$PM_{2.5}$ concentrations based on near-surface visibility in the Northern Hemisphere from 1959 to 2022

We thank the referees for the constructive and helpful comments. We have carefully thought about the comments, made corresponding revisions to the manuscript and the datasets, and checked the manuscript carefully, which have substantially improved the manuscript and the datasets.

Main modifications:

- Collected more $PM_{2.5}$ concentrations data (371 sites with more than 3-year observations) from openAQ in the Northern Hemisphere in Section 2.2.6, increasing the coverage in the NH.
- Used visibility data from ISD instead of the original visibility data in Section 2.3, which resulting in more than 1000 stations added than previous version. Based on ISD visibility, the distances decrease significantly. And the upper limit is set to 100 km.
- Added the comparisons on the daily/monthly scale and before/after 2010 in Section 4.1, to evaluate the predictive ability of the model and the consistency of estimated $PM_{2.5}$ concentration.
- Used GAMM to analyze the interannual trends and spatial patterns on the regional scale due to irregular site distribution in Section 5.
- Adjusted the structure and content of the manuscript. And all figures and tables have been modified or replaced.

Response to Anonymous Referee #2

Hao et al. utilized a machine learning method to estimate a long-term global $PM_{2.5}$ dataset based on visibility data at a site scale. Comprehensive validation and analysis have confirmed the reliability and value of this dataset. However, there are some major issues that must be addressed before considering the manuscript for publication. The specific comments are as follows.

Comment 2.1. *L23-31: The representativeness of spatially distributed sparse station monitoring data for average concentrations on a national scale needs careful consideration. In China, $PM_{2.5}$ monitoring stations are predominantly located in urban areas, where concentrations tend to be higher than in rural areas. Additionally, the methodology for calculating trends warrants clarification. Calculating regional trends across these locations is challenging due to the uneven distribution of monitoring sites. Chang et al. (2017) noted that the European network is more sparsely populated across its northern and eastern regions and therefore a simple average of the individual trends at each site does not yield an accurate regional trend. More robust conclusions could be drawn when estimating the spatiotemporal full-coverage dataset. Reference: Kai-Lan Chang, Irina Petropavlovskikh, Owen R. Cooper, Martin G. Schultz, Tao Wang; Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia. Elementa: Science of the Anthropocene 1 January 2017; 5 50. doi: https://doi.org/10.1525/elementa.243*

- **Response 2.1:**

Thank you for your suggestion. We have used GAMM (Chang et al., 2017) to analyze the regional trends and spatial patterns in Section 5.

Comment 2.2. *L39-141: The content is repeated in the caption of Figure 1.*

- **Response 2.2:**

Thank you for your correction. We have made modifications.

Comment 2.3. L197: Does "2000" in sites as of 2000 refer to 2022 or 2020? Figure 1 indicates the sites in China have existed for only about ten years.

- **Response 2.3:**

Thank you for your correction. We have made modifications.

Comment 2.4. L332: Please provide the full name of the abbreviation "CART".

- **Response 2.4:**

Thank you for your correction. We have made modifications.

Comment 2.5. *How are $PM_{2.5}$, visibility and meteorological data matched spatially, and what is the distance between $PM_{2.5}$ and meteorological monitoring stations? Are there multiple $PM_{2.5}$ sites that match the same meteorological and visibility stations, thereby providing the same features and different labels for the samples of these sites? This scenario is counterfactual.*

- **Response 2.5:**

We have added details on the spatiotemporal matching between visibility station and $PM_{2.5}$ site in Section 2.4.

Comment 2.6. *The verification method for the machine learning model may not be convincing, even if the cross-validation based on samples was used. Given the study aims to establish a long-term $PM_{2.5}$ dataset, especially for historical periods lacking surface monitoring, the temporal generalization performance of the model is crucial. It is necessary to evaluate the performance based on data from the period not included in the training dataset. For instance, the model could be trained on data from before 2020 and tested on data from after 2020.*

- **Response 2.6:**

We sort the sample data by time, the first 80% of sample data is the training set, and the last 20% is the test set, which has been stated in section 2.6.

Comment 2.7. *L615: "Elevation of Meteorological Station" should be corrected to "Elevation of Visibility Station" in Figure 9. The same problem occurs in Figure 10.*

- **Response 2.7:**

Thank you for your correction. We have made modifications.

Comment 2.8. *L805: There is no section 2.6.3, please check the full text.*

- **Response 2.8:**

Thank you for your correction. We have made modifications.