A homogenized daily *in situ* PM$_{2.5}$ concentration dataset from national air quality monitoring network in China

Kaixu Bai$^{1,2,3}$, Ke Li$^3$, Chengbo Wu$^3$, Ni-Bin Chang$^4$, Jianping Guo$^5$*

1Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, China

2Institute of Eco-Chongming, 20 Cuiniao Rd., Chongming, Shanghai, China

3School of Geographic Sciences, East China Normal University, Shanghai, China

4Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL, USA

5State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

*Correspondence to: Dr./Prof. Jianping Guo (jpguocams@gmail.com)
Abstract

In situ PM$_{2.5}$ concentration observations have long been used as critical data sources in haze related studies. Due to the frequently occurred haze pollution events, China started to monitor PM$_{2.5}$ concentration nationwide routinely from the newly established air quality monitoring network. Nevertheless, the acquisition of these invaluable air quality samples is challenging given the absence of public available data download interface. In this study, we provided a homogenized in situ PM$_{2.5}$ concentration dataset that was created using hourly PM$_{2.5}$ data retrieved from the China National Environmental Monitoring Center (CNEMC) via a web crawler between 2015 and 2019. Methods involving missing value reconstruction, change point detection, and bias adjustment were applied sequentially to deal with data gaps and inhomogeneities in raw PM$_{2.5}$ observations. After excluding records with limited temporal coverage, a homogenized PM$_{2.5}$ concentration dataset comprising of 1,309 five-year long daily PM$_{2.5}$ data series was eventually compiled. This is the first thrust to homogenize in situ PM$_{2.5}$ observations in China. The trend estimations derived from the homogenized dataset indicate a spatially homogeneous decreasing tendency of PM$_{2.5}$ across China at a mean rate of about -7.6% per year from 2015 to 2019. In contrast to raw PM$_{2.5}$ observations, the homogenized data record not only has a complete data integrity but is more consistent over space and time. This homogenized daily in situ PM$_{2.5}$ concentration dataset is publicly accessible at https://doi.pangaea.de/10.1594/PANGAEA.917557 (Bai et al., 2020a), which can be applied as a promising dataset for PM$_{2.5}$ related studies such as PM$_{2.5}$ mapping, human exposure risk assessment, and air quality management.

Keywords: PM$_{2.5}$; Data homogenization; Bias correction; In situ observation; Air quality indicators
Introduction

A consistent PM$_{2.5}$ concentration dataset is vital to the analysis of variations in PM$_{2.5}$ loadings over space and time as well as in support of its risk analysis for air quality management, meteorological forecasting, and health-related exposure assessment (Lelieveld et al., 2015; Yin et al., 2020). Ground-based monitoring network is commonly built to measure concentrations of air pollutants in due course across the globe. Suffering from extensive and severe haze pollution events in the past few years (Guo et al., 2014; Ding et al., 2016; Wang et al., 2016; Cai et al., 2017; Huang et al., 2018; Luan et al., 2018; Ning et al., 2018), China launched the operational ambient air quality sampling late in 2012 on the basis of the sparsely distributed aerosol observation network. To date, this in situ network has been enlarged to cover almost all major cities in China consisting of about 1500 monitoring stations. Concentrations of six key air pollutants including PM$_{2.5}$, PM$_{10}$, NO$_2$, SO$_2$, CO, and O$_3$, are routinely measured on an hourly basis while the sampled data are released publicly online by the China National Environmental Monitoring Center (CNEMC) since 2013.

Although in situ PM$_{2.5}$ concentration data have played critical roles in improving our understanding of regional air quality variations and relevant influential factors (Yang D. et al., 2018; Yang Q. et al., 2019; Zheng et al., 2017), little concern was raised to the quality of such dataset itself (Bai et al., 2019a, 2019c; He and Huang, 2018; Zhang et al., 2019, 2018; Zou et al., 2016). Meanwhile, few studies provided a detailed description of the accuracy or bias level (uncertainty) of the observed PM$_{2.5}$ data in recent years (You et al., 2016; Guo et al., 2017; Shen et al., 2018). The primary reason lies in the fact that neither quality assurance flag nor metadata information documenting the uncertainty other than data samplings were provided, making such quality assessment infeasible.

The data quality, in particular the data homogeneity, is of critical importance to the exploration of the given dataset, especially for trend analysis (Bai et al., 2019c; C. Lin et al., 2018; Liu et al., 2018; Ma et al., 2015) and data integration (Bai et al., 2019b, 2020b; T. Li et al., 2017; Zhang et al., 2019) in which a homogeneous dataset is absolutely essential for downstream applications. Since two distinct kinds of instruments are used in the current air quality monitoring network to measure near surface PM$_{2.5}$ concentration around China (Bai et al., 2020), imperfect instrumental calibration and intermittent
replacement of instruments may thus introduce obvious issue of discontinuity in PM$_{2.5}$ observations. Such inhomogeneity may result in large uncertainty and even biased results in the subsequent analysis, especially in context-based and data driven PM$_{2.5}$ concentration mapping (Bai et al., 2019b, 2019a; He and Huang, 2018; Wei et al., 2020), in which in situ PM$_{2.5}$ concentration observations are used as the ground truth to characterize complex relationships with other possible contributing factors.

Given the absence of an open access and quality assured in situ PM$_{2.5}$ concentration dataset in China, in this study, we attempted to generate a long-term coherent in situ PM$_{2.5}$ concentration dataset for scientific community to use in future applications. A set of methods involving missing value reconstruction, change point detect, and bias adjustment were geared up seamlessly in a big data analytic manner to improve the data integrity and to remove discontinuities present in raw PM$_{2.5}$ observations. Such an analytical process is also referred to as data homogenization in data science or big data analytics (Cao and Yan, 2012; Wang et al., 2007). To our knowledge, this is the first thrust to homogenize a large-scale dataset of in situ PM$_{2.5}$ concentration observations in China. In the following sections, we will introduce the data source as well as detailed big data analytics methods used for the creation of a homogenized PM$_{2.5}$ concentration dataset.

2 In situ PM$_{2.5}$ concentration observations

In this study, the hourly PM$_{2.5}$ concentration data sampled at more than 1,600 state-controlled air quality monitoring stations across China from January 1, 2015 to December 31, 2019 were utilized. The PM$_{2.5}$ concentration data are routinely measured on an hourly basis using instruments such as beta-attenuation monitors and Tapered Element Oscillating Microbalance (TEOM) analyzer. The ordinary instrumental calibration and quality control are performed according to the national ambient air quality standard of GB3095-2012 and HJ 618–2011 (Guo et al., 2009, 2017). Generally, TEOM can measure PM$_{2.5}$ concentration within the range of 0–5,000 μg m$^{-3}$ at a resolution of 0.1 μg m$^{-3}$, with precisions of ±0.5 μg m$^{-3}$ for 24-h average and ±1.5 μg m$^{-3}$ for hourly average (Guo et al., 2017; Xin et al., 2012). All PM$_{2.5}$ measurements are publicly released online by the National Urban Air Quality Real-time Publishing Platform (http://106.37.208.233:20035/) under the China National Environmental Monitoring Center (CNEMC) within one hour after the direct sampling.
Although these data sampling are publicly released, the acquisition of these valuable samplings always exhibits a big challenge because no data download interface is provided to the public by the CNEMC website. Therefore, it is impossible for users to retrieve the historical observations from the given website. Rather, science community has to count on other measures such as an automatic web crawler for the retrieval of these online updated data samples from the data publishing platform. Nevertheless, the archived data records through such an approach suffered from significant data losses due to various unexpected reasons like power outage and internet interruption. Consequently, the data integrity becomes problematic and further treatments like gap filling are thus essential to accounting for such defects at least.

Moreover, hourly PM$_{2.5}$ concentration observations that were sampled at five embassies of United States in China from January 2015 to June 2017 were used as an independent dataset to evaluate the fidelity of the homogenized PM$_{2.5}$ concentration dataset. Geographic locations of these five embassies have been shown in Table S1. These PM$_{2.5}$ data were measured independently under the U.S. department of state air quality monitoring program and can be acquired from the http://www.stateair.net/. To be in line with the homogenized dataset, the hourly PM$_{2.5}$ concentration data were aggregated to the daily level by averaging the 24-h observations sampled on each date while daily averages were calculated only for days with more than 12 valid samples of a possible 24-h.

3 Homogenization of in situ PM$_{2.5}$ concentration data

For the creation of a long-term coherent in situ PM$_{2.5}$ concentration dataset, it is necessary to create an analytical framework of the big data analytics which seamlessly gears up several methods as a whole for the purposes of gap filling, change point detection, and discontinuity adjustment. Figure 1 shows a schematic illustration of the general workflow toward generating a homogenized dataset. The whole process can be outlined as follows.

(1) It is necessary to perform the essential quality control and gap filling on the raw PM$_{2.5}$ observations so that the bias arising from large outliers and resampling errors due to incomplete observations can be reduced.
Figure 1. A schematic flowchart for the creation of a homogenized daily in situ PM$_{2.5}$ concentration dataset.

(2) Short-term time series due to sites relocation were temporally merged to attain a long-term record. Then, PM$_{2.5}$ concentration time series with a temporal coverage of less than four-year during the study period were excluded and the quality-controlled observations of hourly in situ PM$_{2.5}$ concentration were resampled to daily and monthly scales to initiate the homogeneity test.

(3) Reference time series were constructed for each long-term PM$_{2.5}$ concentration record using data measured at adjacent monitoring sites in the surroundings. For PM$_{2.5}$ concentration records failing to produce a reference series, no homogeneity test was performed for such datum due to the absence of reference series.

(4) The discontinuity identified in each daily long-term PM$_{2.5}$ concentration time series were adjusted using the quantile-matching (QM) adjustment method according to the detected change points in the monthly record with the support of reference series.
Post-processing measures such as nonpositive value correction and another round gap filling were further performed on the homogenized records to improve the quality to attain a quality-assured in situ PM$_{2.5}$ concentration dataset. More details of each method are described in the following subsections.

3.1 Quality control

Given the possibility of the presence of abnormal samplings, it is essential to removing the outliers detected in the original PM$_{2.5}$ observations to reduce the false alarm rate in change point detection during the subsequent homogeneity test. Specifically, hourly PM$_{2.5}$ concentration data values meeting one of the following criteria were excluded: 1) out of the range between 1 and 1,000 μg m$^{-3}$, and 2) more than three standard deviations from the median of observations within a 15-h time window. Both criteria aimed to remove large outliers which could result in biased daily averages. Overall, 3.46% of PM$_{2.5}$ samples were treated as outliers which were then excluded accordingly (filled with Nan to indicate missing values).

3.2 Gap filling and resampling

As indicated in a recent study (Bai et al., 2020), missing value related data voids become a big obstacle in the raw PM$_{2.5}$ observations that were retrieved from the CNEMC website as PM$_{2.5}$ observations on 40% of sampling days suffered from data losses due to unexpected reasons. To reduce the impact of missing value related resampling (from hourly to daily) bias on the subsequent homogeneity test, we filled those missing value related data gaps that were found in each 24-h PM$_{2.5}$ observation by applying the DCCEOF method developed very recently (Bai et al., 2020b). Such a gap filling effort enabled us to improve the percentage of days without missingness during the study time period from 58.8% to 97.3%.

In spite of the improvement of data integrity after gap filling, the resultant PM$_{2.5}$ time series remain temporally discrete due to the emergence of several long-lasting (e.g., more than 24 consecutive hours) data missing episodes. Also, the hourly time series are still too noisy to be handled by the available homogeneity test software due to the significant variability of PM$_{2.5}$ over space and time. In such context, the hourly PM$_{2.5}$ concentration records were resampled to daily and monthly scales to
initiate the homogeneity test. Moreover, the monthly series was primarily used to detect the possible change points while the daily series was adjusted in reference to the corresponding reference series based on the change points detected from the monthly series. To avoid large resampling bias, monthly averages were calculated only for those with at least 20 valid daily means of a possible month at each site. The frequency of missing values in each month was also calculated as a possible metadata information to further examine the detected change points.

### 3.3 Homogeneity test

A commonly used homogeneity test software, the RHtestsV4 package, was hereby applied to detect the possible discontinuities in raw PM$_{2.5}$ data series that were retrieved from the CNEMC website. As suggested in Wang and Feng (2013), RHtestsV4 is capable of detecting and adjusting change points in a data series with first-order autoregressive errors. Given the low false alarm rate via change point detection and the capability to adjust discontinuity, the RHtests software packages have been widely used to homogenize climate data records such as temperature (Cao et al., 2013; Xu et al., 2013; Zhao et al., 2014), precipitation (Wang et al., 2010a; Nie et al., 2019), and other datum like boundary layer height (Wang and Wang, 2016). Two typical methods, namely the PMTred and PMFred, were embedded in a recursive testing algorithm in RHtestsV4, with the former relying on the penalized maximal $t$ test (PMT) while the latter based on the penalized maximal $F$ test (PMF) (Wang et al., 2007; Wang, 2008a). With the incorporation of these empirical penalty functions (Wang, 2008a, b), the problem of uneven distribution of false alarm rate is largely alleviated with the aid of RHtestsV4. In contrast to the PMF which works without a reference series, the PMT uses a reference series to detect change points and the results are thus far more reliable (Wang, 2008a, b). The way to generate reference series will be described in the next subsection. Also, the RHtestsV4 is capable of making essential adjustments to the detected discontinuities by taking advantage of the QM adjustment method (Wang and Feng, 2013).

Here the PMT method rather than the PMF was used to detect change points given the higher confidence of the former method in change point detection due to the involvement of reference series (Wang and Feng, 2013). To ensure the reliability of detected discontinuities, change point was defined
and confirmed at a nominal 99% confidence level, and the data records were then declared to be homogeneous once no change point was identified. Subsequently, the QM adjustment method was applied to correct PM$_{2.5}$ observations with evident drifts with the support of reference series, namely, to homogenize PM$_{2.5}$ concentration data series. To avoid large sampling uncertainty in the estimate of QM adjustments, the $Mq$ (i.e., the number of categories on which the empirical cumulative distribution function is estimated) was automatically determined by the software to ensure adequate samples for the estimation of mean difference and probability density function. Meanwhile, the number to determine the base segment (i.e., $I_{adj}$) was set to 0 so that datum in other segments were all adjusted to the segment with the longest temporal coverage.

### 3.3.1 Construction of reference series

A good reference series is vital to the relative homogeneity test because it helps pinpoint possible discontinuities in each base series (the data series to be tested) as well as determine the performance of the subsequent data adjustment. In general, reference series can be organized by using one specific record either measured at the adjacent station or aggregated from multiple adjacent observations (Cao and Yan, 2012; Peterson and Easterling, 1994; Xu et al., 2013; Wang et al., 2016). The most straightforward method is to use the neighboring data series either measured at the nearest station or series that are highly correlated with the base series (Peterson and Easterling, 1994; Cao and Yan, 2012; Wang and Feng, 2013). Such methods, however, fail to take the repetitiveness of the neighboring series into account since the neighboring series may also suffer from discontinuities.

To avoid the misuse of inhomogeneous PM$_{2.5}$ concentration records in constructing reference series, a complex yet robust data integration scheme was developed to screen, organize, and construct reference series for each in situ PM$_{2.5}$ concentration data series. For each daily PM$_{2.5}$ concentration data series, all the neighboring series were firstly identified from its surroundings with a lag distance as large as of 50 km. No reference series was constructed once there was no neighboring series available within the given radius and in turn the homogeneity of the given record was not examined. Otherwise, both correlation coefficient (R) and coefficient of variation (CV) were calculated between the given base series and each selected neighboring series to assess their representativeness (Shi et al.,...
2018; Rodriguez et al., 2019). Then, neighboring series with R>0.8 and CV<0.2 were selected as candidates to construct the reference series for a given base series.

The reference series was then constructed by averaging both the base and the candidate series at each observation time if there was only one candidate series. For the situation with more than one candidate series, the empirical orthogonal function (EOF) analysis was applied to these multiple candidates and then the original fields were reconstructed with the leading principal components when the accumulated variance explained by them exceeded 80%. This was expected to reduce the possible impacts of abnormal observations and short-term discontinuities in the neighboring candidates on the resultant reference series. Subsequently, the reference series were organized and constructed through a spatial weighting scheme as each reconstructed record was assigned a spatially resolved weight according to their relative distances to the base series over space. Here we applied a Gaussian kernel function to estimate the weight of one neighboring observation on the other in space and such a scheme has been proven to be effective in assessing the spatial autocorrelation of PM$_{2.5}$ concentration (Bai et al., 2019b). Mathematically, the reference series can be constructed from the following equations:

$$ PM_{ref} = \sum_{i=1}^{N} \frac{w_i * PM_{cand}^i}{\sum w_i} $$

(1)

$$ w = \exp\left(\frac{-d^2}{2h^2}\right) $$

(2)

where $PM_{ref}$ and $PM_{cand}$ denote the reference and candidate series, respectively. $N$ is the total number of candidate series while $w$ is the spatial weight assigned to each candidate series and $d$ is the spatial lag distance between the base and the corresponding candidate series. $h$ is a spatial correlation length that is used to modulate the relative influence of a distant observation on the data measured at the base site. In this study, an empirical value of 50 km was assigned according to the estimated semi-variogram results (Bai et al., 2019b).

For any record having neighboring series within 50 km but poorly correlated (R<0.8 or CV>0.2) to all its neighbors (meaning the base series differ from the neighbors), the reference series were created by following the same procedures as those detailed above by taking the nearest neighbor as the base series. For the situation with only one candidate series available, it is logical to compare both the base and the candidate series against another data to check which one should be corrected. It was noted
that the PM$_{2.5}$ time series estimated from the MERRA-2 aerosol reanalysis in the same way as described in He et al. (2019) was used. The one more correlated to this external PM$_{2.5}$ time series was then used as the reference (deemed as homogeneous) while the other was considered as the base series (i.e., implies to be adjusted). Such an inclusive scheme empowered us to screen and construct reference series for 1,262 long-term PM$_{2.5}$ concentration records across the board. In contrast, no reference series were constructed for 47 isolated records.

### 3.3.2 Post-processing measures

Several post-processing measures were applied to the adjusted data records to further improve the quality of this dataset. Since nonpositive values may appear in the QM adjusted data series if the original values are close to zero (Wang et al., 2010b), nonpositive values were replaced with the smallest valid PM$_{2.5}$ concentration amount measured at each monitoring site during the study period. Subsequently, the data gaps in the adjusted datum due to long-lasting missingness were filled by first calibrating the corresponding data values in the reference series measured on the same date (if available) to the homogenized datum level. The modified quantile-quantile adjustment (MQQA) method proposed in Bai et al. (2016) was hereby used given its adaptive data adjustment principle. For the predicted values, such MQQA scheme rendered higher accuracy than those interpolated from data values measured on adjacent dates because PM$_{2.5}$ concentration is spatially more correlated than in the temporal domain (Bai et al., 2019b). For the remaining data gaps, those missing values were reconstructed in a similar procedure as the DCCEOF method (Bai et al., 2020b). Note that the matrix used for EOF analysis in the context of DCCEOF was constructed using the neighboring data series measured within a radius of 100 km with a temporal lag of 30 days at most. Finally, all data values were rounded to integer to be in line with the original PM$_{2.5}$ concentration observations.

### 4 Results and discussion

#### 4.1 Descriptive statistics

Prior to data homogenization, we first need to exclude those short-term and less reliable records. Figure 2 shows the temporal variations of the number of air quality monitoring stations deployed in China during 2015–2019 as well as the spatial patterns of the frequency of missing values for each
long-term PM$_{2.5}$ concentration record. It shows that a total of about 1,630 air quality monitoring stations had been deployed in China before 2020. Nevertheless, about 1,500 sites routinely providing PM$_{2.5}$ observations were kept up in operation since 2015 (Figure 2a). By referring to the data continuity of PM$_{2.5}$ observations, it is noticeable that 100 monitoring stations had been withdrawn before 2020 because no PM$_{2.5}$ observations were provided for more than three consecutive months since the release of their last valid data (Figure 2b). Meanwhile, 42 pairs of stations were found to be relocated since new stations at nearby started to provide PM$_{2.5}$ observations soon after the suspension of the original site. This is also corroborated by the temporal lags of PM$_{2.5}$ observations between original and newly deployed stations as many of them were found to have a time lag less than 15-day. Also, 94 sites were found with limited data records due to short temporal coverage (newly deployed). Finally, 1,353 long-term PM$_{2.5}$ concentration records were identified with their first valid data released even earlier than 2015. In regard to the frequency of missing value, it is indicative that data gaps were obvious in these long-term PM$_{2.5}$ concentration records, with about 6% of hourly data values missed on $\sim$47% of sampling days on average. This also motivates us to fill such data gaps to improve the data integrity.

**Figure 2.** Spatial and temporal patterns of air quality monitoring stations in China. (a) Temporal variations of the total number of air quality monitoring stations in China. (b) Spatial patterns of the frequency of missing value in each long-term hourly PM$_{2.5}$ concentration record measured from January 1, 2015 to December 31, 2019. Stations were categorized into distinct groups according to their data length and temporal continuity. The frequency of missingness was calculated as the ratio of
the number of missing values in each PM$_{2.5}$ concentration record to the total number of samplings from the time of the release of the first valid data to December 31, 2019.

4.2 Homogenization of in situ PM$_{2.5}$ data

A total of 1,395 long-term (with five-year observations) PM$_{2.5}$ concentration records were acquired with the inclusion of 42 temporally merged data series at those relocated stations. After removing those suffering from more than three consecutive months data losses, 1,309 long-term yet consecutive PM$_{2.5}$ concentration records were obtained. The homogeneity test was finally performed on 1,262 records due to the availability of reference series. Figure 3 shows the spatial patterns of the total number of change points detected in 1,262 monthly PM$_{2.5}$ concentration records. The ubiquitous change points imply that there is an obvious inhomogeneity in this in situ PM$_{2.5}$ concentration dataset. About 57% (719 out of 1,262) of records failed to pass the homogeneity test due to the presence of change points. Given the overall good agreement between the base and reference series (refer to Figure S1 for the correlation coefficient and root mean square error between them), it indicated that these PM$_{2.5}$ concentration records did suffer from evident discontinuities. Meanwhile, the vast majority (~80%) of the inhomogeneous PM$_{2.5}$ records suffered from no more than two change points (Figure 3), suggesting the mean shift could be the primary reason for the detected discontinuities. Moreover, 20 records were even found suffering from no less than five significant change points, indicating phenomenal discontinuities in these records.
Figure 3. Spatial patterns of the total number of change points detected in each long-term yet consecutive PM$_{2.5}$ concentration records. Gray dot indicates there was no change point detected in this PM$_{2.5}$ concentration record.

Figure 4 shows the temporal variability of the number of change points detected in monthly PM$_{2.5}$ concentration records. As indicated, change points were detected in every specific month of the year from May 2015 to July 2019, especially in late spring (e.g., May), in which change points were more likely to be detected (Figure 4b). This is attributable to the seasonality of PM$_{2.5}$ loading in China as high PM$_{2.5}$ concentrations are always observed in the winter whereas low values in the summer. Consequently, change points were detected with larger chance during the chronic transition periods (e.g., spring to summer). In addition, it is noteworthy that a large volume of change points was detected in early 2015, indicating the existence of phenomenal discontinuities during this period (Figure 4a). After checking the temporal variations of PM$_{2.5}$ concentration, findings indicate that PM$_{2.5}$ observations varied with large deviations among each other during this period. This could be linked to the imperfect instrument calibration or irregular operation in the early stage.

Figure 4. Temporal variations of the number of change points detected in (a) each specific month from 2015 to 2019 and (b) each month of the year. National mean PM$_{2.5}$ concentration in each month of the year was calculated based on PM$_{2.5}$ data measured at our selected 1309 sites during 2015–2019.

Due to the lack of essential metadata information, it is a challenge for us to verify each detected change point through a manual inspection. Rather, the variations in the base and reference series was explored to identify the possible reasons for the detected discontinuities. Figure 5 presents three typical
inhomogeneous PM$_{2.5}$ time series with different number of change points. The inter-comparisons between the base and reference series indicate an overall good agreement among them in terms of the long-term variation tendency. However, obvious drifts were still phenomenal in their residual series, which were even more evident by referring to their mean-shift series. For example, both the residual and mean-shift series shown in Figure 5d clearly illustrate a typical discontinuity as there was an obvious departure of mean PM$_{2.5}$ concentration level during the period of January to October 2016. In contrast, the Figures. 5b and 5e present another typical inhomogeneity as statistically significant decreasing trend was found in the residual series with monthly PM$_{2.5}$ concentration deviations decreased from nearly 5 $\mu$g m$^{-3}$ to –4 $\mu$g m$^{-3}$ step wise. Such inhomogeneity would undoubtedly result large bias in the trend estimations over that region. The bottom panel (Figures. 5c and 5f) shows the change points detected in the merged PM$_{2.5}$ time series at a pair of relocated sites. It is noteworthy that the detected discontinuity should be largely ascribed to the inconsistency emerged in the first data series rather than due to the site relocation.

**Figure 5.** Temporal variations of three typical inhomogeneous PM$_{2.5}$ concentration records during 2015–2019. (Top) Significant deviations during a short time period, (middle) long-term chronic drifts with statistically significant varying trend detected in the residual series, (bottom) discontinuity due to
site relocation. The left panel compares the base series with the reference and the neighboring series used to compose the reference while the right panel shows the residual series between the base and reference series as well as their mean-shift series.

Figure 6 shows the estimated linear trends for PM$_{2.5}$ residual series that failed to pass the homogeneity test. Approximately 89% of the residual series were found exhibiting statistically significant linear trends, suggesting the vital importance to homogenize such PM$_{2.5}$ concentration records as the trend estimations at these stations could be prone to large bias if no essential adjustments are performed. Further comparisons of the percentage of data gaps between homogeneous and inhomogeneous records (Figure S2) as well as the spatial distance between the base and the reference series (Figure S3) indicate that both the frequency of data gaps and spatial distance have no obvious impact on the change point detection. In other words, the detected change points have no linkage with neither missing value frequency nor spatial distance between the base and neighboring series, suggesting a high confidence level of the identified discontinuities in these PM$_{2.5}$ concentration records.

**Figure 6.** Trend estimations for the residual PM$_{2.5}$ concentration data series that failed to pass the homogeneity test during 2015–2019. The solid circles indicate trends are statistically significant at the 95% confidence level.
Given the emergence of obvious discontinuities in more than half of the selected long-term PM$_{2.5}$ concentration records, the QM adjustment method was applied to correct the discontinuities detected in each PM$_{2.5}$ concentration record. Figure 7 shows an example of homogenization on PM$_{2.5}$ concentration data series that suffered from evident drifts from its reference (large drifts shown in Figure 5d). The inter-comparisons of PM$_{2.5}$ concentration data between the base and reference series indicate that the PM$_{2.5}$ concentration level was obviously underestimated by the raw observations compared with the reference, especially during the middle of 2016 (Figure 7a). Such evident drifts were remarkably diminished after the homogenization (Figure 7b), which shows a good agreement of the mean PM$_{2.5}$ concentration level between the homogenized datum and the reference series.

**Figure 7.** Comparison of daily mean PM$_{2.5}$ concentration before and after homogenization at one monitoring site in Guangdong province (24.69°N/113.60°E) from November 2015 to December 2016 (large drifts shown in Figure 5d).

### 4.3 Validation with independent dataset

In this study, PM$_{2.5}$ observations that were collected independently by five consulates of United States distributed in five major Chinese cities between 2015 and 2017 were used to evaluate the consistency of the derived PM$_{2.5}$ concentration records. Figure 8 shows site-specific comparisons of
daily PM$_{2.5}$ concentration between homogenized and observed data in Beijing, Shanghai, Chengdu, Shenyang, and Guangzhou, respectively. It is indicative that the homogenized daily PM$_{2.5}$ concentration data were in good agreement with PM$_{2.5}$ observations sampled at US consulates, with a correlation coefficient value of $>0.95$ and root mean square error of $<15 \ \mu g \ m^{-3}$. Given the independent measurement of PM$_{2.5}$ concentration data at US consulates, we argue that the homogenized PM$_{2.5}$ records are accurate enough in characterizing the variability of PM$_{2.5}$ loadings in China. It is also noteworthy that the homogenized PM$_{2.5}$ records are temporally complete whereas missing values are found in PM$_{2.5}$ observations sampled at US consulates.

Figure 8. Comparisons of the homogenized PM$_{2.5}$ concentration (red) against PM$_{2.5}$ observations (blue) measured at five consulates of United States in China from January 2015 to June 2017. (a~e) Temporal variations of daily PM$_{2.5}$ concentration and (f~j) the associated scatter plots.
4.4 Trend estimations from the homogenized dataset

A homogenized data record is essential to trend analysis. Figure 9 presents the annual mean concentration of PM$_{2.5}$ across China from 2015 to 2019. As shown, there is a phenomenal reduction of PM$_{2.5}$ concentration in the past five years, especially in the North China Plain as the annual mean PM$_{2.5}$ concentration decreased from more than 100 $\mu$g m$^{-3}$ in 2015 to about 60 $\mu$g m$^{-3}$ in 2019. To evaluate the benefits of data homogenization on PM$_{2.5}$ trend estimations, PM$_{2.5}$ trends estimated from both the raw observations and homogenized dataset were compared. Prior to trend analysis, each PM$_{2.5}$ concentration record was standardized in reference to its annual cycle to reduce the impacts of seasonality and spatial variations. Figure 10 shows a site-specific comparison of PM$_{2.5}$ trend estimations derived from raw observed and homogenized datasets during 2015–2019. In general, trend estimations from both datasets showed an evident decreasing tendency of PM$_{2.5}$ concentration from 2015 to 2019. However, PM$_{2.5}$ trends derived from raw observations exhibit obvious inhomogeneity over space, which is clearly evidenced by the antiphase trend estimations even at adjacent stations, especially for those with positive trends whereas all adjacent neighbors exhibited negative trends. Such antiphase trend estimations in a very small region also demonstrate the existence of obvious inhomogeneity in raw observed in situ PM$_{2.5}$ concentration dataset.

After homogenization, the phenomena of antiphase trend estimations over the local region was substantially diminished, resulting in a spatially much more homogeneous decreasing tendency of PM$_{2.5}$ concentration across China (Figure 10b). This can be also evidenced by the enlargement of national mean PM$_{2.5}$ decreasing trend estimations (increased from 7.01 to 7.25), in particular the decreased variations in trend values (uncertainty reduced from 0.25 to 0.22) and the increased number of PM$_{2.5}$ records with statistically significant varying trends (1,208 versus 1,248). These results collectively demonstrate the effectiveness of the QM adjustment method in mitigating such inhomogeneity, which also highlight the critical importance of data homogenization to account for discontinuities in this in situ PM$_{2.5}$ concentration dataset. Overall, our results indicate an obvious decreasing trend of PM$_{2.5}$ concentration in China in the past five years at a mean rate of $-7.25 \pm 0.22\%$ per year. Compared with other regions of interest (ROIs) such as Pearl River Delta (PRD, refer to...
Figure S4 for the location, PM$_{2.5}$ loading over Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), Sichuan Basin (SCB), and Central China (CC) decreased even more prominently (Table 1).

![Figure 9](https://example.com/fig9.png)

**Figure 9.** Annual mean PM$_{2.5}$ concentration derived from the homogenized daily PM$_{2.5}$ concentration dataset at 1,309 monitoring stations in China between 2015 and 2019. The North China Plain was outlined by the red rectangle in panel (f).

![Figure 10](https://example.com/fig10.png)

**Figure 10.** Linear trends for (a) raw observed and (b) homogenized daily PM$_{2.5}$ concentration data during 2015–2019. Solid circles indicate trends are statistically significant at the 95% confidence
interval. Numbers shown in the lower left of each panel indicate the overall trend derived from (top) all available stations and (bottom) the stations with significant trends at the 95% confidence interval while the numbers shown in brackets are the corresponding number of data records. Each PM$_{2.5}$ time series were standardized by its mean annual cycle during the study period to account for spatial variations of PM$_{2.5}$.

Table 1. Regional trend estimations for PM$_{2.5}$ concentration over five major ROIs in China during 2015–2019 before and after homogenization. Uncertainty in trend estimations were characterized at the 95% confidence interval. Locations of these ROIs can be found in Figure S4.

<table>
<thead>
<tr>
<th>ROI</th>
<th>Raw observation (% a$^{-1}$)</th>
<th>Homogenized record (% a$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing-Tianjin-Hebei (BTH)</td>
<td>-9.03 ± 0.78</td>
<td>-9.19 ± 0.69</td>
</tr>
<tr>
<td>Yangtze River Delta (YRD)</td>
<td>-7.07 ± 0.54</td>
<td>-7.33 ± 0.40</td>
</tr>
<tr>
<td>Central China (CC)</td>
<td>-8.47 ± 0.51</td>
<td>-8.58 ± 0.41</td>
</tr>
<tr>
<td>Sichuan Basin (SCB)</td>
<td>-7.39 ± 1.02</td>
<td>-7.84 ± 0.89</td>
</tr>
<tr>
<td>Pearl River Delta (PRD)</td>
<td>-4.30 ± 0.51</td>
<td>-4.60 ± 0.39</td>
</tr>
</tbody>
</table>

To further assess the improvement of the data quality after homogenization, the daily in situ PM$_{2.5}$ concentration records at a $1^\circ \times 1^\circ$ grid cell resolution were grouped across China. In each grid cell, the regional mean correlation coefficient among PM$_{2.5}$ concentration time series and standard deviation of PM$_{2.5}$ trends were estimated from the raw observed and homogenized daily PM$_{2.5}$ concentration time series, respectively. Their relative differences were then calculated to show the improvements of data homogeneity within each grid cell. As shown in Figure 11, the correlation among PM$_{2.5}$ concentration datum was enhanced ubiquitously after homogenization, especially in the southwest of China (e.g., Yunnan) where obvious inhomogeneity was observed in the raw PM$_{2.5}$ observations (Figure 10a). Meanwhile, the standard deviation of PM$_{2.5}$ trends within each grid cell was also substantially reduced, even by more than two folds in the magnitude (Figure 11b). These results also demonstrate the critical need to homogenize the observed PM$_{2.5}$ concentration data from a large-scale monitoring network to reduce temporal inconsistency and spatial inhomogeneity that were not even noticed before.
5 Data availability

The raw observations of \textit{in situ} PM$_{2.5}$ concentration data in China used in this study were retrieved via a web crawler from the National Urban Air Quality Real-time Publishing Platform (http://106.37.208.233:20035) between 2014 and 2019. Given the deployment of many new monitoring sites in 2014, we decided to generate a coherent PM$_{2.5}$ concentration dataset starting from 2015 to include as many records as possible. The homogenized daily \textit{in situ} PM$_{2.5}$ concentration dataset developed in this study is publicly accessible at https://doi.pangaea.de/10.1594/PANGAEA.917557 (Bai et al., 2020a).

6 Conclusions

In this study, a homogenized yet temporally complete daily \textit{in situ} PM$_{2.5}$ concentration dataset in China was generated based on the discrete hourly PM$_{2.5}$ concentration records that were retrieved from the National Urban Air Quality Real-time Publishing Platform using a web crawler during the period of 2015–2019. To create such a coherent dataset, a set of analytic methods were geared up seamlessly and applied sequentially to the retrieved raw PM$_{2.5}$ concentration records, involving quality assessment.
control, gap filling, data merging, change point detection, and bias correction. This new dataset would help scientific community better elucidate the temporal and spatial variability of haze pollution in China in the recent years, which is expected to improve the understanding of underlying causes.

The raw PM$_{2.5}$ concentration records were found to be suffering from phenomenal inhomogeneity caused by data consistency and temporal coverage as well as the relocation and repeal of a bunch of monitoring stations. It indicated that more than half of the long-term PM$_{2.5}$ concentration records failed to pass the homogeneity test, given the presence of significant change points. Further investigation confirms that large yet short-term mean shifts and chronic drifts are two primary reasons for the detected discontinuities.

Based on the homogenized dataset, the long-term trends of PM$_{2.5}$ concentration in China were estimated. In contrast to the inhomogeneous trend estimations that were derived from raw PM$_{2.5}$ concentration records, the homogenized dataset yielded a spatially much more homogeneous decreasing tendency of PM$_{2.5}$ across China at a mean rate of about $-7.3\%$ per year. Such an improvement of homogeneity was also evidenced by the enhanced correlation and reduced standard deviation of trend estimations between homogenized PM$_{2.5}$ concentration time series in the surroundings. These results clearly demonstrate the benefit of data homogenization on the improvement of the quality of this PM$_{2.5}$ concentration dataset as evident discontinuities have been removed after homogenization. Overall, our work clearly reveals the presence of evident discontinuities in the *in situ* PM$_{2.5}$ concentration records measured in China, and the homogenization actions are imperative to take in order to attain a long-term coherent PM$_{2.5}$ concentration dataset that can be used to advance PM$_{2.5}$ pollution related policy making and public health risk assessment.

**Author contributions**

The study was completed with cooperation between all authors. JG and KB conceived of the idea behind generating homogenous PM$_{2.5}$ dataset across China; KB and KL conducted the data analyses and KB wrote the manuscript; All authors discussed the experimental results and helped reviewing the manuscript.
Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (grant 41701413) and the National Key Research and Development Program of China (grant 2017YFC1501401). The authors are grateful to China National Environmental Monitoring Center (http://www.cnemc.cn/en/) and the embassy of United States in China (http://www.stateair.net/) for releasing the sampled air quality data publicly online. We also want to express our sincere thanks to Dr. Yang Feng in the Expert Team on Climate Change Detection and Indices (ETCCDI) (http://etccdi.pacificclimate.org/software.shtml) for providing the RHtestsV4 software package.
References


Bai, K., Li, K., Chang, N.-B., Gao, W., 2019b. Advancing the prediction accuracy of satellite-based PM$_{2.5}$ concentration mapping: A perspective of data mining through in situ PM$_{2.5}$ measurements. Environ. Pollut. 254, 113047. https://doi.org/10.1016/j.envpol.2019.113047


Bai, K., Ma, M., Chang, N.-B., Gao, W., 2019c. Spatiotemporal trend analysis for fine particulate matter concentrations in China using high-resolution satellite-derived and ground-measured PM$_{2.5}$ data. J. Environ. Manage. 233, 530–542. https://doi.org/10.1016/j.jenvman.2018.12.071


Huang, X., Wang, Z., Ding, A., 2018. Impact of aerosol-PBL interaction on haze pollution: multiyear...
https://doi.org/10.1029/2018GL079239

https://doi.org/10.1175/2009BAMS2736.1


https://doi.org/10.1093/nsr/nwx117

https://doi.org/10.4209/aaqr.2017.11.0437


https://doi.org/10.3390/ijerph15071487


Nie, H., Qin, T., Yang, H., Chen, J., He, S., Lv, Z., Shen, Z., 2019. Trend analysis of temperature
and precipitation extremes during winter wheat growth period in the major winter wheat planting area of China. Atmosphere. 10, 240. https://doi.org/10.3390/atmos10050240


