https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

Deep-Learning Spatial Principles from Deterministic Chemical Transport Model for Chemical Reanalysis: An Application in China for $PM_{2.5}$

Baolei Lyu[1], Ran Huang[2], Xinlu Wang[2], Weiguo Wang[3], Yongtao Hu[4]

[1]Huayun Sounding Meteorological Technology Co. Ltd., Beijing 100081, P. R. China
[2]Hangzhou AiMa Technologies, Hangzhou, Zhejiang 311121, P. R. China
[3]I.M. System Group, Environment Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, Maryland 20740, United States
[4]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

Correspondence to: Baolei Lyu (baoleilv@foxmail.com), Ran Huang (ranhuang2019@163.com)

Abstract. Well-estimated air pollutant concentration fields through data fusion are critically important to compensate the observations that are only sparsely available, especially over non-urban areas. Previous data fusion methods generally used statistical models to relate target observations and supporting data variables at known stations. In this study, we built a new data fusion paradigm by designing a dedicated deep learning framework to learn multi-variable spatial correlations from Chemical Transport Model (CTM) simulations, before using it to estimate $PM_{2.5}$ reanalysis fields from station observations. The model was composed of two modules, which include an explainable PointConv operation to pre-process isolated observations and a regression grid-to-grid network to reflect correlations among multiple variables. The model was evaluated in two aspects of reproducing $PM_{2.5}$ CTM simulations and generating reanalysis/fused $PM_{2.5}$ fields. First, the fusion model was able to well reproduce CTM simulations from sampled station CTM data items with an average R^2 = 0.94. Second, the fusion model achieved good performance with R^2=0.77 and R^2=0.83 respectively evaluated at the stringent city-level and station-level. The generated reanalysis $PM_{2.5}$ fields have complete spatial coverage within the modelling domain and at daily time scale. One significant benefit of our fusion framework is that the model training does not rely on observations, which can be used to predict $PM_{2.5}$ fields in newly-setup observation networks such as those using portable sensors. The fusion model has high computing efficiency (<1s/day) in predicting $PM_{2.5}$ concentrations due to acceleration using GPU. As an alternative to generate chemical/meteorological reanalysis fields, the method can be readily applied for other simulated variables that with measurements available.

1 Introduction

Pollutant concentration fields with high accuracy are important for evaluating health effects, climate changes and agricultural studies (Bell et al., 2007; Donkelaar et al., 2015; Gao et al., 2017). Long-term and reliable air quality dataset could also be used to assess pollutant emission control measures (Wang et al., 2010). Data fusion method has been widely used to obtain accurate and spatially complete datasets, such as fusing air quality model simulations and station air pollutant observations to estimate fine-scale air pollutant concentration fields (Berrocal et al., 2012; Rundel et al., 2015).

In previous studies, there exists a general paradigm to develop well-estimated air pollutant concentration fields. In this paradigm, complex statistical models were trained to depict non-linear relationships between observations and proxy data and other supporting variables at the locations of observation sites (Berrocal et al., 2012; Lyu et al., 2019; Chu et al., 2016). The widely used proxy data are Aerosol Optical Depth (Lv et al., 2016), chemical transport model (CTM) simulations (Lyu et al., 2019) and other geophysical variables. Popular statistical models include machine learning models of linear mixed effect model

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

(Hao et al., 2015), random forest (Brokamp et al., 2018; Huang et al., 2021), deep neural networks (Qi et al., 2018), and ensembled models (Xiao et al., 2018). The fitted model was then used to predict concentration field of target variables in the whole area directly or through other spatial spreading techniques such as Bayesian estimation (Xu et al., 2016), partial linear regression (Wang et al., 2016) and distance-constrained interpolations (Chang et al., 2014; Friberg et al., 2016).

Even though many high-quality datasets have been developed through deliberately designed statistical models and abundant explanatory variables, there are scientific gaps following this paradigm to develop air pollutant fields. First, these models usually rely on long-term and large-scale station observations for training, especially those complex time and space resolved models (Feng et al., 2020; Huang et al., 2021). For newly setup or temporally mobile observation networks, there would be limited datasets for training an effective model. Second, most of the previous methods cannot well fuse multi-variable observations from different monitoring networks. For example, stations in air quality and meteorology observation networks are usually not spatially aligned. The observations in two networks could not be well directly fused in current models. Instead, meteorology reanalysis data were often used as important explanatory variables in previous fusion model (Geng et al., 2015; Ma et al., 2015; Wei et al., 2021). However, in real-time operational data fusion applications, these reanalysis data would be unavailable or requiring intensive computations. Last but not the least, for most of the previous methods that fusing CTM simulations, they rely on relatively highly accurate and stable simulations to achieve good fusion performance (Tong and Mauzerall, 2006). Consistency in CTM parameters, configurations and inputs are also strictly required to achieve good data fusing performance. Especially in near-real-time operational data fusion applications, adjoint models are often required to be running simultaneously (Friberg et al., 2016).

To address these scientific gaps, this study developed a new deep-learning-based model framework to estimate reanalysis from station observations by learning spatio-temporal correlations from deterministic models. Distinct from the existing data fusion models, we do not use CTM simulations in regression directly as proxy of the real pollutant concentrations. Instead, the deep learning network was trained with only CTM model simulations to learn the dynamic correlations, which is backed by the CTM's first principals, between the simulations at the randomly selected grid points that mimic monitoring locations and at the whole grid cells. The data fusion/reanalysis is then achieved by applying the learned dynamic correlations with real observational data in the prediction procedure. The model framework is fundamentally an alternative of generating chemical/meteorological reanalysis fields but without rerunning CTMs with data assimilation.

2 Data and Methods

2.1 CTM Simulations

In this study, our data fusion model was trained to learn spatial correlations of multiple variables from CTM simulations. The simulated $PM_{2.5}$ and other meteorological variables in 2016~2020 were produced using a modeling system that consists of three major components: The meteorology component (WRFv3.4.1) provides meteorological fields, the emission component provides gridded estimates of hourly emissions rates of primary pollutants that matched to model species, and the CTM component (CMAQ v5.0.2(Byun and Schere, 2006)) solves the governing physical and chemical equations to obtain 3-D pollutant concentrations fields at a horizontal resolution of 12 km. We used the simulated daily mean surface layer predictions of $PM_{2.5}$ concentrations, RH, and WS. The data covered the whole China with a size of 372×426 grid cells. Simulation data covering the 2016~2019 period was used as the training dataset, while the 2020 simulation data was used for evaluation.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

2.2 Ground Observations

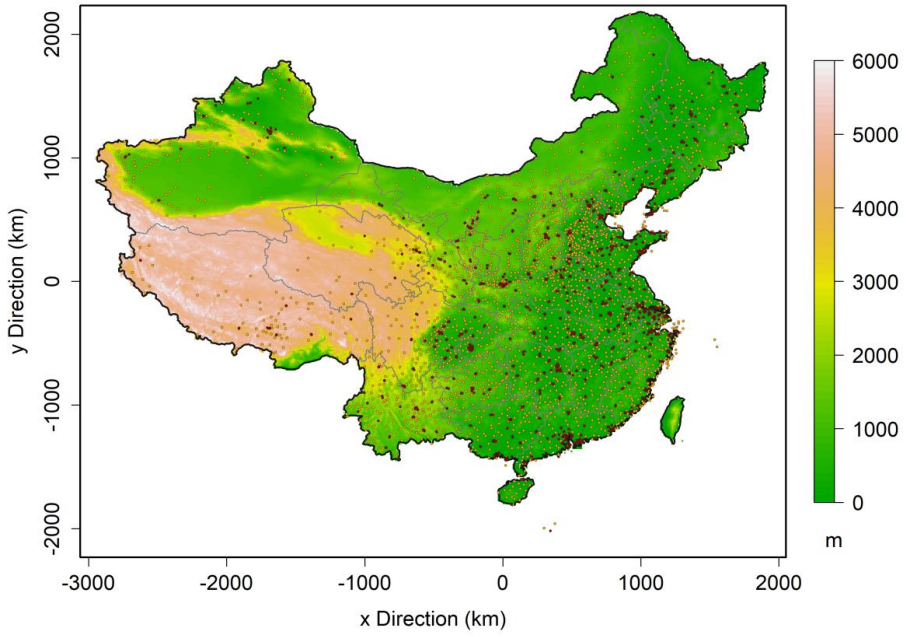


Figure 1: The map of the study area with elevation in color. Dark red dots represent the national $PM_{2.5}$ monitors and orange dots refer to national meteorological stations.

The air quality and meteorology observations were only used in predicting fused data fields. $PM_{2.5}$ observations in 2020 from the China National Environmental Monitoring Center (CNEMC) (http://106.37.208.233:20035/) were used, with the monitoring network as exhibited in Figure 1. Meteorological variables of daily mean relative humidity (RH) and wind speed (WS) for the same period at national meteorological observing stations were obtained from the China Meteorology Agency (CMA) network (Figure 1). The raw data of both $PM_{2.5}$ and meteorology data were hourly, which were averaged to daily mean if there are more than 18 valid hourly observations in a day at the local time at each monitor. Each of these data items at each were assigned to a grid that was defined same as used in the aforementioned CTM simulations. For the sites that co-located in a same grid cell, their averages were also used. It should be noted that those grid cells, which do not have valid observations within them, were filled with zero.

Geographical variables such as the surface height of Digital Elevation Model (DEM), land use and land cover (LULC) (Zhang et al., 2020) were also used in this study for fusion. These data variables were also resampled to the afore-mentioned grid.

2.3 Deep Learning Data Fusion Framework

The objective of obtaining spatially complete air pollutant field from point observations can be regarded as a downscaling problem, which indicates that values in gap areas among stations need to be optimally estimated from known sparse measurements based on physical or statistical constrains. Most previous studies use statistical methods to relate observations with other supporting variables at stations (Di et al., 2016; Beloconi et al., 2016). In this study, we built a point-to-grid model by learning from CTM simulations to generate gridded data fusion fields from station observations.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

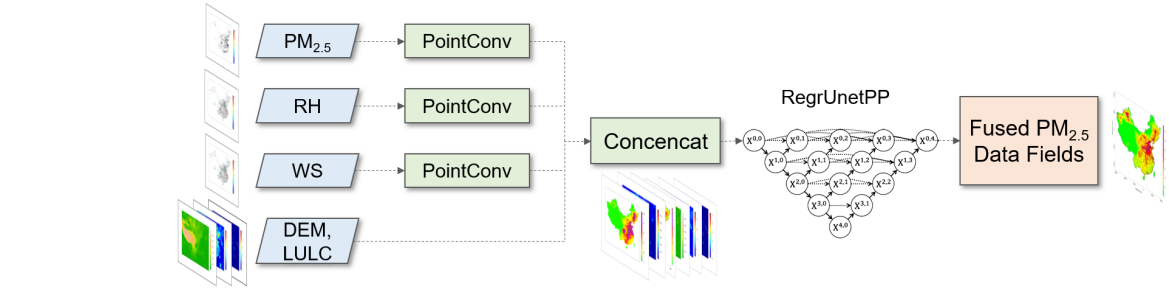


Figure 2: Data fusion framework using station observations of multiple variables to obtain gridded fields of $PM_{2.5}$.

A new deep learning model framework (Figure 2) was designed to fulfill the task of point-to-grid data fusion and downscaling. This model includes two successive point convolutional (PointConv) operations and a deep learning backbone fusion module. The PointConv is designed for handling spatially isolated and irregular station observations. In traditional convolutional operations, the 3×3 moving sum kernels were often used, which would lose effectiveness when it handles station observations. For example, when convolutional kernels coincide with grid cells without observations, the result will be zero. However, if the kernels coincide with grid cells with dense observations, the results will become significantly larger (Qi et al., 2018). To solve the problem, we proposed a novel and interpretable operation PointConv to handle isolated station observations of multiple variables. The successive PointConv operation is defined as follows,

$$PointConv_1 = \frac{Conv(w_{n1},x)}{Conv(w_{n1},x\_one)+ e^{-5}} \quad (1)$$

$$PointConv_2 = \frac{Conv(w_{n2},PointConv_1-x)}{Conv(w_{n2},x\_one)+ e^{-5}} \quad (2)$$

$$PC = PointConv_1 + PointConv_2 \quad (3)$$

Where w_n refers to a convolutional kernel with a size n. The $Conv(w_n, x)$ in Eq. (1) refers to the traditional convolution on x, which is station observations assigned to pre-defined grid cells. The x_one was binarized from x by replacing grid cells with valid observation data in x as 1. The PointConv was conducted for the second time by mimicking successive analysis procedures as in Eq. (2). The PointConv kernel size in the two steps was determined to be 21 and 11 respectively for n_1 and n_2. This model framework has the following features and advantages compared to conventional convolutions.

1) The weighted average of isolated data is implemented rather than weighted sum,
2) Large-size kernels are used to well reflect spatial correlations in a large area,
3) Successive PointConv operations are implemented to reflect local variations,
4) Multi-variable observations from different networks are handled simultaneously.

The PointConv kernels in well-trained models are expected to have larger values in the center area and lower values in the outer area. With the PointConv module, the spatially complete gridded data set are constructed, denoted as PC in Eq. (3). By binding results of PointConv with other static supplementary data such as DEM and LULC, input data to data fusion module RegrUnetPP is built as exhibited in Eq. (4).

$$\hat{y}_{PM_{2.5}} = RegrUnetPP\left(Concencat\left(PC_{PM_{2.5}}, PC_{RH}, PC_{WS}, DEM, LULC\right)\right) \quad (4)$$

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left|y_{PM_{2.5},i} - \hat{y}_{PM_{2.5},i}\right| \quad (5)$$

The operation $Concencat$ refers to appending different data variables as one multiple-layer data item. The $\hat{y}_{PM_{2.5}}$ refers to the estimated $PM_{2.5}$ concentrations, $y_{PM_{2.5},i}$ refers to the original CTM simulations of $PM_{2.5}$ with N equals to the number of total grid cells.

The fusion module can be any grid-to-grid deep learning model to estimate fused $PM_{2.5}$ concentrations $\hat{y}_{PM_{2.5}}$. Here we used

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021


Geoscientific Model Development Discussions

a regression Unet++ ($RegrUnetPP$) model (Eq.4) which was revised from the original Unet++ model (Zhou et al., 2018). The Unet++ model was designed as an Encoding-Decoding type network developed from Unet (Ronneberger et al., 2015). Many skip-connection modules (Yamanaka et al., 2017) were added in the Unet++ to fully explore spatial correlations in different scales while keeping abundant details in output results. RegrUnetPP was constructed by replacing the SoftMax activation layers with the ReLU layers and adopting a mean absolute error (MAE) loss function (Eq.5) instead of the original MaxEntropy function.

Model Training

The model was trained with the WRF-CMAQ simulations of $PM_{2.5}$, RH, and WS, together with geophysical covariates of DEM and LULC. In the training data, nominal point-wise 'station' data were constructed by randomly sampling 1500~2500 data points from gridded simulation data separately for each variable at each time, while raw spatially complete $PM_{2.5}$ simulation data were used as the target gridded "truth" data. The spatial correlations of CTM simulations are backed by physical and chemical principles comprehensively represented in the WRF-CMAQ model. The fusion model was trained with the WRF-CMAQ CTM simulations within China from 2016 to 2019 for 20000 iterations with a batch size of 10 when the loss function became stable running on a NVIDIA RTX GeForce 2080Ti GPU card. It should be highly noted that the observation data were not involved in the model training procedure at all. In the model prediction procedure, actual station observations will be used as input to generate fused $PM_{2.5}$ concentration fields.

Model Evaluation

In general, the evaluation was conducted for 2020 as it is independent from the training data period of 2016~2019. Specifically, the fitted fusion model was evaluated in two aspects. Firstly, its capabilities to predict the fully gridded model simulations from isolated sampled grid cells were assessed with the CTM simulation data in 2020. In this aspect, the station-wise CTM $PM_{2.5}$ simulations were constructed by sampling those grid cells with observations stations from raw gridded simulations. By feeding these point-wise simulations and supporting static variables into the fusion model, spatially completed grided data are obtained. The fused simulation data are then compared against the corresponding raw CTM $PM_{2.5}$ simulations. The comparison was performed in each day, since there are sufficient data items in daily simulations. It should be noted that only those grid cells located in mainland China area were compared. Statistical metrics of coefficient of determinant (R^2), root mean square error (RMSE) and normalized mean absolute error (NME) were calculated for performance evaluation.

For the second aspects, data fusion model performance was evaluated with station observations using two cross validation methods. Specifically, Leave-Stations-Out cross-validation methods (LSCV)(Lv et al., 2016) and stringent ten-fold Leave-Cities-Out cross-validation (LCCV) were used. In the LCCV method, all cities with $PM_{2.5}$ stations were randomly split into ten groups, while in the LSCV method all stations were randomly split into ten groups. $PM_{2.5}$ observations in one group of stations were used as independent evaluation data, while the data in remaining nine groups were used for data fusion. This process was iteratively performed ten times. Considering that the air quality stations are mostly clustered in cities' urban area, the LCCV method will better reflect the model's performance in predicting $PM_{2.5}$ concentrations in the remote rural areas than the station-based LSCV method. Statistical metrics of R^2, RMSE, and NME are also used for statistical measures.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

3 Results and Discussions

3.1 Model Parameters

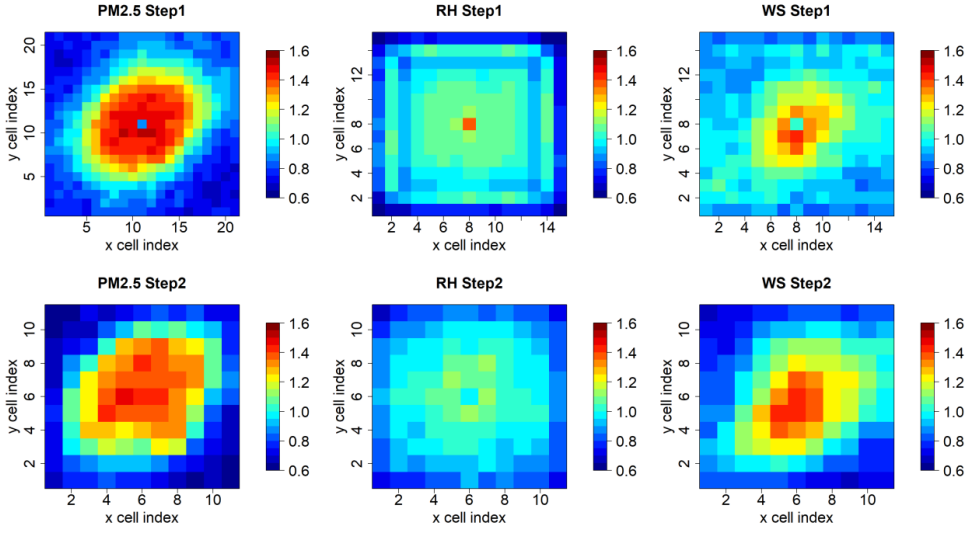


Figure 3: PointConv kernels for $PM_{2.5}$, RH and WS.

The PointConv was interpretable due to its dedicated design to implement a process like an interpolation from station observations to remove imbalanced sparsity and clustering effects of station data items. In other words, the kernel resembles to covariance function on distances. Larger values of PointConv kernels in the central area for different steps (Figure 3), indicate spatial correlations are stronger in neighboring stations (Shepard, 1968). The PointConv kernels values in the central area are respectively around 1.5, 1.1, and 1.4 for $PM_{2.5}$, RH, and WS in both steps. The kernels' distribution also revealed that the influencing distance for $PM_{2.5}$ is around 6 grid cells, which was equivalent to 72 kilometers in terms of the 12 km resolution. For RH, the spatial correlations are weak considering that kernels were more spatially uniform as exhibited in Figure 3. For wind speed, it exhibited a stronger locality indicated by the smaller hot spot with a radius around 4 grid cells (~48 kilometers). The kernels were generally isotropic with slightly larger values in the northeast-southwest direction than in other directions, which could be caused by topographic and climatic patterns in China.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

3.2 Model Performance for Reproducing Simulation Fields

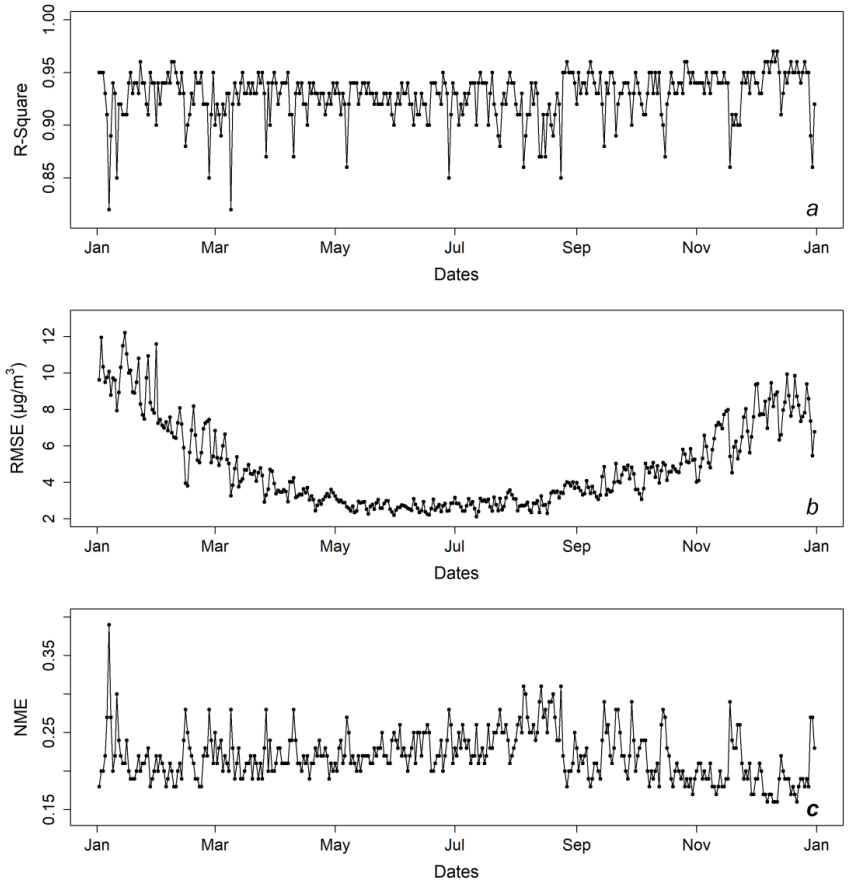
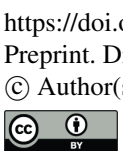


Figure 4: Daily a) R^2, b) RMSE and c) NME values evaluated between prediction CTM $PM_{2.5}$ simulations and raw gridded simulations.

The data fusion model has very high accuracies in predicting/reproducing fully gridded CTM $PM_{2.5}$ simulations as exhibited in Figure 4, even though data items in only ~800 grid cells in which with observation stations located were used to estimate data values in the nation-wide 158472 grid cells. The average of daily R^2, RMSE and NME values were respectively 0.94, 4.85 μg/m^3 and 0.22 in 2020. The good evaluation metrics indicate the high accuracies of the trained deep learning data fusion model reproducing the spatial correlations of multiple variables in the CTM model. The model has stable performance in terms of R^2 and NME values. There are occasional days where R^2 values are at low levels of ~0.85. In these days, $PM_{2.5}$ pollutions patterns generally changes fast (Figure S2 in the SI), which were generally less trained compared to those days with stable pollution patterns.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021



Geoscientific Model Development Discussions

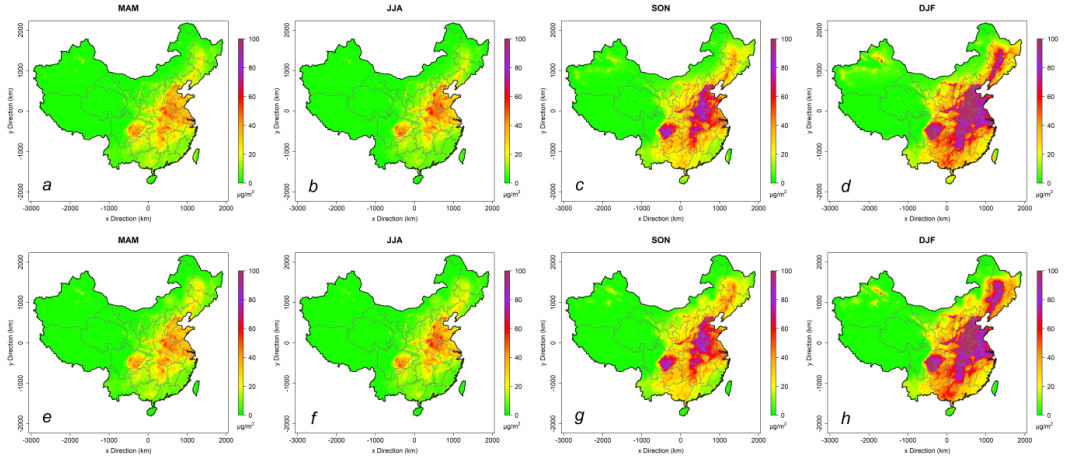


Figure 5: The seasonal average $PM_{2.5}$ concentrations of the raw CTM simulations (*a* to *d*, first row), and the reproduced simulations using data fusion models (*e* to *h*, second row).

By comparing the raw CTM simulations of daily average $PM_{2.5}$ and the reproduced$PM_{2.5}$ fields from station-wise simulations, it can be easily concluded that they exhibited high correlations and similarities as shown in Figure 5. The fusion model fully reproduced the raw CTM simulations in terms of concentration levels, spatial patterns, and fine-scale hot spots, indicting the data fusion model's capabilities to encode high-level and detailed spatial correlations. By giving the fusion model only a small portion of simulations that at sparsely scattered points, it can reproduce the entire whole domain simulation dataset accurately.

3.3 Model Performance for Generating Reanalysis Fields

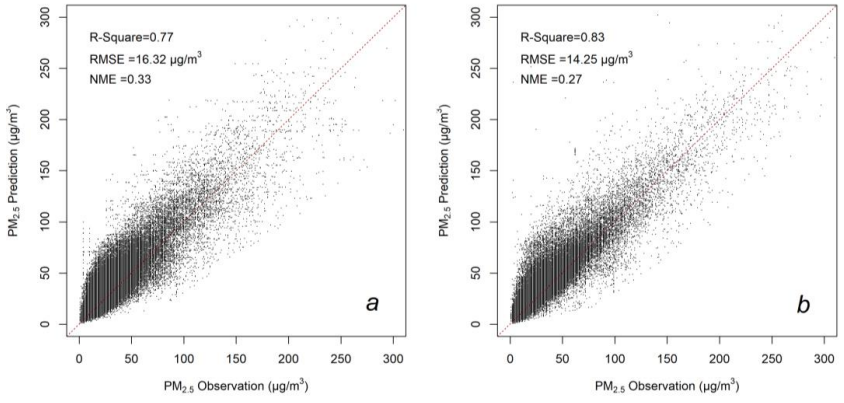


Figure 6: Scatter plots of predictions versus observations evaluated respectively by the method of a) LCCV and b) LSCV.

We implemented the data fusion model with the observations in 2020 to generate the fused fields for evaluation. The evaluation results exhibited good performance with R^2=0.77 for the LCCV method and R^2=0.83 for the LSCV method (Figure 6), with the RMSE values respectively 16.32 and 14.25 μg/m^3. Considering that most grid cells were located within city urban areas, actual model performance should be in between the metrics evaluated by LCCV and LSCV, which is 0.77~0.83 for R^2, 14.25~16.32 μg/m^3 for RMSE and 0.27~0.32 for NME. Previous studies tend to underestimate $PM_{2.5}$ concentrations in high pollution scenarios (Di et al., 2016; Senthilkumar et al., 2019). Our data fusion method predicted high level $PM_{2.5}$ concentrations very well, with NME for $PM_{2.5}$ concentration higher than 150 μg/m^3 being small of 0.19 and 0.14 respectively

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

for LCCV and LSCV. It worth noting that there exist increased errors from reproducing CTM simulations (R^2=0.93) to generating reanalysis fields (R^2=0.77~0.83). The difference of 0.1~0.16 should be mainly attributed to CTM simulation uncertainties of $PM_{2.5}$ spatial correlations compared to actual observed correlations.

Our model has good performance comparing to the previous studies that used the spatial cross-validation method. For example, Lyu et al. (2019) used an ensemble deep learning model to build relations between CTM simulations and observations of $PM_{2.5}$ in China with a performance of R^2=0.64 and RMSE=24.8 μg/m^3 using a station-level evaluation method (2019). Xue et al. (2020) fused AOD, CTM simulations and ground observations with a complex multi-stage model and achieved a good performance of R^2=0.81 with LSCV method (2017). Xiao et al. (2018) built up an ensemble machine learning model to predict $PM_{2.5}$ at 0.1° resolution with an accuracy of R^2=0.76 (2018). Huang et al. (2021) used a multi-stage random-forecast-based model to predict very high-resolution data set and achieved R^2 = 0.92 with the LSCV method (2021).

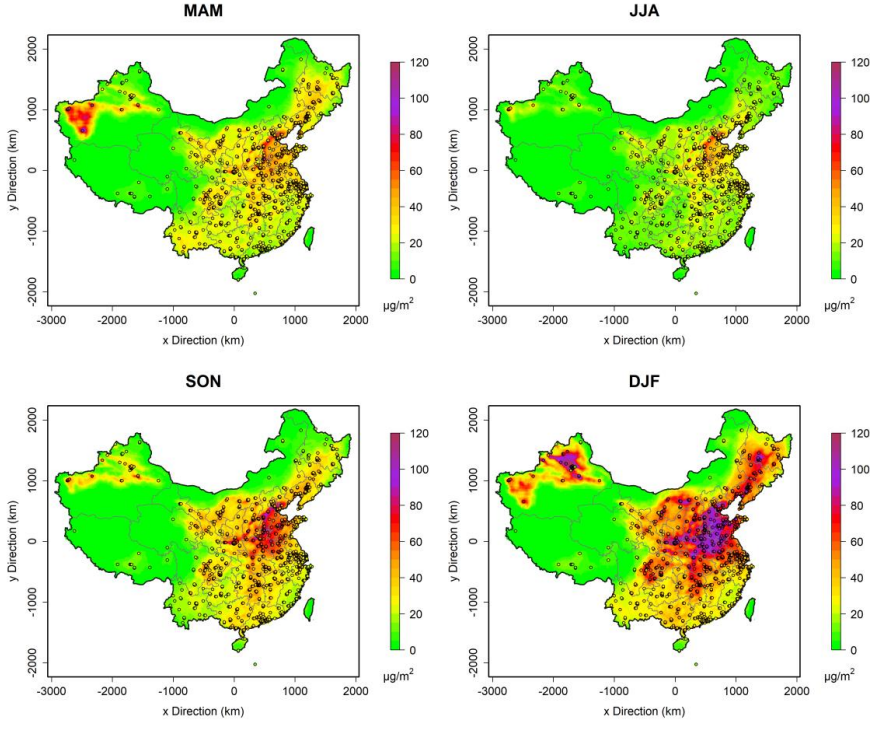


Figure 7: Reanalysis $PM_{2.5}$ concentration fields in four seasons in 2020. Circles with filled colors represent monitoring sites and corresponding observations.

Daily fused $PM_{2.5}$ fields for 2020 were obtained with the model framework. Considering the fused/reanalysis fields are complete in space, the high pollution levels in winter are well revealed in details in the North China Plain (NCP), and in the long-narrow basin areas of Shanxi and Shaanxi provinces (Figure 7). Besides, unlike most previous studies (Huang et al., 2021), $PM_{2.5}$ concentrations in high-altitude clean Tibetan Plateau region are predicted low, same as observed. High pollution levels in the middle-west Inner Mongolia area of Hohhot, Baotou, and Yinchuan were also well captured in four seasons.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Geoscientific Model Development Discussions

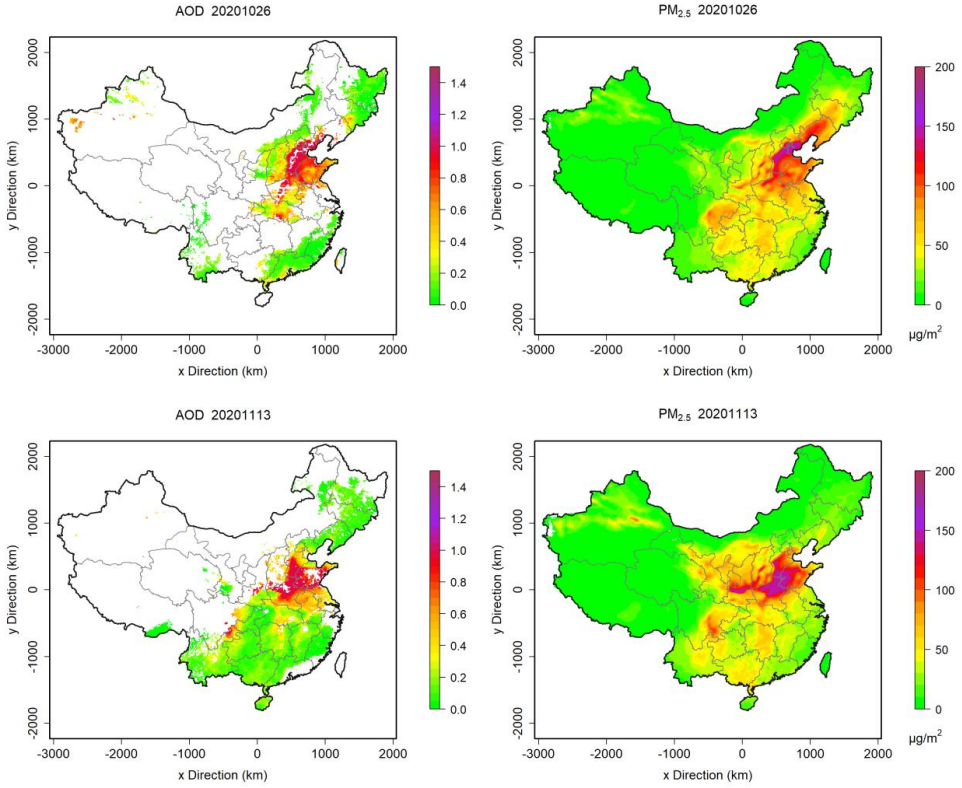


Figure 8: Comparison between MODIS AOD and $PM_{2.5}$ fusion data on October 26 and November 13 in 2020.

To further evaluate the spatial distributions of the fused $PM_{2.5}$ fields, we compare them with the MODIS AOD distributions at the days with large spatial coverages (Figure 8 and Figure S1 in the supplementary material). Spatial distributions of the fused $PM_{2.5}$ and AOD show high similarity to each other. For example, on October 26, high $PM_{2.5}$ concentrations coincide with high AOD values in NCP, especially in the areas along the east edge of the Taihang Mountains. In detail, $PM_{2.5}$ concentrations and AOD values are both relatively low in the Yimeng Mountains that located in the middle south of Shandong province. The high $PM_{2.5}$ concentrations in the basin area of Shanxi province are also higher than the surrounding area, consistent with that of AOD. On November 13, $PM_{2.5}$ concentrations are extremely high in NCP and high in central China areas of Hubei and Hunan provinces, same as that of AOD values. Besides, in the northeast regions of Yunnan province, both $PM_{2.5}$ and AOD values are high. The spatial coincidence of high $PM_{2.5}$ concentrations and AOD values supported the high accuracy of the fused data.

4 Discussions

In this study, $PM_{2.5}$ fields are fused from multiple observational variables using a novel deep learning data fusion model framework. As we have demonstrated, the method can accurately reproduce the whole domain CTM simulations from a small portion of simulations selected at sparse locations. By fully utilizing such learned spatial correlations that simulated by the CTM models, it can accurately generate spatio-temporally complete fused fields from using observation at sparse locations as well.

In previous studies, all variables need to be spatially paired at stations first to train regression models (Lyu et al., 2019; Ma et al., 2016; Xue et al., 2019). To use data from different networks, interpolation and analysis/reanalysis need to be carried out

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021


Geoscientific Model Development Discussions

which is disconnected from the data fusion model. Here with the two PointConv modules, it can fuse station data variables from different observation networks, even when they are not spatially aligned at collocations. The successive PointConv modules were able to process station data for each variable independently before the fusion. The PointConv modules were trainable as part of the whole deep learning data fusion model. Without data pairing procedure, the model training and prediction procedure became straightforward that only requires a same spatial grid setting for all input variables.

This model was fitted with model simulation data by learning daily spatial patterns from long-term CTM simulations. It has two benefits. First, the trained deep learning data fusion model can represent and reflect spatial correlations between $PM_{2.5}$ (and any other model species/variables as well) and its supporting variables, by retaining physical and chemical principles in the WRF-CMAQ model. Hence, the method can be readily applied for other CTM simulated species that with measurements available. Second, it doesn't need any observation data sets to train the model. This is quite beneficial for data fusion applications, especially when station networks are newly set up or observations are from mobile or portable sensors. The data fusion models used in previous studies are very complex requiring long-term observations (Wei et al., 2021; Huang et al., 2021; Xue et al., 2019), which make it difficult to be reproducibly used in new applications. Conversely, our method is straightforward to use and can be easily examined by inter-comparison with other methods. It provides a pre-trained deep learning model for its application in other studies. To run this model, users only need air quality observations, meteorological observations, and static variables.

Considering that the model training process is to fully learn the general spatial correlations among different variables, CTM simulations theoretically do not need to be very accurate in model inputs but need to be consistent in model configurations of physical and chemical processes. For example, input changes of pollutant emissions and meteorological fields within the CTM simulations are allowed and should be encouraged to cover a wider range of emissions and meteorological scenarios for better training. Comparatively in other observation-simulation regression methods, both model configurations and inputs are usually required to be unchanged in training data sets (Xue et al., 2017; Hao et al., 2015). In fact, in our method, larger variations of input emissions and meteorological conditions can be even more beneficial to help improve the robustness of the trained model in applications with the dramatical change of emissions or extreme climate conditions. However, substantial changes of atmospheric physical and chemical principles in models can deteriorate the model performance because it modifies the simulated correlations between different variables. The fused output data sets have high accuracy while following the known spatio-temporal principles represented in the state-of-the-art air quality models, which will be beneficial for further studies of downscaling, nowcasting and model forecast post-processing based on fused data set.

The model framework in this study has very high computational efficiency, with computing time for one-time fusion far less than 1*s* running on a consumer GPU card of NVIDIA 2080Ti. The calculation time will not increase much by enabling processing near-real-time data in a large area.

Data and code availability. The CTM simulation data and fused datasets can be accessed by contacting the corresponding authors Baolei Lyu (baoleilv@foxmail.com) and Ran Huang (ranhuang2019@163.com). The land use and land cover data are available at Data Sharing and Service Portal of Chinese Academy of Science (http://data.casearth.cn/en/sdo/detail/5ebe2a9908415d14083a4c24). The source code and a pre-trained model file of the exact version used to produce the results used in this paper is available at https://doi.org/10.5281/zenodo.5152567 on Zenodo (Lyu, 2021). The configuration files for running models of WRF v3.4.1 and CAMQ v5.0.2 are also available at https://doi.org/10.5281/zenodo.5152621 (Hu, 2021).

Author contributions. BL and YH conceived the study. BL developed the model and codes. RH and XW contributed the CTM simulation data. BL and RH collected the observation data. BL analyzed data and wrote the paper with contributions from YH, RH, WW and XW. RH managed the project.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research has been in part supported by the National Key R&D Program of China (2018YFC0214000) and the AiMa R&D Project (R#2016-004) of Hangzhou AiMa Technologies. The findings in this research do not necessarily reflect the views of the sponsors.

References

Bell, M. L., Goldberg, R., Hogrefe, C., Kinney, P. L., Knowlton, K., Lynn, B., Rosenthal, J., Rosenzweig, C., and Patz, J. A.: Climate change, ambient ozone, and health in 50 US cities, Climatic Change, 82, 61-76, 2007.

Beloconi, A., Kamarianakis, Y., and Chrysoulakis, N.: Estimating urban PM_{10} and $PM_{2.5}$ concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data, Remote Sensing of Environment, 172, 148-164, http://dx.doi.org/10.1016/j.rse.2015.10.017, 2016.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M.: Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality, Biometrics, 68, 837-848, 2012.

Brokamp, C., Jandarov, R., Hossain, M., and Ryan, P.: Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model, Environmental Science & Technology, 2018.

Byun, D. and Schere, K. L.: Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system, Applied Mechanics Reviews, 59, 51-77, 2006.

Chang, H. H., Hu, X., and Liu, Y.: Calibrating MODIS aerosol optical depth for predicting daily $PM_{2.5}$ concentrations via statistical downscaling, Journal of Exposure Science and Environmental Epidemiology, 24, 398-404, 2014.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., and Xiang, H.: A Review on Predicting Ground $PM_{2.5}$ Concentration Using Satellite Aerosol Optical Depth, Atmosphere, 7, 129, 2016.

Di, Q., Koutrakis, P., and Schwartz, J.: A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression, Atmospheric Environment, 131, 390-399, 2016.

Donkelaar, A. V., Martin, R. V., Brauer, M., and Boys, B. L.: Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter, Environmental Health Perspectives, 123, 135-143, 2015.

Feng, L., Li, Y., Wang, Y., and Du, Q.: Estimating hourly and continuous ground-level $PM_{2.5}$ concentrations using an ensemble learning algorithm: The ST-stacking model, Atmospheric environment, 223, 117242.117241-117242.117213, 2020.

Friberg, M. D., Zhai, X. D., Holmes, H., Chang, H. H., Strickland, M., Sarnat, S. E., Tolbert, P. E., Russell, A. G., and Mulholland, J. A.: Method for Fusing Observational Data and Chemical Transport Model Simulations to Estimate Spatiotemporally-Resolved Ambient Air Pollution, Environmental Science & Technology, 2016.

Gao, M., Saide, P. E., Xin, J., Wang, Y., Liu, Z., Wang, Y., Wang, Z., Pagowski, M., Guttikunda, S. K., and Carmichael, G. R.: Estimates of Health Impacts and Radiative Forcing in Winter Haze in Eastern China through Constraints of Surface $PM_{2.5}$ Predictions, Environmental Science & Technology, 2017.

Geng, G., Zhang, Q., Martin, R. V., van Donkelaar, A., Huo, H., Che, H., Lin, J., and He, K.: Estimating long-term $PM_{2.5}$ concentrations in China using satellite-based aerosol optical depth and a chemical transport model, Remote Sensing of Environment, 166, 262-270, 2015.

Hao, H., Chang, H. H., Holmes, H. A., Mulholland, J. A., Klein, M., Darrow, L. A., and Strickland, M. J.: Air Pollution and Preterm Birth in the U.S. State of Georgia (2002-2006): Associations with Concentrations of 11 Ambient Air Pollutants

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Estimated by Combining Community Multiscale Air Quality Model (CMAQ) Simulations with Stationary Monitor Measurements, Environmental Health Perspectives, 2015.

Hu, Y.: Configurations for running 12 km resolution WRF-CMAQ simulations in China, Zenodo, https://doi.org/10.5281/zenodo.5152621, 2021.

Huang, C., Hu, J., Xue, T., Xu, H., and Wang, M.: High-Resolution Spatiotemporal Modeling for Ambient $PM_{2.5}$ Exposure Assessment in China from 2013 to 2019, Environmental Science & Technology, 55, 2152-2162, 2021.

Lv, B., Hu, Y., Chang, H. H., Russell, A. G., and Bai, Y.: Improving the Accuracy of Daily $PM_{2.5}$ Distributions Derived from the Fusion of Ground-level Measurements with Aerosol Optical Depth Observations, a Case Study in North China, Environmental Science & Technology, 50, 4752, 2016.

Lyu, B., Hu, Y., Zhang, W., Du, Y., Luo, B., Sun, X., Sun, Z., Deng, Z., Wang, X., Liu, J., Wang, X., and Russell, A. G.: Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved $PM_{2.5}$ Exposure Fields in 2014–2017, Environmental Science & Technology, 53, 7306-7315, 2019.

Lyu, B.: Chemical Reanalysis Model with Deep Learning from CTM simulations, Zenodo, https://doi.org/10.5281/zenodo.5152567, 2021.

Ma, Z., Liu, Y., Zhao, Q., Liu, M., Zhou, Y., and Bi, J.: Satellite-derived high resolution $PM_{2.5}$ concentrations in Yangtze River Delta Region of China using improved linear mixed effects model, Atmospheric Environment, 2016.

Ma, Z., Hu, X., Sayer, A. M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., and Liu, Y.: Satellite-Based Spatiotemporal Trends in $PM_{2.5}$ Concentrations: China, 2004–2013, Environmental Health Perspectives, 124, 2015.

Qi, Z., Wang, T., Song, G., Hu, W., and Li, X.: Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality, IEEE Transactions on Knowledge and Data Engineering, 30, 2285-2297, 2018.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, 2015//, 234-241,

Rundel, C. W., Schliep, E. M., Gelfand, A. E., and Holland, D. M.: A data fusion approach for spatial analysis of speciated $PM_{2.5}$ across time, Environmetrics, 26, 515–525, 2015.

Senthilkumar, N., Gilfether, M., Metcalf, F., Russell, A. G., Mulholland, J. A., and Chang, H. H.: Application of a Fusion Method for Gas and Particle Air Pollutants between Observational Data and Chemical Transport Model Simulations Over the Contiguous United States for 2005–2014, International Journal of Environmental Research and Public Health, 16, 3314, 2019.

Shepard, D.: Geography and the Properties of Surfaces. A Two-Dimensional Interpolation Function for Computer Mapping of Irregularly Spaced data, ACM, 1968.

Tong, D. Q. and Mauzerall, D. L.: Spatial variability of summertime tropospheric ozone over the continental United States: Implications of an evaluation of the CMAQ model, Atmospheric Environment, 40, 3041-3056, 2006.

Wang, L., Jang, C., Zhang, Y., Wang, K., Zhang, Q., Streets, D., Fu, J., Lei, Y., Schreifels, J., He, K., Hao, J., Lam, Y.-F., Lin, J., Meskhidze, N., Voorhees, S., Evarts, D., and Phillips, S.: Assessment of air quality benefits from national air pollution control policies in China. Part II: Evaluation of air quality predictions and air quality benefits assessment, Atmospheric Environment, 44, 3449-3457, http://dx.doi.org/10.1016/j.atmosenv.2010.05.058, 2010.

Wang, M., Sampson, P. D., Hu, J., Kleeman, M., Keller, J. P., Olives, C., Szpiro, A. A., Vedal, S., and Kaufman, J. D.: Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and $PM_{2.5}$, Environmental science & technology, 50, 5111-5118, 2016.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sensing of Environment, 252, 112136, 10.1016/j.rse.2020.112136, 2021.

https://doi.org/10.5194/gmd-2021-253
Preprint. Discussion started: 10 August 2021

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An Ensemble Machine-Learning Model To Predict Historical $PM_{2.5}$ Concentrations in China from Satellite Data, Environmental Science & Technology, 52, 13260-13269, 10.1021/acs.est.8b02917, 2018.

Xu, Y., Serre, M. L., Reyes, J., and Vizuete, W.: Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: A National Application, Environmental Science & Technology, 2016.

Xue, T., Zheng, Y., Geng, G., Zheng, B., Jiang, X., Zhang, Q., and He, K.: Fusing Observational, Satellite Remote Sensing and Air Quality Model Simulated Data to Estimate Spatiotemporal Variations of $PM_{2.5}$ Exposure in China, Remote Sensing, 221, 2017.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., and Zhang, Q.: Spatiotemporal continuous estimates of $PM_{2.5}$ concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations☆, Environment International, 123, 345-357, 2019.

Yamanaka, J., Kuwashima, S., and Kurita, T.: Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network, Neural Information Processing, Cham, 2017//, 217-225,

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., and Mi, J.: GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery, Earth System Science Data Discussion, https://doi.org/10.5194/essd-2020-182, 2020.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation, Cham, 3-11,