu}_{i,k})^2$ are the sample mean and variance of the kth pollutant at site s_i, respectively.

We want to locate a sequence of change-points $\tau^{(i)} = \left\{\tau_1^{(i)}, \ldots, \tau_{N_i}^{(i)}\right\}$ at each site i ($i = 1, \ldots, n$). We define a change-point $\tau_j^{(i)}$ as a time point such that the distributions of PM_{10} and other pollutants differ on the two adjacent intervals $t \in [\tau_{j-1}^{(i)}, \tau_j^{(i)})$ and $t \in [\tau_j^{(i)}, \tau_{j+1}^{(i)})$, $(j = 2, \ldots, N_i - 1)$.

To allow use of both the PM_{10} data and its correlation with other pollutants, we define $z_{it} = (y_{it,1}, Q_{it})'$ where

$$\hat{Q}_{it} = (2h)^{-1} \sum_{j=t-h}^{t+h} (y_{ij,1} - \bar{y}_{it,1})(y_{ij,(1,3,4,5)} - \bar{y}_{it,(1,3,4,5)})',$$

$\bar{y}_{it,1} = (2h+1)^{-1} \sum_{j=t-h}^{t+h} y_{ij,1}$, and $\bar{y}_{it,(1,3,4,5)} = (2h+1)^{-1} \sum_{j=t-h}^{t+h} y_{ij,(1,3,4,5)}$ are the moving window averages of PM_{10} and a vector contains four pollutants (PM_{10}, NO_2, SO_2 and CO), respectively, and h is the window size.

For each t in a generic time interval $[a, b]$, we construct the Cumulative Sum (CUSUM) statistics (Page, 1954):

$$S_{it}(a, b) = \frac{1}{\sqrt{b-a+1}} \left(\sum_{k=a}^{t} z_{ik} - \frac{t-a}{b-a+1} \sum_{k=a}^{b} z_{ik} \right) \quad \text{and} \tag{1}$$

$$\Omega_i(a, b) = \frac{1}{b-a+1} \sum_{t=a}^{b} S_{it}(a, b)^T W_i(a, b)^{-1} S_{it}(a, b), \tag{2}$$

where W_i is a known or estimated invertible semi-positive definite and symmetric weighting matrix. In this study, we choose it as a diagonal matrix by setting the off-diagonal elements in estimated long-run covariance matrix of $S_{it}(a, b)$ as 0. The estimated change-point over $[a, b]$ is

$$\hat{\tau}(a, b) = \arg\max_{a \le t \le b} S_{it}(a, b)^T W_i^{-1}(a, b) S_{it}(a, b). \tag{3}$$

Naturally, there are multiple occurrences of dust events at the sites over the study periods. Thus, a major task in the first stage is site-wise multiple change-points detection, which is achieved by applying Equation 3 to a collection of sub-intervals in $[0, T]$. See Algorithm 1 in Text S3 of Supporting Information S1 for more details and a pseudo code for the multiple change-points detection procedure.

The multiple change-points detection creates site-wise change-point segments $\left\{\tau^{(i)}\right\}_{i=1}^{n}$ where $\hat{\tau}^{(i)} = \left\{\tau_1^{(i)}, \ldots, \tau_{N_i}^{(i)}\right\}$ collects within episode time points at site s_i. Before performing spatial-temporal clustering, our algorithm evaluates the mean PM_{10} level and determines whether or not a pollution event is a case during $[\tau_j^{(i)}, \tau_{j+1}^{(i)}]$. For $t \in [0, T]$, we define a piece-wise mean-value function at a site s_i

$$\hat{f}_{s_i}(t) = \sum_{j=0}^{N_i} \bar{x}_{1,j}(s_i) I(t \in [\tau_j^{(i)}, \tau_{j+1}^{(i)}])$$

where $\bar{x}_{1,j}(s_i)$ is the mean PM_{10} value at site s_i over the interval $[\tau_j^{(i)}, \tau_{j+1}^{(i)}]$, and we take $\tau_0^{(i)} = 0$ and $\tau_{N_i+1}^{(i)} = T$. By thresholding $\hat{f}_{s_i}(t)$ at a level ζ, we attain a dust event indicator function

$$V(s_i, t) = I(\hat{f}_{s_i}(t) \ge \zeta)$$

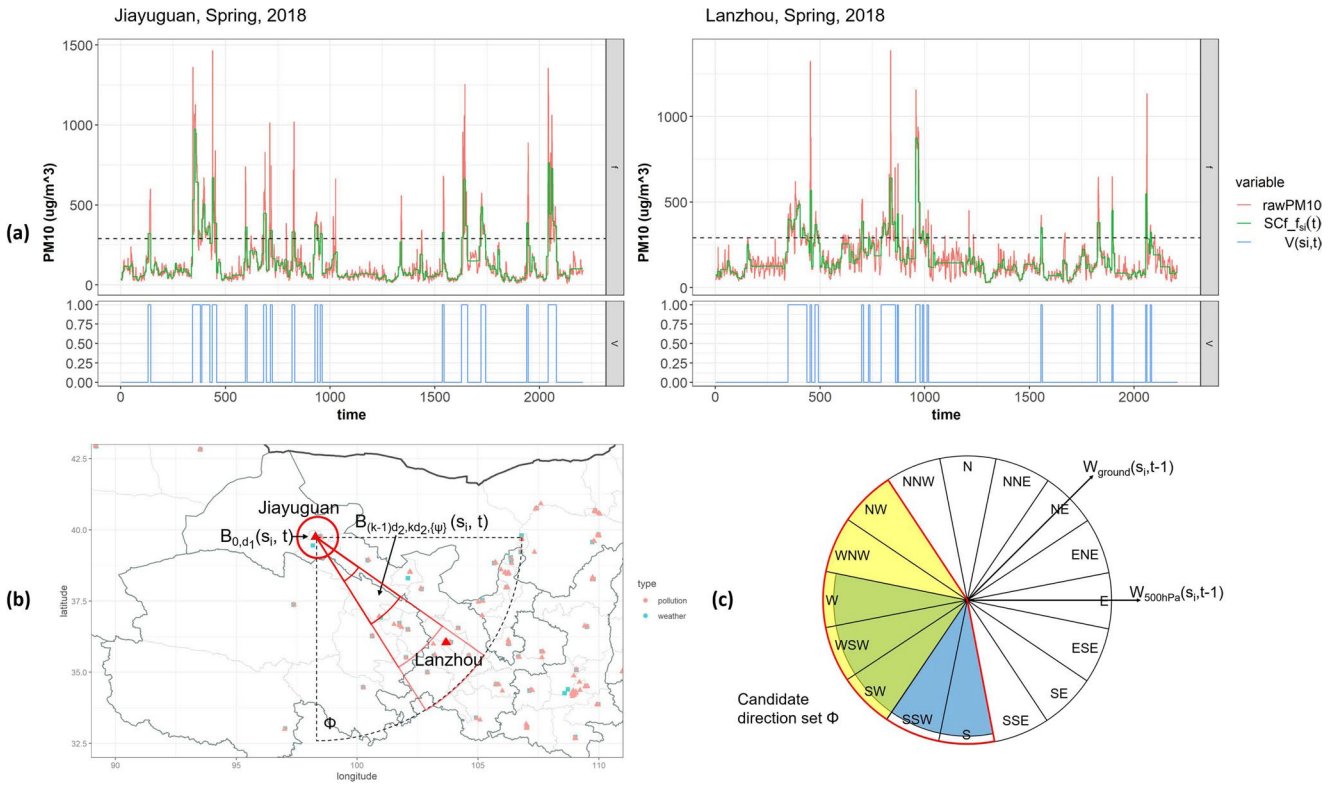

Figure 2. (a) Raw PM_{10} time series (red), the associated segmented concentration functions $\hat{f}_{s_i}(t)$ (green) and the dust-event indicator $V(s_i, t)$ (blue in lower panels) for spring 2018 at sites G620201 (Jiayuguan) and G620102 (Lanzhou) both in Gansu province. The dashed horizontal lines in (a) mark the threshold level $\zeta = 290$ μg/m^3 determined from the training data. (b) A demonstration for the bottom-up searching areas in Steps 1 and 2 of Stage Two algorithm with the candidate searching direction set Φ is between the dashed black lines, and the search sets $\left\{B_{(k-1)d_2,kd_2,\{\phi\}}(s_i,t)\right\}_{k=1}^{K}$. (c) Directional set Φ (in color) when the ground wind direction $W_{\text{ground}}(s_i, t-1) = SW$ and the 500 hPa pressure level wind direction $W_{500\,\text{hPa}}(s_i, t-1) = W$, where Φ is the union of the wind directional sets at the ground (blue) and at 500 hpa (yellow).

where $I(A)$ is the indicator function, which is 1 if condition A is met and 0 otherwise. That is, if the mean PM_{10} level over a detected interval exceeds ζ, such a period is flagged as a dust event. Figure 2a provides plots of exemplary $\hat{f}_{s_i}(t)$ and $V(s_i, t)$. Here, ζ as a tuning parameter will be determined by the Receiver Operating Characteristic Curve (ROC) calculated from the labeled (classified) data, whose details are given in Text S1 of Supporting Information S1. Data-oriented tuning enriches our spatial-temporal clustering approach as a semi-supervised one—an added feature of the proposed method.

3.2. Stage Two: Dynamic Clustering of Dust Affected Sites

The identified site-wise process $V(s_i, t)$ reflects dust situation over time at site s_i. Clearly, the site-wise detection helps detect and study dust events as spatial-temporal clusters of pollution. Nevertheless, not all site-wise events are due to a dust storm, as some may be idiosyncratic local pollution events. The motivations of the stage two algorithms are thus two-fold. First, it aims at capturing the complex dynamic behaviors of dust events, with no constraint or assumption on their shape, number or size, while maintaining computational efficiency. The arbitrary shape is benefited by merging small pieces of affected areas together from bottom to up; second, the algorithm is designed for achieving adequate robustness capable of differentiating the dust from the local pollution events, resulting in few false alarms.

The stage two algorithm requires a single scanning over all available time points in $[0, T]$. During the scanning, once a detected event is encountered at a site, a spatial-temporal candidate set is constructed by including sites nearby for evaluation and confirmation. When forming the spatial-temporal candidate set, we incorporate the site-wise wind directions which are denoted by $W(s_i, t)$ taking 16 possible values, as shown in Figure 2c. Then a majority vote rule is developed to decide whether or not such a set belongs to a dust storm event.

Efficiently generating candidate event sets is a key step that considers the site profile and wind information. To introduce our solution, let $g(s', s)$ denote the relative position of s' to s, for instance g (Beijing, Shanghai) = NW and g (Shanghai, Beijing) = SE. Let $d(s', s)$ be the Euclidean distance between sites s' and s. Then, the Φ-directional r_1, r_2-spatial-temporal neighbor of a site s at time t is

$$B_{r_1,r_2,\Phi}(s,t) = \left\{(s',t') : r_1 \le d(s',s) \le r_2 \text{ and } g(s',s) \in \Phi \text{ and } t' \in \{t-1, t\}\right\}. \quad (4)$$

Here $B_{r_1,r_2,\Phi}(s,t)$ contains the neighboring sites of s at times t and $t-1$ within a distance range $[r_1, r_2]$ in the wind direction Φ, which provides a device to refine the nearby sites according to the relative position of the sites while incorporating wind direction. A special case is when Φ contains all 16 wind directions, and we denote it by $B_{r_1,r_2}(s,t)$, which encompasses all the $[r_1, r_2]$ spatial neighbors of (s, t).

The direction set Φ is important in defining the candidate search set in Equation 4. Existing studies in Northwest China show evidence that the long-range transport of dust aerosols may be dominated by the upper level wind (Huang et al., 2008). Indeed, the impacts of ground level wind are more concentrated in the dust source regions for lifting the dust aerosols. Therefore, both the wind directions at the 500 hPa pressure level and near surface are used to decide the candidate Φ. As shown in Figure 2c, Φ is the union of the two sets of wind directions.

We next present a majority voting rule with a threshold level $\lambda \in (0, 1)$ for forming event clusters over a search region R when most of the sites within it are experiencing the event.

Definition 1 For a spatial-temporal candidate set R containing $|R|$ number of sites, the λ-majority voting rule is

$$\mathrm{CL}(R, \lambda) = I\left\{\frac{\sum_{(s,t)\in R} V(s,t)}{|R|} > \lambda\right\}.$$

As an example, consider $R = B_{0,r}(s_i, t)$ corresponding to a spatial-temporal region near site s_i at a time t. Then whether or not a clustered event occurs near s_i can be assessed by examining the value of $\mathrm{CL}\{B_{0,r}(s_i, t)\}$, if more than λ percentage of the sites are having the event, we declare the region $B_{0,r}(s_i, t)$ is (part of) a cluster of a detected event. If we confine the search direction to *NE* by assigning $\Phi = \{\mathrm{NE}\}$, then $\mathrm{CL}\left\{B_{r_1,r_2,\Phi}(s_i,t)\right\}$ is used to reflect the popular votes.

The stage two algorithm consists of three steps:

1. (Locating) All s_i and t ($i = 1, \ldots, n$; $t = 1, \ldots, T$) are scanned once. For each $V(s_i, t) = 1$ located, $B_{0,d_1}(s_i, t)$ with $d_1 = 20$ KMs is constructed. Then the λ-majority voting rule is applied to determine whether or not a cluster of events is the case ($\mathrm{CL}\left\{B_{0,d_1}(s_i,t)\right\} = 1$ or 0).
2. (Determining Spatial Coverage) For each $B_{0,d_1}(s_i, t)$ that has passed the majority vote in step 1, a set Φ is created based on wind directions $W(s_i, t-1)$ at the ground level and 500 hpa as in Figure 2c. Facilitated by Φ, a directional step-wise process by including more sites and repeating the λ-majority voting is conducted for extensively determining the spatial-temporal coverage of this clustered events; see Figure 2b for an illustration, and Figure S2 in Supporting Information S1 for more details.
3. (Merging) When there are (s, t) belonging to multiple clustered events, we merge those clusters as one. This results in the final output of the event clusters $\{C_1, \ldots, C_M\}$; see Figure 3 for a demonstration.

The pseudo-code of the stage two algorithm is presented in Text S3 of Supporting Information S1.

3.3. Semi-Supervised Tuning Parameter Selection and Validation

The two-stage procedure contains tuning parameters: Ω^*—the threshold of the CUSUM statistics, ζ—the threshold in deciding dust episodes at each site, and λ in the majority vote rule. Other hyperparameters include the moving window width h used to calculate the moving covariance in Stage one, the searching radius d_1 and d_2, and the maximum recursive searching depth K, in Stage two. The methods used for selecting these tuning parameters are detailed in Text S1 of Supporting Information S1.

It is noted that the labels of the dust and other pollution events are not part of the original ground air quality monitoring data as they contain pollution information only. The procedure used to obtain the training data is by a manual validation process that identifies and labels the spatial-temporal events using other source of information

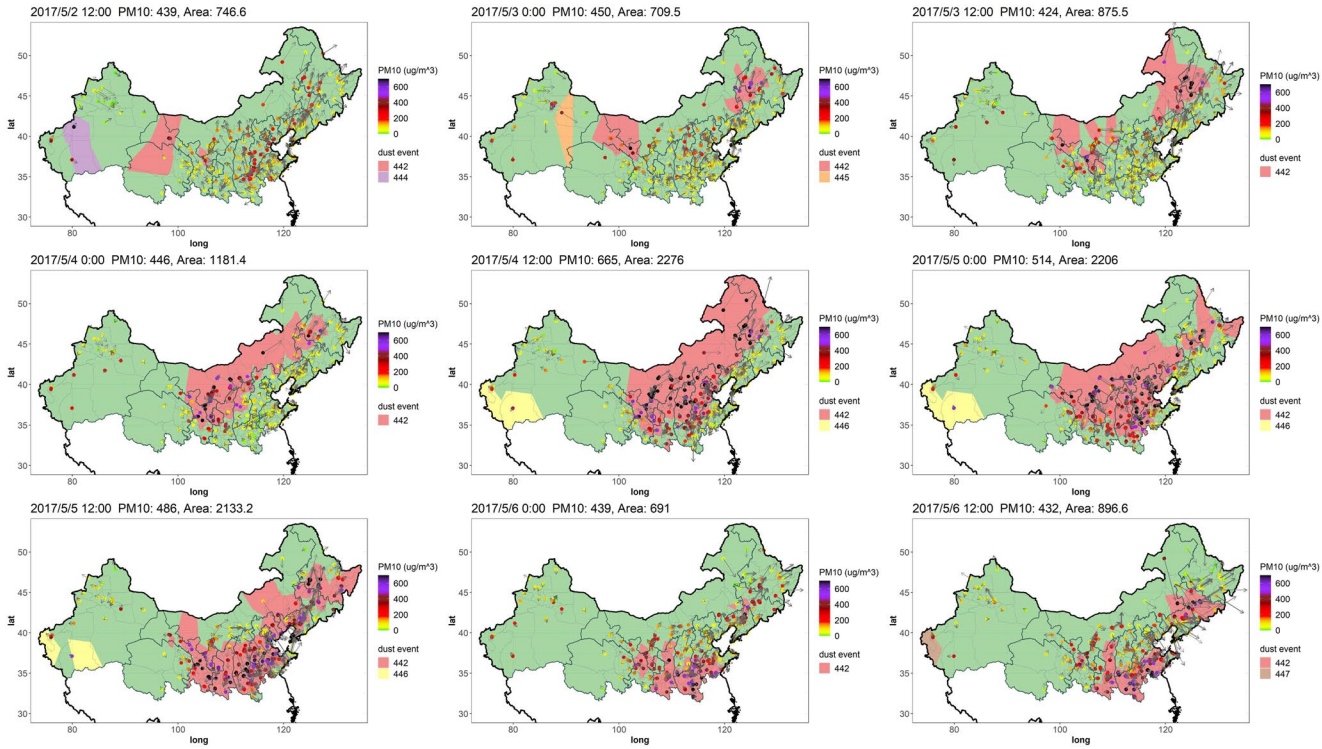

Figure 3. Nine snapshots (every 12 hr) on the detected dust-event areas (yellow) for two dust storms from May 2 to 6, 2017. The color points correspond to air quality sites with color indicating PM_{10} concentration and arrow denoting the wind direction and speed. The dust affected regions are drawn based on the Voronoi diagrams constructed using the detected sites from the two stage procedure. Mean PM_{10} concentration ($\mu g/m^3$) and affected area (10^3 KM^2) in dust events region are also reported for each snapshot.

such as the CMA and the media reports. The labeled training data were used for back-testing of the detection procedure for the dust events.

In the labeled data, two dust processes happened in North China from May 2 to 7, 2017, which gave us an opportunity to evaluate the proposed method, and provides understanding to the structure of the dust events. Figure 3 gives nine Snapshots of the dust storm taken every 12 hr with site-wise PM_{10} concentration and ground wind vectors. For ease of presentation, we converted the point observation data to a colored polygon in Figure 3 by the Voronoi diagram, where different colors indicate different dust events. The affected area at each time step was automatically selected by our algorithm. Dust storm A was first detected in Gansu and Qinghai provinces. Twelve hours later Dust B emerged from Horqin sandy land in Northeast China, which stayed by itself for more than 24 hr before merging with Dust A before the 60th hr. Dust A gained momentum between the 36th and 48th hr when it became a major event over Gansu, Ningxia and West Inner-mongol area. Driven by a strong north-westerly wind, the dust storm traveled east and south to reach Shanxi, Beijing-Tianjin-Hebei (BTH) and East Inner-Mongol region and merged with a small dust event in Heilongjiang province at 00:00 a.m. on 4 May. After spent around 48 hr in the densely populated north-central China the dust storm dispersed by the northwest wind and entered Eastern China with decreased strength.

Excessive PM_{10} were observed in the monitor network during the 96 hr, where the maximum PM_{10} reached 1,721 $\mu g/m^3$ at 14:00 on 4 May in Xilingole in East Inner-Mongolia. The dust cluster moved fast in the study region with varying shape and the affected area. The frequent merging and splitting of dust storms were affected by the wind and terrain. It is obvious that we can hardly use the existing spatial-temporal abnormality detection methods (Lee et al., 2020; Neill et al., 2005) for the task of dust event detection since they make too strong assumptions in both shape and dynamics.

To verify the effectiveness of the proposed dust event detection procedure, we applied the two-stage spatial-temporal clustering algorithm on the labeled data consisting of dust events as well as data from the usual polluted and clean periods. The labels (classifications) allow us to access the true positive and false negative rates of the

Table 1
Cross-Tabulation of the Correct and Incorrect Classification Rates on the Hourly States (Dust Storm or Otherwise) at the Monitoring Sites as Verified via the Labeled Data Based on the Five-Fold Cross-Validation Using Only PM_{10}, and Adding PM_{10}'s Covariances With the Four Pollutants in the Dust-Event Detection Algorithm

Setting			Statue		
			Dust	Clean	Pollution
Out-of-sample predictions	PM_{10} only	Dust	85.0%	2.5%	9.3%
		No dust	15.0%	97.5%	90.7%
	PM_{10} with four covariances	Dust	95.7%	1.2%	2.9%
		No dust	4.3%	98.8%	97.1%
	PM_{10} with four covariances (source restricted)	Dust	94.1%	1.0%	1.0%
		No dust	5.9%	99.0%	99.0%
	Total counts		89,661	2,000,258	32,561

Note. That "source restricted" refers to exclude the identified events originated from those known areas with more pollution events but dust events are rarely from. The total counts of the spatial-temporal observations corresponding to each category of the labels are reported at the last row.

proposed procedure, where a true positive means correct identification of a dust process while a false negative means failure to detect one.

The labeled data set we created provides a platform for evaluating the out-of-sample classification accuracy of our approach on the three statues of the data: dust, pollution and clean. The evaluation was made via a five-fold cross-validation (CV) procedure that randomly split the labeled data to five segments and alternatively leave out one segment as the validation set and the other four segments as the training set. We utilized the four-segment training data for the tuning parameter selection for the proposed two-stage algorithm first, and then applied the trained two-stage detection procedure to predict the statue of every datum of the leave-out segment. Such a practice was repeated five times so that each segment had been a validation set once.

Table 1 provides the classification accuracy using the aforementioned five-fold CV procedure based on the labeled data by employing only the PM_{10} and both PM_{10} and its covariance with the four pollutants (PM_{10}, NO_2, SO_2 and CO) in the dust event detection procedure. The performance is based on the true positive and false negative criteria. Compared with using only PM_{10} in the dust-event detection, adding the covariance information enhanced the true positive rate by 10.7%. More strikingly, it reduced the false positive rate by 6.4% for the pollution events and 1.3% for the clean processes, which mounted for more than 51% reduction in the overall false positive rates. This is achieved by the ability of better distinguishing the dust from the pollution episodes with the use of the covariance information. We note an additional fact that pollution happens more often near the region surrounding the major urban and industrial centers, and in contrast, it is known that dust events rarely originate from this area. If we exclude the identified dust events originated from this area, the false positive rate was further reduced from 2.9% to 1%, as reported in the third setting (source restricted) of Table 1. Meanwhile, this gain in the specificity is accompanied by a relatively small cost of the sensitivity to the dust process detection. It is noted that as the number of observations per class (dust or non-dust) was imbalanced, and both methods achieved low false positive rates as the number of no dust records was relatively large.

4. Features of Dust Events in North China

4.1. Overview

We evaluate the dust event situation in North China in the spring season from 2015 to 2020 by the proposed dust detection method to provide understanding on the source, intensity, duration, and severity as well as the transportation network of the dust events. Identifying the source region will help locating the areas for desertification control, while moving trajectory detection provides dust transportation pathways useful for building a dust storm early warning system. We show that with the detected dust episodes, these can be addressed numerically at different levels of spatial and geographical resolution.

Before laying out our findings, we define some quantities that are informative for characterizing the aforementioned features. These quantities are all calculated from the outcomes of our methods.

We define the source region, denoted by R^k_{source}, of the kth dust episode to be the affected area during the first 10% of its entire period of occurrence. This is motivated by the study of R. Zhang et al. (2005) and X.-X. Zhang et al. (2018), which found that dust events could move out of their source regions in the first 15% of the episode time. Specifically, recall that a detected dust event, denoted as C_k as defined in Step 3 of the stage two algorithm in Section 3.2, is described as $C_k = \left\{R^k_t, t \in T_k\right\}$ where $R^k_t \subset S$ represents the affected area of C_k at time t in the event time interval $T_k = [B_k, E_k]$ with the beginning (ending)-time B_k (E_k). Hence, R^k_{source} is the union of those R^k_t with t belonging to the first 10% of $[B_k, E_k]$.

For a location s_i, let

$$H(s_i) = \sum_{k=1}^{M} \sum_{t=B_k}^{E_k} I(s_i \in R^k_t) \text{ and } F(s_i) = \sum_{k=1}^{M} I(s_i \in R^k_{source}) \tag{5}$$

be the total number of hours that s_i appeared in all the detected dusts episodes, and the frequencies that s_i appeared in the source regions, respectively. Here, $H(s_i)$ is a s_i-specific dust duration measure, and $F(s_i)$ reflects how often dust events originate from an area containing s_i. For a region, for example, a province, denoted by $\mathcal{D}$, that contains multiple locations, let $\mathcal{K}_D = \left\{k \in \{1, \ldots, M\} : \mathcal{D} \cap R^k_{source} \neq \varnothing\right\}$ be the collection of dust episodes whose source contain at least a part of $\mathcal{D}$. Let

$$\begin{aligned} \mathrm{D}(\mathcal{D}) &= \frac{1}{|\mathcal{K}_D|} \sum_{k \in \mathcal{K}_D} |T_k|, \mathrm{P}(\mathcal{D}) = \frac{1}{|\mathcal{K}_D|} \sum_{k \in \mathcal{K}_D} \frac{1}{|T_k|} \sum_{t \in T_k} \frac{1}{|R^k_t|} \sum_{s_i \in R^k_t} x_1(s_i, t), \\ \mathrm{A}(\mathcal{D}) &= \frac{1}{|\mathcal{K}_D|} \sum_{k \in \mathcal{K}_D} \frac{1}{|T_k|} \sum_{t \in T_k} \mathrm{Area}(R^k_t), \end{aligned} \tag{6}$$

where $|M|$ is the cardinality of the set M, and Area(M) returns the area of M. Specifically, D($\mathcal{D}$) is a duration measure by the average length of the dust events originated from the region $\mathcal{D}$, P($\mathcal{D}$) is the averaged PM_{10} concentration measurements of the dust episodes, A($\mathcal{D}$) provides the average area that the dust events covered in this region.

4.2. Source Region, Frequency, and Duration

We applied proposed dust event detection algorithm to find the dust events over the study region in the springs from 2015 to 2020, namely obtaining $C_k = \left\{R^k_t, t \in T_k\right\}$ which contained the spatial and temporal coverage of the dust events. Based on the information in C_k, those summary statistics given in Equations 5 and 6 can be calculated to gain information on the source region, frequency and duration of the dust events in North China over the 6 years.

The Outer Mongolia is a major source region for dust events affecting East Asia. However, as the ground monitoring data in the Outer Mongolia are not readily accessible or available, it may well be that some of the sand events sourced to Inner Mongolia (for instance Alashan desert) were originated in the Outer Mongolia. It should be noted that what we had detected were effectively the early affected areas.

Figures 4a and 4b display the geographical distribution of the source region in the site-wise scale with the frequencies and the cumulative affected hours of the dust episodes originated in the sites over the six springs from 2015 to 2020, calculated according to Equation 5, while Figures 4c and 4d describe the transportation network from the source to down-stream provinces and counts of out of the source province episodes versus those remained within the source provinces, respectively. The figure shows that Taklamakan and Alashan (consists of Gansu, Western Mongolia and Ningxia, part of Gobi desert in China side) deserts, and Horqin sandy land were the three main source regions in China. The most frequent source region was Taklamakan desert occupying 42.1% of the total dust events over the six springs, followed by 23.4% for Alashan and 16.6% for Horqin, obtained by aggregating the dust frequencies in Figure S4 of Supporting Information S1. The air quality sites surrounding Taklamakan desert observed excessive and persistent PM_{10} levels during the detected episodes. For instance, Hotan city located in the southwest edge of Taklamakan experienced 61 events with 6,945 cumulative dust hours over the six springs, mounting to an average of 48.2 days each spring.

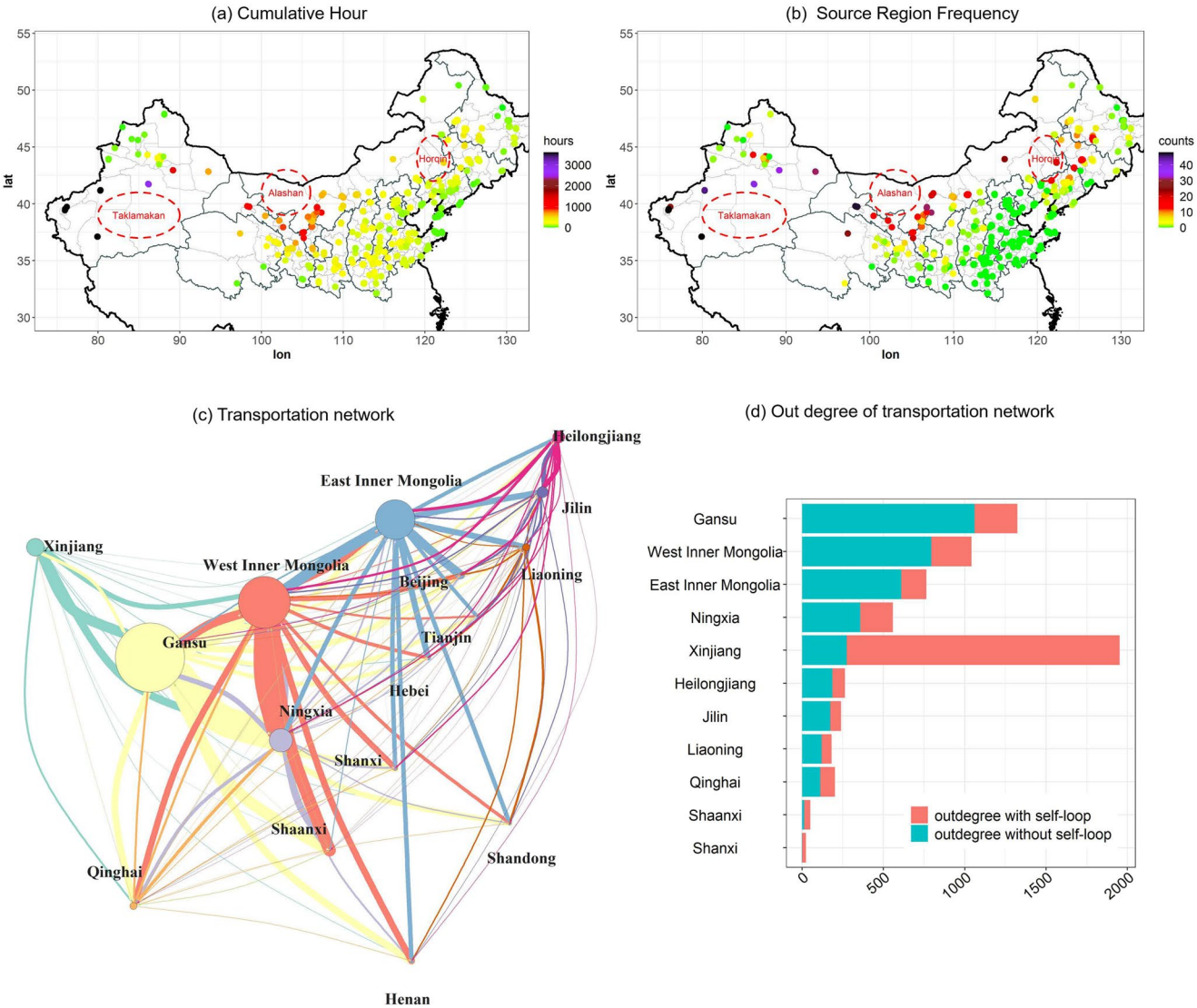

Figure 4. Cumulative affected hours (a) and frequencies (b) of dust events originated (first 10% of episode hours) at the air quality sites in spring time from 2015 to 2020 according to Equation 5; provincial transportation network of the dust events (c) and its out-degree with or without self-loop from 2015 to 2020 (d). In the network in (c), vertices (representing provinces) are organized to reflect the geographical location of the provinces with the radius of the circular vertices determined by their out-degrees (the severity of transported dust events originated in the province). The edges are drawn with the color of the source provinces with their width corresponding to the number and duration of the downstream provinces transported. Self-loops are not counted in the network.

Horqin sandy land as a source region is less prominent in the discussion of dust events in North China. The events originated in Horqin was characterized by shorter average duration (17.8 hr, standard error (SE): 2.1 hr) as compared with the average 25.9 hr duration (SE: 2.1 hr) in North China, as derived from Figure S6 of Supporting Information S1. However, it is an important source for the dust events due to its high frequency of dust events as shown in Figures 4b and 5a.

Figure 5 reports the annual frequencies, the average duration, the average within episode PM_{10} concentration and the affected area of the dust episodes with respect to the five main source regions in the spring from 2015 to 2020, calculated via Equation 6. More detailed results on these aspects are available in Figures S3 and S4 in Supporting Information S1 with respect to the source region, and Figures S5 and S6 in Supporting Information S1 for annual cumulative dust duration in the city and province levels, respectively. Figure 5 shows that the dust events originated in Xinjiang tended to have longer duration and larger affected area than those originated from the other areas. In contrast those originated in Horqin had shorter duration and less affected areas among the five source region. However, the within episode average PM_{10} concentration were largely similar across the source regions. Despite

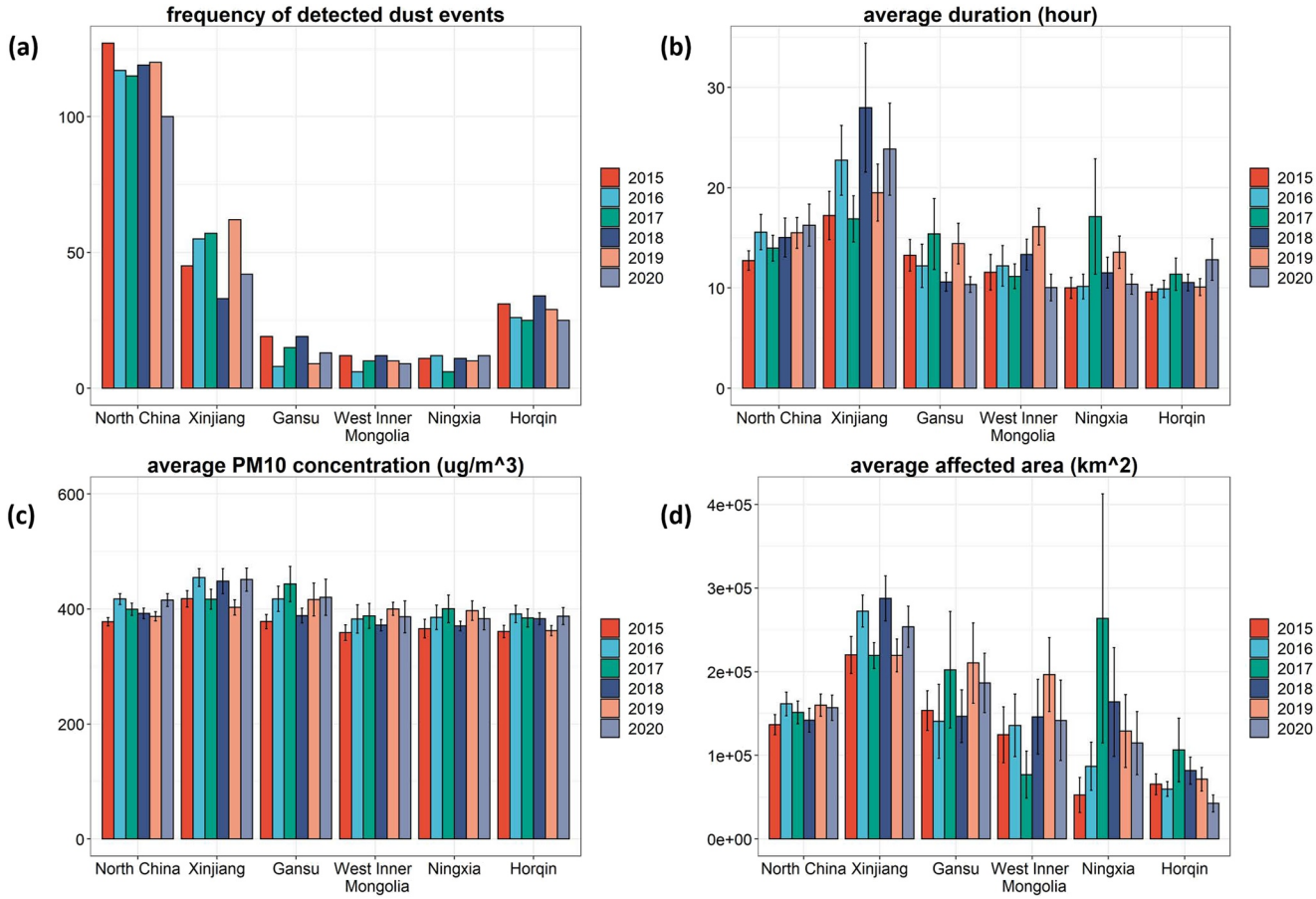

Figure 5. (a) Yearly frequencies, (b) the average duration, (c) the average within episode PM_{10} concentration, and (d) the average affected area of the detected dust events according to Equation 6 with respect to the source provinces or regions, and North China as a whole, respectively. The duration and the PM_{10} concentration were averaged by the number of affected sites and the duration of the dust events, respectively, while the affected area were averaged by the hourly affected area within the episodes.

some annual variations for Ningxia and Xinjiang, there were in general no clear temporal trend with respect to the four measures for each source regions and North China as a whole. This suggests the intensity and severity of the dust event in North China as a whole had been static, although there were some variation in the provincial level. There was a drop in the number of dust events in North China in 2020 largely due to a drop in Xingjinag. As the events in Xingjiang were largely confined to itself, the number of episodes which can be transported out of the source regions was largely unchanged in 2020.

4.3. Dynamic Characteristics

Having studied the static features of the dust events in North China, we consider their dynamic aspects via the transportation networks, which are constructed based on the detected episodes. The networks allow understanding on the dynamic transportation of the dust events in North China. Specifically, we build directed graphs to summarize the transportation network as shown in Figure 4c, where the nodes of different colors represent the source provinces and the directed edges of the same color count for all transportation from the source to downstream provinces. The width of the edges reflects the frequency of the transportation, which is related to the notion of out-degree of a directed edge. The out-degree from nodes A to B is the number of directed connections from A to B. Collection of all out-degrees among the provinces form the adjacency matrix. The adjacent matrix associated with the network shown in Figure 4c is reported in Table S3 of Supporting Information S1, as well as those for the yearly transportation networks and their associated adjacency matrices.

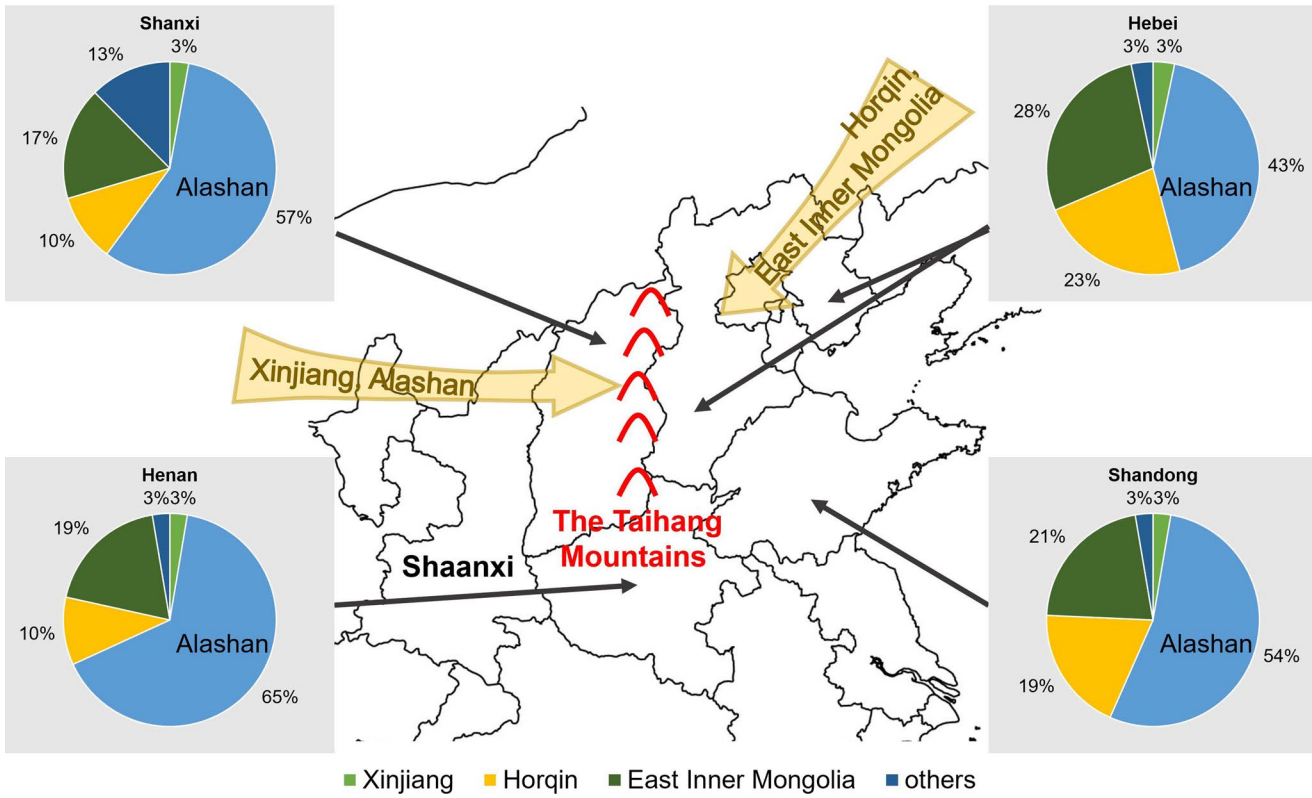

Figure 6. Contributions of the main dust source regions to the city-level averaged dust duration in four down-stream provinces (Shanxi, Hebei, Henan, and Shandong) in spring time from 2015 to 2020. The Taihang Mountain and these downstream provinces are labeled in the map.

Figures 4c and 4d provide the transportation network from the source to down-stream provinces and break-down of events between those that moved out the source provinces (the out-degree) and those stayed within the source region. It shows that the dust events originated in Taklamakan desert, despite having the highest frequencies, were less likely to move out of Xinjiang. This may be interpreted by the special terrain of Xinjiang where the Taklamakan desert is located at the center of Tarim Basin surrounded by mountains, as well as Xingjiang's big geographical area and isolation. Alashan desert (consists of Gansu, Ningxia and Western Inner Mongol), which is part of Gobi desert was the second ranked source region. However, it is the most influence source region to the more populated part of China, with the out-degree of vertices in Gansu, Ningxia, and the West Inner Mongolia (west of $111°E$) contributed 59.8% of the total out-degrees from 2015 to 2020 in North China. For comparison, Horqin, East Inner Mongolia(east of $111°E$) and Xinjiang contributed 13.0%, 16.5% and 7.4% of the total out-degree, respectively, over the same period. Thus, dust storm control on the Alashan desert would be the most important in the dust mitigation effort in North China.

We obtained the contribution of dust-events from the source regions to the downstream provinces from the adjacency matrices. Figure 6 reports the contributions to the four major downstream provinces, Hebei, Henan, Shandong, and Shanxi, in North China overt the six springs. Alashan was the biggest source region contributing 43%–65% of the dust events of the four provinces. The influence of terrain undulation on dust transport can be seen from Figure 6. The Taihang Mountains that partitions Shanxi and Hebei in the north-south direction, blocked some of the dust events originated in the western source regions like Xinjiang and Alashan, and made them having a reduced contribution on Hebei (43%) than that on Shanxi (57%). The dust events originated from the west regions had to make a detour through Shaanxi to affect Henan and Shandong. This demonstrates the effectiveness of our algorithm as it reflected the influence of terrain on the dust transport. It is also observed that Hoqin had quite significant effects on Hebei and Shandong, made to 19%–23% contributions. It is noted that the transported effects on the downstream provinces were not necessarily negatively correlated with the distance, for instance Alasan's contributions to Hebei versus Shandong and Henan, as the wind may lost momentum to make the dust stayed in a region longer.

5. Changing Correlation Pattern

In this section, we study the correlation patterns between PM_{10} and the other five pollutants before and after the start of a dust event. A dust event typically elevates PM_{10} rapidly while the other pollutants may not change much, which may alter the pattern of dependence before and after the dust events. Our study shows a clear decreasing trend in the correlation between PM_{10} and NO_2, SO_2 and CO at the start of the dust event, respectively. This result provides the underlying reason why the proposed algorithm with the covariance information between PM_{10} and other pollutants led to better the dust-event detection as shown in Table 1.

Suppose that $\tau_j^{(i)}$ is a start time of a detected episode C_j at s_i, we extract the time series of the PM_{10} and the correlation coefficients over a time window $[\tau_j^{(i)} - 24, \tau_j^{(i)} + 72]$. To reduce the noise in the time series, we applied the one sided Nadaraya-Watson kernel smoothing estimator to generate the smoothed time series with the 95% confidence intervals (Härdle, 1990) as shown in Figure 7. The smoothing was conducted with the Gaussian kernel with the bandwidth h selected by the cross-validation method, and the confidence intervals are constructed by the block bootstrap that resamples on the estimated residuals (Davison & Hinkley, 1997; Hardle & Mammen, 1993; Liang et al., 2015); see Text S5 of Supporting Information S1 for details. Figure 7 displays the smoothed correlations and the pollutant levels 24 hr before and 72 hr after the start of the dust events by averaging over all sites (North China), sites at Beijing-Tianjin-Hebei (BTH) region and sites located at Gansu, Ningxia and west Inner Mongolia (Alashan). Having the three regions is to compare and identify any differences in the pattern of the correlation among the source and the downstream regions.

The right panels of Figure 7 reveal that as a dust event arrives, there were rapid increase in PM_{10} and $PM_{2.5}$, decrease in SO_2, NO_2, and CO and rather steady ozone level, propelled by increase in the wind speeds at the surface and 500hpa with the ground's wind increased more (47%) than that at the 500 hPa (around 35%). There was increase of the wind speed at 500 hPa which started about 12 hr earlier than the dust events as shown in Figure S9 in Supporting Information S1. These led to the correlation pattern reported on the left panels, which show that PM_{10}'s correlation with $PM_{2.5}$ was largely steady before and after the start of the dust event, with slightly higher (less) correlation in Alashan (downstream Beijing-Tianjin-Hebei) region after the start of the dust event. In the meanwhile, as suggested by the right panel of Figure 7, the relative increase of PM_{10} was higher than that of $PM_{2.5}$, which is consistent with the observation that the dust events were dominated by PM_{10} with a smaller $PM_{2.5}$ to PM_{10} ratio during the storm (Alghamdi et al., 2015; Hoffmann et al., 2008). The correlation pattern between PM_{10} and other pollutants and wind speed were quite robust with the affected area of dust events, as the cross-province dust events (dust events affected more than one provinces) showed similar pattern compared with those results from all dust events (See Figure 7 and Figures S8–S10 in Supporting Information S1).

The most interesting feature was the finding of the significant reduction in the correlations between the three gaseous pollutants (SO_2, NO_2 and CO) and PM_{10} with the correlation dropped from around 0.3 before the coming of the dust to below 0 soon after the dust's arrival and then recovered to around 0 after the 5th hr and then recovered to the pre-episode level in about 2 days.

There are two opposing effects of dust on ozone as is known in the literature (He & Carmichael, 1999; Wang et al., 2020). First, the dust particulates can destruct ozone by heterogeneous reaction at the surfaces of the dust aerosols and may decrease the generation of ozone by photochemistry (Bonasoni et al., 2004; He & Carmichael, 1999). But as BTH region is VOC-sensitive, the decrease of ozone precursors, NO_x, could lead to an increase of ozone as discussed in Tong et al. (2020) and Q. Zhang et al. (2020). Li et al. (2017) discussed the net effect of dust to ozone concentration based on WRF-Chem model in a case study in Nanjing, and concluded a negative effect of dust on ozone. However, our study reveals that the ozone level was less affected by the dust as the correlation of PM_{10} and O_3 after the start of the dust events largely stay around zero. There was even a slight gradual elevation in the correlation between PM_{10} and ozone after the dust events, especially in the BTH region, which replicated an early observation Xie et al. (2005) made in Beijing, and was in line with a recent study by Wang et al. (2020) for the aerosol-radiation interactions. This deserves further study as there are hidden factors and uncertainties in modeling the heterogeneous chemistry processes in ozone generation.

Motivated by Figure 7, we assessed the covariance structures in the rapid increasing stage of dust episodes with respect to their longitudes in Figure S11 of Supporting Information S1. We partitioned the study region with respect to the longitude from 75°E to 125°E in every 10°, and the latitude ranged from 35°N to 40°N. Thus, we roughly obtained five segments from Northwest to Northeast China and from the source region of dust storms

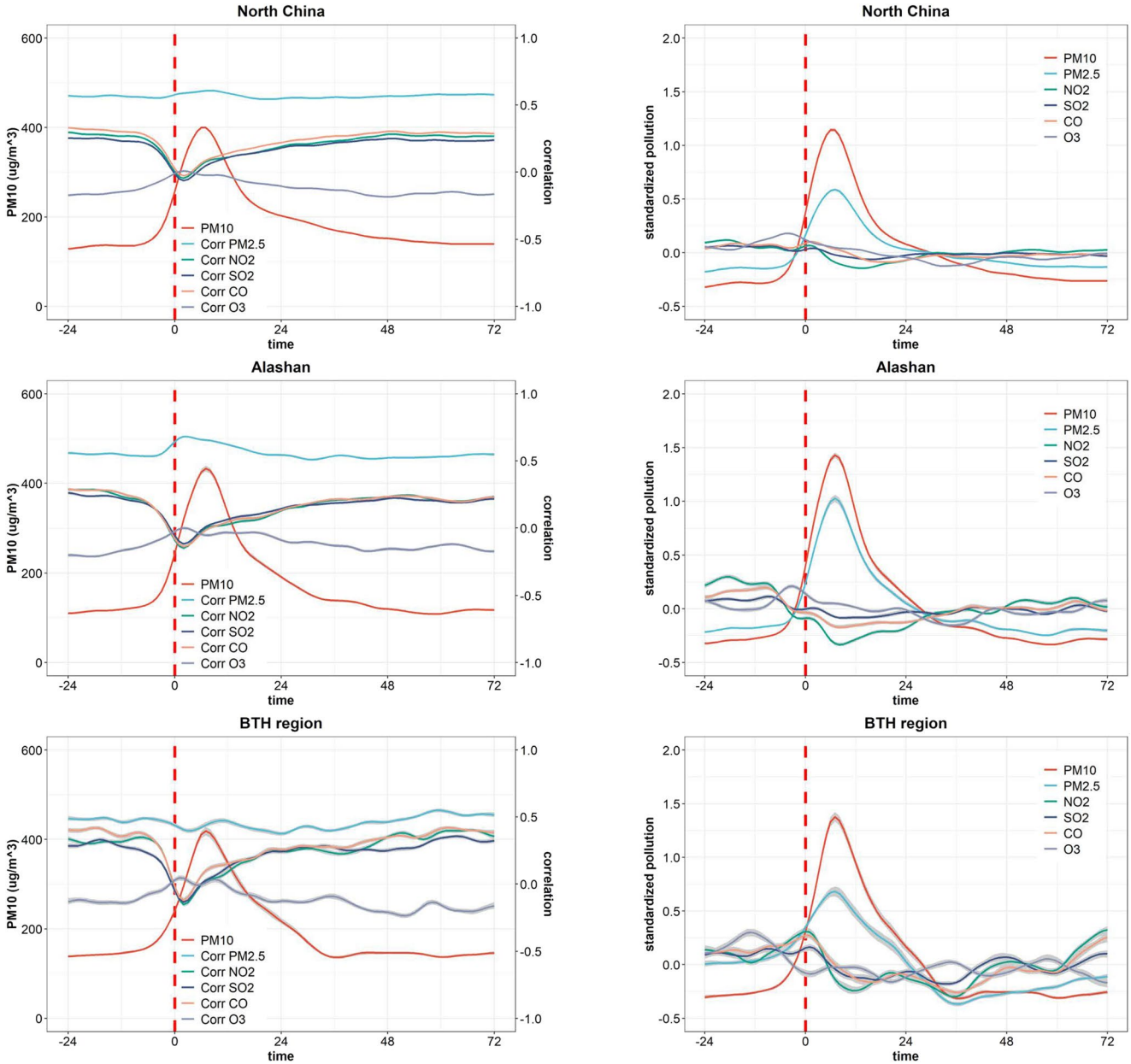

Figure 7. Averaged time series of several covariables and correlation coefficients in different regions at the beginning of dust events. Left panels: PM_{10} and its correlation coefficients with five other pollutants. Right panels: standardized concentration of the six pollutants by first subtracting their respective means and then divided by the standard deviations. The red dashed lines denote the starting time of the dust events. One sided Nadaraya-Watson kernel smoothing method was applied with the Gaussian kernel and the smoothing bandwidth selected by the cross-validation.

to the highly developed industrial area. Because there is no air quality sites from 85°E to 95°E, Figure S11 in Supporting Information S1 actually showed only four segments. The rapid increasing stage is defined from $t = 0$ when the algorithm detected the start of a dust episode to 6 hr after the start time. It can be seen from Figure S11 in Supporting Information S1 that the two source regions (left two segments) showed higher correlations between PM_{10} and $PM_{2.5}$ with a negative correlation between PM_{10} and O_3. Besides, the correlation between PM_{10} and other three anthropogenic pollutants were all negative except in Xinjiang province, which may be explained by its low human emission level. The correlation between PM_{10} and O_3 was increasing in the two right segments in the more industrialized regions where the dust was largely transported from the source region. This suggests that the correlation between PM_{10} and other pollutants was influenced by the background emission intensity and human activities, which led to the spatial heterogeneity as shown in Figure 7.

Table 2
Average Concentrations (Standard Errors) of Airborne Particulate Matters PM_{10} and $PM_{2.5}$ in Spring From 2015 to 2020 in Three Aspects Based on All Observations (All Time), Observations Excluding the Dust Periods (No Dust) and Observations During Dust Periods (Dust) in Six Down-Stream Provinces

	Mean spring PM_{10} (μg/m^3)				Mean spring $PM_{2.5}$ (μg/m^3)			
Province	All time	No dust	Dust	Net effect	All time	No dust	Dust	Net effect
Beijing	101.7 (1.13)	96.9 (0.97)	418.1 (9.35)	4.7 4.9%	59.9 (0.82)	58.1 (0.78)	177.9 (2.45)	1.8 3.1%
Tianjin	110.3 (1.23)	107.5 (1.07)	575.4 (19.97)	2.8 2.6%	59.2 (0.77)	58.5 (0.75)	163.9 (5.52)	0.6 1.1%
Hebei	118.5 (0.59)	114.6 (0.54)	402.1 (4.87)	3.9 3.4%	54.7 (0.37)	53.7 (0.36)	123.3 (2.36)	0.9 1.7%
Henan	122.9 (0.53)	118.8 (0.45)	377.9 (3.61)	4.2 3.5%	59.5 (0.28)	58.7 (0.28)	108.3 (1.80)	0.8 1.4%
Shandong	117.1 (0.49)	113.7 (0.44)	390.7 (4.98)	3.3 2.9%	54.4 (0.25)	53.6 (0.25)	116.6 (2.00)	0.8 1.4%
Shanxi	114.0 (0.58)	109.3 (0.51)	389.2 (4.39)	4.7 4.3%	49.8 (0.30)	48.7 (0.29)	115.3 (2.00)	1.1 2.3%

Note. The net effects were the difference between "all time" and "no dust"' in the absolute and relative terms.

6. Contribution to PM_{10} and $PM_{2.5}$ From Dust Events

Based on the detected dust events by the two-stage procedure, we assess the contribution of the dust events on ambient particulate matters PM_{10} and $PM_{2.5}$'s concentrations. Compared with the existing results such as (Pederzoli et al., 2010), our method can be employed at higher temporal resolution (using hourly observation) and higher spatial resolution (at each air quality site). The scope of the temporal and spatial domains can be readily extended over longer periods and over larger areas.

The enumeration of the dust event's contribution is made as follows. At each site s_i, we computed three forms of the average PM_{10} and $PM_{2.5}$, respectively. One is the whole series average including the dust periods (raw average), the other is the average levels excluding the dust periods (background level), and finally the average levels within the dust periods.

The impacts of the transported dust on the air quality as represented by PM_{10} and $PM_{2.5}$ levels on the four downstream provinces (Hebei, Shanxi, Shandong and Henan) as well as the two mega-cities, Beijing and Tianjin, are reported in Table 2. It shows that the dust events can elevate the average springtime PM_{10} and $PM_{2.5}$ concentrations by 3.4% and 1.7% for Hebei, 3.5% and 1.4% for Henan, 2.9% and 1.4% for Shandong, respectively, while its influence was higher in Shanxi by 4.3% and 2.3% ,respectively, possibly by the front shielding effect of Taihang mountains. The contributions to Beijing was 4.9% and 3.1%, respectively, which maybe explained by its locations to the northern edge of the downstream region.

A more complete result is shown in Figure 8, which gives the difference and its corresponding percentage of contributions (compared with background average) the raw and background data in the 6 years spring time from 2015 to 2020. The contribution of dust events is highest at Xinjiang province, which contributed to a 51.3 μg/m^3 (59.2%) increasing in ground PM_{10} concentration and a 19.4 μg/m^3 (55.0%) increasing in $PM_{2.5}$. As mentioned in Section 4.2, this is mainly because the extremely high affected hours of dust events around Taklamakan desert. In China's capital Beijing, the affect of dust events is also huge as they caused a 4.7 μg/m^3 (4.9%) and 1.8 μg/m^3 (3.1%) increasing in 6 years average spring PM_{10} and $PM_{2.5}$ concentrations, respectively. There is also a downward trend of dust contribution from source regions to their downstream areas. For details of the provincial as well as annual results, see Table S10 in the separately uploaded spreadsheet attachment.

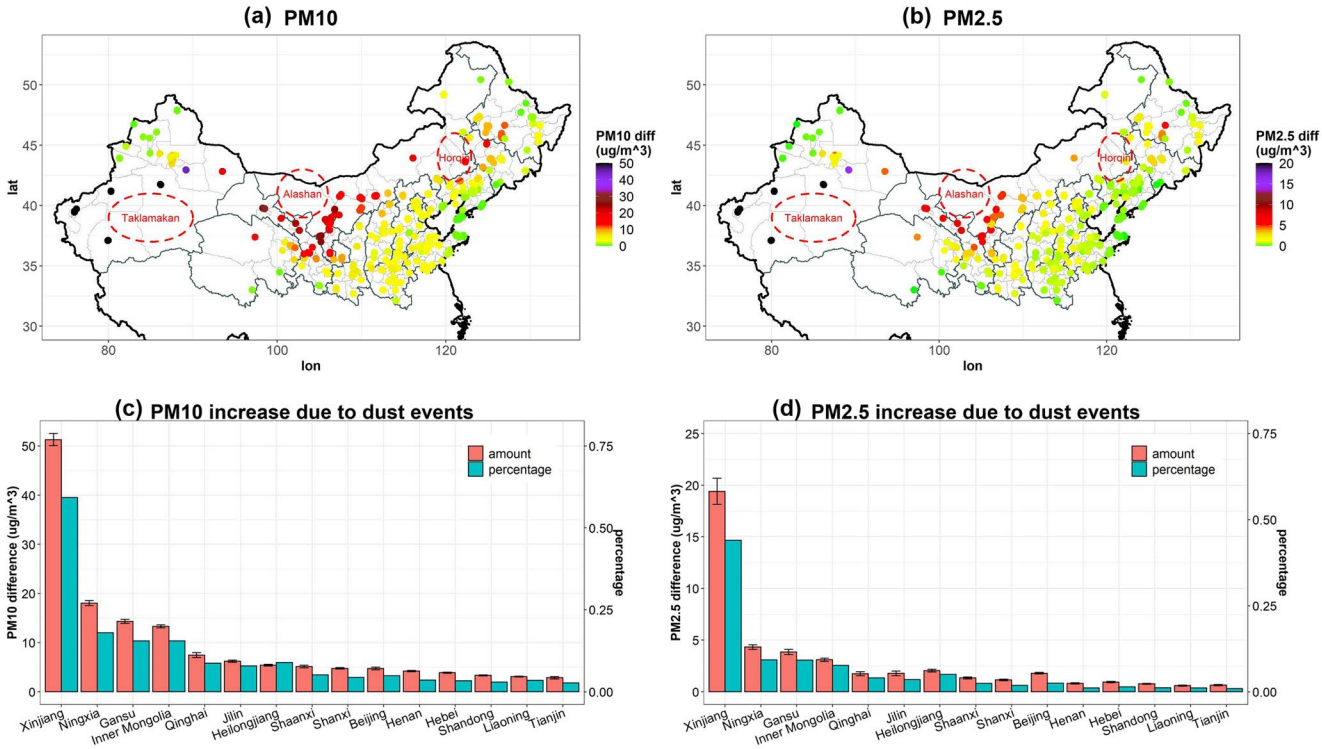

Figure 8. Contributions of dust events to the 6 years spring PM_{10} and $PM_{2.5}$ concentrations. Panels (a and b) sitewise concentration differences for PM_{10} and $PM_{2.5}$. (c) and (d) provincial results of concentration differences. The percentages are compared with background (no dust) concentration.

7. Conclusion

Our study provides a novel two-stage detection and tracking procedure for spatial-temporal moving dust events. This procedure is highly adaptable with no constrains on the shape, size, duration and the number of the target events, and can be achieved with reasonable computing complexity. Empirical study showed that the approach attained high accuracy in dust event detection with the ability to track and identify the affected region of dust events from the source region to the end of the event. It is also demonstrated that by using covariance information better discrimination between dust and pollution events can be attained.

The spatial resolution and coverage of the proposed approach is limited by the spatial density of the air quality monitoring network. The current air quality network in China has varying spatial density between the less populated source regions and the more populated downstream regions. As the provinces in China are building their respective provincial monitoring networks, the spatial disparity in the distributions of the monitoring sites will be alleviated over time. Transnational monitoring data, like data from the Outer Mongolia, could also be helpful for accurate source region detection in future. Another limitation is that, the observed PM_{10} concentrations are truncated by an upper bound (typically at $1500\mu g/m^3$), which may under-estimate the most severe sand storms, but was higher enough for our algorithm whose threshold was 290 $\mu g/m^3$. Finally, in the stage one, our algorithm detects change points separately while the epidemic change point detection strategy can be applied to detect both start and end time simultaneously (Hao et al., 2021), which can be considered in future works.

The proposed approach could lead to future works which may focus on the network structure of the dust events and find the causal relationship between the meteorology and dust event in the identified source region. Indeed, predicting future dust events is an interesting but not stated objective in the current investigation. Viewing from both the spatial and temporal coverage, we believe that satisfactory short term predictions can be extended. By incorporating available information—geographical locations, weather condition and weather forecast, current occurrence of dust events, the chance of the future occurrence of a dust storm at a given city can be quantified by using, for example, the techniques for investigating future time-to-events (Elashoff et al., 2016; Lawless, 2003; Meeker & Escobar, 1998). Longer time-horizon prediction with broader objectives is also possible. For example,

if we focus more on the associations between the dust storms and the meteorology, then the overall severity of the dust storms can be predicted by considering the meteorological variables as covariates and build the relationship between them and the dust event.

Data Availability Statement

The ERA-5 reanalysis data used in this study were collected from the ECMWF website (https://apps.ecmwf.int/). We archived our pollution data in the repository (https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data).

Acknowledgments
The authors thank three anonymous reviewers for comments and suggestions which have improved the content and presentation of the paper. The research was partially supported by National Natural Science Foundation of China Grants 92046021, 12026607, and 12071013, China's National Key Research Special Program Grant 2016YFC0207701, and LMEQF at Peking University.

References

Alghamdi, M. A., Almazroui, M., Shamy, M., Redal, M. A., Alkhalaf, A. K., Hussein, M. A., & Khoder, M. I. (2015). Characterization and elemental composition of atmospheric aerosol loads during springtime dust storm in western Saudi Arabia. *Aerosol and Air Quality Research*, *15*(2), 440–453. https://doi.org/10.4209/aaqr.2014.06.0110

Bener, A., Abdulrazzaq, Y., Al-Mutawwa, J., & Debuse, P. (1996). Genetic and environmental factors associated with asthma. *Human Biology*, *68*(3), 405–414.

Bonasoni, P., Cristofanelli, P., Calzolari, F., Bonafè, U., Evangelisti, F., Stohl, A., et al. (2004). Aerosol-ozone correlations during dust transport episodes. *Atmospheric Chemistry and Physics*, *4*(5), 1201–1215. https://doi.org/10.5194/acp-4-1201-2004

Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Yang, J., Chen, P.-Y., et al. (2019). Extreme gradient boosting model to estimate $PM_{2.5}$ concentrations with missing-filled satellite data in China. *Atmospheric Environment*, *202*, 180–189. https://doi.org/10.1016/j.atmosenv.2019.01.027

CMA. (1979). *Regulations of surface meteorological observation*. Meteorological Press.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Duczmal, L., & Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, *45*(2), 269–286. https://doi.org/10.1016/s0167-9473(02)00302-x

Elashoff, R., Li, G., & Li, N. (2016). *Joint modeling of longitudinal and time-to-event data*. CRC Press LLC.

Guo, J., Xu, H., Liu, L., Chen, D., Peng, Y., Yim, S. H., et al. (2019). The trend reversal of dust aerosol over East Asia and the North Pacific Ocean attributed to large-scale meteorology, deposition, and soil moisture. *Journal of Geophysical Research: Atmospheres*, *124*(19), 10450–10466. https://doi.org/10.1029/2019jd030654

Han, L., Tsunekawa, A., Tsubo, M., & Zhou, W. (2013). An enhanced dust index for Asian dust detection with MODIS images. *International Journal of Remote Sensing*, *34*(19), 6484–6495. https://doi.org/10.1080/01431161.2013.802055

Han, Z., Li, J., Xia, X., & Zhang, R. (2012). Investigation of direct radiative effects of aerosols in dust storm season over East Asia with an online coupled regional climate-chemistry-aerosol model. *Atmospheric Environment*, *54*, 688–699. https://doi.org/10.1016/j.atmosenv.2012.01.041

Hao, N., Niu, Y. S., Xiao, F., & Zhang, H. (2021). A super scalable algorithm for short segment detection. *Statistics in Biosciences*, *13*(1), 18–33. https://doi.org/10.1007/s12561-020-09278-z

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge University Press.

Hardle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, *21*(4), 1926–1947. https://doi.org/10.1214/aos/1176349403

He, S., & Carmichael, G. R. (1999). Sensitivity of photolysis rates and ozone production in the troposphere to aerosol properties. *Journal of Geophysical Research: Atmospheres*, *104*(D21), 26307–26324. https://doi.org/10.1029/1999jd900789

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoffmann, C., Funk, R., Sommer, M., & Li, Y. (2008). Temporal variations in PM_{10} and particle size distribution during Asian dust storms in Inner Mongolia. *Atmospheric Environment*, *42*(36), 8422–8431. https://doi.org/10.1016/j.atmosenv.2008.08.014

Huang, J., Minnis, P., Chen, B., Huang, Z., Liu, Z., Zhao, Q., et al. (2008). Long-range transport and vertical structure of Asian dust from CALIPSO and surface measurements during PACDEX. *Journal of Geophysical Research: Atmospheres*, *113*(D23). https://doi.org/10.1029/2008jd010620

Huang, J., Wang, T., Wang, W., Li, Z., & Yan, H. (2014). Climate effects of dust aerosols over East Asian arid and semiarid regions. *Journal of Geophysical Research: Atmospheres*, *119*(19), 11398–11416. https://doi.org/10.1002/2014jd021796

Krasnov, H., Katra, I., & Friger, M. (2016). Increase in dust storm related PM_{10} concentrations: A time series analysis of 2001–2015. *Environmental Pollution*, *213*, 36–42. https://doi.org/10.1016/j.envpol.2015.10.021

Lawless, J. (2003). *Statistical models and methods for lifetime data*. Wiley.

Lee, J., Sun, Y., & Chang, H. H. (2020). Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, *31*(2), e2578. https://doi.org/10.1002/env.2578

Li, M., Wang, T., Han, Y., Xie, M., Li, S., Zhuang, B., & Chen, P. (2017). Modeling of a severe dust event and its impacts on ozone photochemistry over the downstream Nanjing megacity of eastern China. *Atmospheric Environment*, *160*, 107–123. https://doi.org/10.1016/j.atmosenv.2017.04.010

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H. Z., Zhang, S. Y., et al. (2015). Assessing Beijing's $PM_{2.5}$ pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *471*(2182), 20150257. https://doi.org/10.1098/rspa.2015.0257

Lin, P.-S. (2014). Generalized scan statistics for disease surveillance. *Scandinavian Journal of Statistics*, *41*(3), 791–808. https://doi.org/10.1111/sjos.12063

Ma, X., Wang, J., Yu, F., Jia, H., & Hu, Y. (2016). Can MODIS AOD be employed to derive $PM_{2.5}$ in Beijing-Tianjin-Hebei over China? *Atmospheric Research*, *181*, 250–256. https://doi.org/10.1016/j.atmosres.2016.06.018

Meeker, W., & Escobar, L. (1998). *Statistical methods for reliability data*. Wiley.

Neill, D. B., Moore, A. W., Sabhnani, M., & Daniel, K. (2005). Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 218–227). Association for Computing Machinery. https://doi.org/10.1145/1081870.1081897

Nobakht, M., Shahgedanova, M., & White, K. (2021). New inventory of dust emission sources in central Asia and northwestern China derived from MODIS imagery using dust enhancement technique. *Journal of Geophysical Research: Atmospheres*, *126*(4). https://doi.org/10.1029/2020jd033382

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

Pederzoli, A., Mircea, M., Finardi, S., di Sarra, A., & Zanini, G. (2010). Quantification of Saharan dust contribution to PM_{10} concentrations over Italy during 2003–2005. *Atmospheric Environment*, *44*(34), 4181–4190. https://doi.org/10.1016/j.atmosenv.2010.07.031

Peters, A., Fröhlich, M., Döring, A., Immervoll, T., Wichmann, H.-E., Hutchinson, W., et al. (2001). Particulate air pollution is associated with an acute phase response in men. Results from the MONICA–Augsburg Study. *European Heart Journal*, *22*(14), 1198–1204. https://doi.org/10.1053/euhj.2000.2483

Querol, X., Pey, J., Pandolfi, M., Alastuey, A., Cusack, M., Pérez, N., et al. (2009). African dust contributions to mean ambient PM_{10} mass-levels across the Mediterranean Basin. *Atmospheric Environment*, *43*(28), 4266–4277. https://doi.org/10.1016/j.atmosenv.2009.06.013

Rosenfeld, D., Rudich, Y., & Lahav, R. (2001). Desert dust suppressing precipitation: A possible desertification feedback loop. *Proceedings of the National Academy of Sciences*, *98*(11), 5975–5980. https://doi.org/10.1073/pnas.101122798

Sokolik, I. N., & Toon, O. B. (1996). Direct radiative forcing by anthropogenic airborne mineral aerosols. *Nature*, *381*(6584), 681–683. https://doi.org/10.1038/381681a0

Somers, C. M., Yauk, C. L., White, P. A., Parfett, C. L. J., & Quinn, J. S. (2002). Air pollution induces heritable DNA mutations. *Proceedings of the National Academy of Sciences*, *99*(25), 15904–15907. https://doi.org/10.1073/pnas.252499499

Sowden, M., Mueller, U., & Blake, D. (2018). Review of surface particulate monitoring of dust events using geostationary satellite remote sensing. *Atmospheric Environment*, *183*, 154–164. https://doi.org/10.1016/j.atmosenv.2018.04.020

Sprigg, W. A., Nickovic, S., Galgiani, J. N., Pejanovic, G., Petkovic, S., Vujadinovic, M., et al. (2014). Regional dust storm modeling for health services: The case of valley fever. *Aeolian Research*, *14*, 53–73. (Airborne Mineral Dust Contaminants: Impacts on Human Health and the Environment). https://doi.org/10.1016/j.aeolia.2014.03.001

Tian, Y., Pan, X., Wang, Z., Wang, D., Ge, B., Liu, X., et al. (2020). Transport patterns, size distributions, and depolarization characteristics of dust particles in East Asia in spring 2018. *Journal of Geophysical Research: Atmospheres*, *125*(16). https://doi.org/10.1029/2019jd031752

Tong, P., Zhang, Q., Lin, H., Jian, X., & Wang, X. (2020). Simulation of the impact of the emergency control measures on the reduction of air pollutants: A case study of APEC blue. *Environmental Monitoring and Assessment*, *192*(2), 116. https://doi.org/10.1007/s10661-019-8056-1

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., & Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(1), 49–82. https://doi.org/10.1111/j.1467-985x.2011.00714.x

van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2015). Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental Health Perspectives*, *123*(2), 135–143. https://doi.org/10.1289/ehp.1408646

Wang, Z., Huang, X., Wang, N., Xu, J., & Ding, A. (2020). Aerosol-radiation interactions of dust storm deteriorate particle and ozone pollution in East China. *Journal of Geophysical Research: Atmospheres*, *125*(24). https://doi.org/10.1029/2020jd033601

Wong, M. S., Xiao, F., Nichol, J., Fung, J., Kim, J., Campbell, J., & Chan, P. (2015). A multi-scale hybrid neural network retrieval model for dust storm detection, a study in Asia. *Atmospheric Research*, *158–159*, 89–106. https://doi.org/10.1016/j.atmosres.2015.02.006

Xie, S., Yu, T., Zhang, Y., Zeng, L., Qi, L., & Tang, X. (2005). Characteristics of PM_{10}, SO_2, NO_x and O_3 in ambient air during the dust storm period in beijing. *Science of the Total Environment*, *345*(1–3), 153–164. https://doi.org/10.1016/j.scitotenv.2004.10.013

Yang, Y. Q., Hou, Q., Zhou, C. H., Liu, H. L., Wang, Y. Q., & Niu, T. (2008). Sand/dust storm processes in Northeast Asia and associated large-scale circulations. *Atmospheric Chemistry and Physics*, *8*(1), 25–33. https://doi.org/10.5194/acp-8-25-2008

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., & Wang, W. (2016). National-scale estimates of ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. *Remote Sensing*, *8*(3). https://doi.org/10.3390/rs8030184

Yu, M., Bambacus, M., Cervone, G., Clarke, K., Duffy, D., Huang, Q., et al. (2020). Spatiotemporal event detection: A review. *International Journal of Digital Earth*, *13*(12), 1339–1365. https://doi.org/10.1080/17538947.2020.1738569

Zhang, Q., Tong, P., Liu, M., Lin, H., Yun, X., Zhang, H., et al. (2020). A WRF-Chem model-based future vehicle emission control policy simulation and assessment for the Beijing-Tianjin-Hebei region, China. *Journal of Environmental Management*, *253*, 109751. https://doi.org/10.1016/j.jenvman.2019.109751

Zhang, R., Arimoto, R., An, J., Yabuki, S., & Sun, J. (2005). Ground observations of a strong dust storm in Beijing in March 2002. *Journal of Geophysical Research: Atmospheres*, *110*(D18). https://doi.org/10.1029/2004jd004589

Zhang, X.-X., Sharratt, B., Liu, L.-Y., Wang, Z.-F., Pan, X.-L., Lei, J.-Q., et al. (2018). East Asian dust storm in May 2017: Observations, modelling, and its influence on the Asia-Pacific region. *Atmospheric Chemistry and Physics*, *18*(11), 8353–8371. https://doi.org/10.5194/acp-18-8353-2018

Zhou, C., Gui, H., Hu, J., Ke, H., Wang, Y., & Zhang, X. (2019). Detection of new dust sources in Central/East Asia and their impact on simulations of a severe sand and dust storm. *Journal of Geophysical Research: Atmospheres*, *124*(17–18), 10232–10247. https://doi.org/10.1029/2019jd030753