Environmental Research Letters

LETTER

Spatially and temporally coherent reconstruction of tropospheric NO_2 over China combining OMI and GOME-2B measurements

Qin He[1], Kai Qin[1], Jason Blake Cohen[1,2,3], Diego Loyola[4], Ding Li[1], Jincheng Shi[1] and Yong Xue[1]

[1] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, People's Republic of China
[2] School of Atmospheric Sciences, Sun Yat-Sen University, Zhuhai, People's Republic of China
[3] Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, People's Republic of China
[4] German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Oberpfaffenhofen, Germany

E-mail: qinkai@cumt.edu.cn

Keywords: reconstruction, NO2, gaps, OMI, GOME-2B, machine learning

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
26 July 2020

REVISED
24 October 2020

ACCEPTED FOR PUBLICATION
5 November 2020

PUBLISHED
15 December 2020

Abstract

Tropospheric NO_2 columns retrieved from ozone monitoring instrument (OMI) are widely used, even though there is a significant loss of spatial coverage due to multiple factors. This work introduces a framework for reconstructing gaps in the OMI NO_2 data over China by using machine learning and an adaptive weighted temporal fitting method with NO_2 measurements from Global Ozone Monitoring Experiment–2B, and surface measurements. The reconstructed NO_2 has four important characteristics. First, there is improved spatial and temporal coherence on a day-to-day basis, allowing new scientific findings to be made. Second, the amount of data doubled, with 40% more data available. Third, the results are reliable overall, with a good agreement with Multi-AXis Differential Optical Absorption Spectroscopy measurements (R: 0.75–0.85). Finally, the mean of reconstructed NO_2 vertical columns during 2015 and 2018 is consistent with the original data in the spatial distribution, while the standard deviation decreases in most places over Mainland China. This novel finding is expected to contribute to both air quality and climate studies.

1. Introduction

Nitrogen dioxide (NO_2) is an important trace gas with a short atmospheric lifetime that has impacts on troposphere ozone (Tiwari and Agrawal 2018, Kang *et al* 2020) and secondary nitrate aerosols (Wang *et al* 2011a). Nitrogen oxides ($NO_x \equiv NO + NO_2$) are produced through atmospheric combustion associated with the recent industrialization process in China, with many cities still suffering from harmful air pollution (Richter *et al* 2005, Cohen and Prinn 2011, Zhang *et al* 2012a, 2012b, Zhao *et al* 2013), although there has been a recent decrease in some parts of China over the past few years (Gu *et al* 2013, de Foy *et al* 2016, van der A *et al* 2017, Si *et al* 2019, Wang *et al* 2019, Georgoulias *et al* 2019). Furthermore, since NO_2 is known to have adverse impacts on human health and ecosystems (Atkinson *et al* 2018), it should be monitored dynamically and regularly on a large scale.

Satellite retrieval of NO_2 offers a tool for monitoring the spatial distribution and temporal variation of tropospheric NO_2 (Bovensmann *et al* 1999, Burrows *et al* 1999, Levelt *et al* 2006, Veefkind *et al* 2012, Yang *et al* 2014, Munro *et al* 2016, Kim *et al* 2020, Liu *et al* 2020b). Specifically, the ozone monitoring instrument (OMI) is the first sensor to provide daily global coverage of NO_2 at a relatively high resolution (Levelt *et al* 2018), becoming widely used in air quality applications, including estimation of ground-level NO_2 concentrations (Lamsal *et al* 2008, Lee and Koutrakis 2014, Qin *et al* 2017, 2020, Zhan *et al* 2018, de Hoogh *et al* 2019), spatiotemporal analysis of tropospheric NO_2 columns (Cui *et al* 2019, Si *et al* 2019, Wang *et al* 2019), calculation of top-down NO_x emissions (Liu *et al* 2016, 2017a, Silvern *et al* 2019, Kong *et al* 2019, Adams *et al* 2019), assimilation of observations into chemical transport models (Wang *et al* 2011b, 2020, Silver *et al* 2013, Liu *et al* 2017b). However, the OMI dataset tends to be used

for long-term (monthly, seasonal or annual) analysis because of a row anomaly issue, which sacrifices the temporal resolution and contributes to an artificially high standard deviation (Cohen 2014, Cohen *et al* 2018, Lin *et al* 2020). Another significant issue is the cloud contamination, which can bias the results by obscuring the sensor's view of the lower atmosphere, resulting in significant errors, particularly in polluted regions (Boersma *et al* 2004).

Geostatistical methods such as inverse distance weighting and kriging interpolation are conventionally applied to predict values using neighboring information. They may lead to spatial discontinuity if the gap is large, while also not taking the temporal relationship into account, making the reconstruction of atmospheric products complicated (Shen *et al* 2015). Peng *et al* (2016) extend the adaptive weighted temporal fitting (AWTF) method to reconstruct OMI total ozone columns. However, the lack of information caused by clouds makes it impossible for tropospheric NO_2 data products to meet the premise of AWTF (where 90% of the target data gaps have an overlap with valid values occurring either the previous or next day). de Hoogh *et al* (2019) impute missing OMI NO_2 columns over Switzerland using estimates from a chemical transport model, with a linear regression model explaining 68% of the variability, demonstrating the feasibility of using another NO_2 product to fill the gaps associated with OMI (which has an afternoon overpass), such as the Global Ozone Monitoring Experiment–2B (GOME-2B) (which has a morning overpass).

The relationship between tropospheric NO_2 loadings in the morning and the afternoon changes under different land use and meteorological conditions (Lin *et al* 2010, Wang *et al* 2017a, Weng *et al* 2020, Penn and Holloway 2020) and due to its complexity, ordinary linear regression cannot fit it well (Cohen and Prinn 2011). Thus, we first use a machine learning method to capture the nonlinear relationship, and second apply AWTF to further improve the data coverage. This research aims to combine GOME-2B NO_2 column data with OMI NO_2 data to overcome these known issues and use the additional benefit of a different overpass time, with the goal of producing a more robust, precise, and scientifically useful NO_2 dataset over China.

2. Data and methods

2.1. Data used in the model

OMI onboard Aura (2004–today) has a local equator crossing time at around 13:45 providing early afternoon coverage globally. OMI measures backscattered solar UV, allowing the approximation of the NO_2 vertical column with an overall spatial resolution of 13 km × 24 km at nadir (Levelt *et al* 2006, 2018). This work uses the QA4ECV NO_2 version 1.1 provided by KNMI (Boersma *et al* 2017), which features an enhanced spectral fitting method (Zara *et al* 2018) and improved air mass factor calculation (Lorente *et al* 2017, 2018) and shows a good agreement with ground-based measurements (Boersma *et al* 2018, Liu *et al* 2019) over China with a 12% total uncertainty at Xianghe (Compernolle *et al* 2020).

Several criteria are used to ensure the data quality according to the standard L3 product production process (Krotkov *et al* 2017), including: first, setting the cloud fraction <0.3, second, setting the surface albedo <0.3, and third, setting the solar zenith angle <85°. To further ensure quality, cross-track pixels affected by row anomalies are excluded. This row anomaly was first noticed in June 2007, and since 2009 has affected around one-third of the cross-track positions (http://projects.knmi.nl/omi/research/product/rowanomaly-background.php#overview).

GOME-2 on MetOp-B (2012–today) is a nadir-scanning spectrometer, which measures at around 9:30 local equator crossing time. The default ground pixel size of GOME-2B is 80 km × 40 km in the forward scan, requiring 1.5 d to cover the globe (Munro *et al* 2016). The GDP offline products of version 4.7 and 4.8 provided by DLR in the framework of the EUMETSAT AC-SAF project are used here. To be compatible with each other, all OMI and GOME-2B NO_2 tropospheric columns were binned to 0.25° × 0.25° grids using HARP, a toolkit for binning weighted polygon-shaped remotely sensed measurements, as given by (http://stcorp.github.io/harp/doc/html/index.html).

To ensure that data from two satellites are available, all products in this work are subsequently confined to the same study period, specifically from 1 January 2015 to 31 December 2018, with specific details summarized in table S1.

Hourly concentrations of ground-level NO_2 from 2015 to 2018 from the China National Environmental Monitoring Center (CNEMC) website were averaged during both the morning (0900 and 1000 Local solar time, LST = UTC + 8) and the afternoon (1300 and 1400 LST). The ordinary kriging method (Núñez-Alonso *et al* 2019) was used with month-by-month NO_2 data to produce spatially continuous ground-based NO_2 concentration maps.

Given that the air quality is closely related to meteorological conditions (Guo *et al* 2017, 2019), five instantaneous meteorological variables, including 10 m U and V wind components (u10 and v10; unit: m s^{-1}), the 2 m air temperature (t2m; unit: K), boundary layer height (blh; unit: m), and surface pressure (sp; unit: hPa), at a 0.25° × 0.25° resolution were downloaded from the ERA5 hourly atmospheric reanalysis. Daily meteorological variables were averaged from 1100 to 1300 LST data, so as to coincide with the overpass time of two satellites.

Land cover types are considered to distinguish different types and magnitudes of anthropogenic activity in different places. The specific product contains

nine land types (Cropland, Forest, Grassland, Shrubland, Wetland, Water, Tundra, Impervious Surface, Bareland) from the Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) 2017v1 dataset (Gong *et al* 2013), with a spatial resolution of 30 meters. The two types (impervious surface and cropland) were used for building the model, and the spatial scale was fractionally aggregated type-by-type to a spatial resolution of 0.25° × 0.25°.

2.2. Other data used in this work

The TROPOspheric Monitoring Instrument (TROPOMI) on Sentinel-5 Precursor (2017–today) has a significant raise in spatial resolution (3.5 km × 7 km initially, and 3.5 km × 5.5 km from August 6 2019 onward) and a similar equator crossing time of 13:30 with OMI (Veefkind *et al* 2012). The offline or reprocessing products of version 1.2.0 and 1.2.2 from 1 May 2018 to 31 December 2018 are used to make a comparison with our reconstructed data, and pixels with qa_value <0.75 are filtered to ensure the data quality.

We also used Multi-AXis Differential Optical Absorption Spectroscopy (MAX-DOAS) measurements at two suburban sites (see locations in figure S1 (available online at stacks.iop.org/ERL/15/125011/mmedia)) to evaluate the performance of reconstructed data, including one in Xianghe (116.96° E, 39.75° N) to the south of Beijing, and another at the roof of the School of Environment Science and Spatial Informatics in Xuzhou (117.14° E, 34.22° N). The instrument in Xianghe was designed at BIRA-IASB (Clémer *et al* 2010) and the data before January 2017 can be downloaded from http://uv–vis.aeronomie.be/groundbased/QA4ECV_MAXDOAS/index.php. The instrument in Xuzhou is a MAX-DOAS 2000 which was developed by the Anhui Institute of Optics and Fine Mechanics (Wang *et al* 2017b), and data are available from March 2018. We averaged all valid MAX-DOAS measurements within ±1 h of the OMI overpass time for comparison with the satellite data of the single grid cell where the site is located.

2.3. Reconstruction framework

The framework used to systematically create the reconstructed NO_2 product (see figure 1) is based on two steps. First, the OMI and GOME-2B data are combined using The eXtreme Gradient Boosting (XGBoost) approach, a gradient boosting tree machine learning algorithm (Chen and Guestrin 2016), to form the first reconstruction step (RECON-1). This step regresses by producing a prediction model in the form of an ensemble of weak predictions (Chen and Guestrin 2016, Mitchell and Frank 2017), which has been shown to be applicable in various remote sensing applications (Hengl *et al* 2017, Just *et al* 2018, Shtein *et al* 2019, Li *et al* 2020). Specifically, this approach combines information that exists in GOME-2B to provide information that was otherwise missed by the same OMI pass later that same day.

The second systematic step to create the reconstructed NO_2 product requires filling missing data using the AWTF method (Peng *et al* 2016). This makes use of the information at the target location the day before and the day after ($t-1$ and $t+1$) to predict the target at date t (RECON-2). Specifically, the regression relationship is calculated from the grid cells in the adaptive search window with valid values over the three-day window, effectively considering the spatial-temporal patterns of NO_2. The AWTF method computes RECON-2 with respect to the subset of grid cells, which have neither OMI nor GOME-2B valid data, thus further improving the spatial coverage.

NO_2 data from OMI and GOME-2B, as well as ancillary surface and meteorological data in table S1 are interpolated onto a common grid in space and time. XGBoost is first applied to combine common data, and the relationship is trained by using the features (all listed in table S1 except omi) to predict a target variable (omi). ATWR is subsequently applied to predict missing data using adjacent measurements in space and/or time of RECON-1. The principles of XGBoost and AWTF are summarized in text S1 and text S2.

Before applying the XGBoost tree-based models, in this work the OMI NO_2 column loadings are scaled using a natural log-transformation (figure S2). Using the log-transformed data, a tenfold cross validation to tune the hyperparameters for training the model was conducted to avoid errors due to random sampling. The 6 668 046 matching grid cells from 2015 to 2018 were divided each turn into a training set (90%) and a validation set (10%). The results of the model performance are analyzed using three metrics: the root mean square error (RMSE), the mean average error (MAE), and the coefficient of determination (R^2) (equations (1–3)),

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\widehat{y_i}-y_i)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|\widehat{y_i}-y_i| \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\widehat{y_i}-y_i)^2}{\sum_{i=1}^{m}(\bar{y}-y_i)^2} \tag{3}$$

where $\widehat{y_i}$ and y_i denote the model predicted vertical column and the OMI measured vertical column respectively, $\bar{y}$ denotes the mean of y_i and m denotes the number of samples.

Resulting scatterplots of the training and testing splits from 2015 to 2018 over mainland China

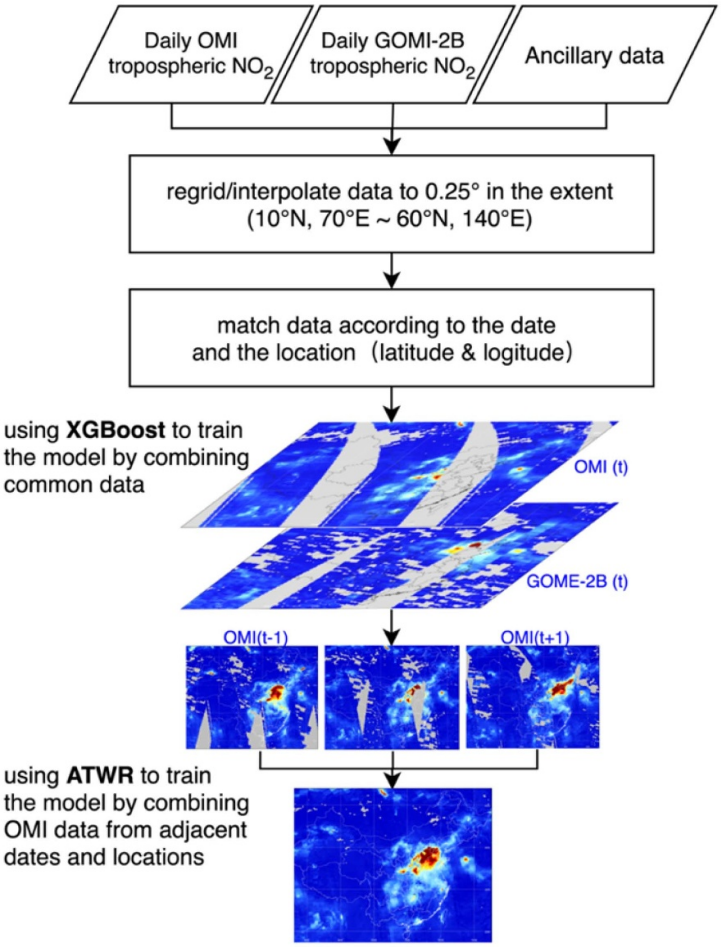

Figure 1. The flowchart of the reconstruction framework.

are given in figure 2. Overall, the model performs well with MAE $= 0.66 \times 10^{15}$ molecules cm^{-2}, RMSE $= 1.54 \times 10^{15}$ molecules cm^{-2}, and $R^2 = 0.89$. However, there is a small underestimation of the NO_2 vertical column at high values (greater than 80×10^{15} molecules cm^{-2}).

3. Results and discussions

3.1. Day-by-day reconstruction and analysis

The reconstructed data show improvement over OMI and either improvement over or similarity to the higher spatial resolution TROPOMI in terms of spatial coverage. First, there is a significant improvement over regions with missing data, due to either the known OMI row anomaly (red color) or extensive cloud cover (blue color), as clearly demonstrated in a specific case from 31 October to 2 November 2018 (figure 3, RECON-1). Second, there is a further reduction of missing data due to the inclusion of information over the three consecutive day period during which at least one input dataset has valid data (such as the grey dashed grid boxes) (figure 3, RECON-2).

The tropospheric NO_2 vertical columns in the reconstructed data are consistent with the various measurements that exist within the area of overlap, without showing distinct boundaries. RECON-2 and TROPOMI both identify most of the hot spots of tropospheric NO_2, with the following caveats. Although the spatial resolution of TROPOMI is superior to OMI where both products exist, there are also regions over which TROPOMI has no data (due to cloud contamination or lower data quality), but the reconstructed OMI NO_2 product does have data, such as hotspots in the Pearl River Delta, Henan, and Xinjiang, and the lesser polluted regions in Yunnan and Inner Mongolia. One result is that RECON-2 tropospheric NO_2 columns are on an average higher than TROPOMI (16.0%), consistent with the fact that such measurements overall tend to underestimate actual NO_2 (Verhoelst *et al* 2020) and the fact that RECON-2 tends to have a lower fraction of missing data over highly polluted regions, as compared to TROPOMI.

An important insight is made by exploring and analyzing the reconstructed NO_2 data from 15 March 2017 (figure 4). This date was selected because there is a total absence of OMI data from the 14 to the 16 due to acquisition problems, which would normally

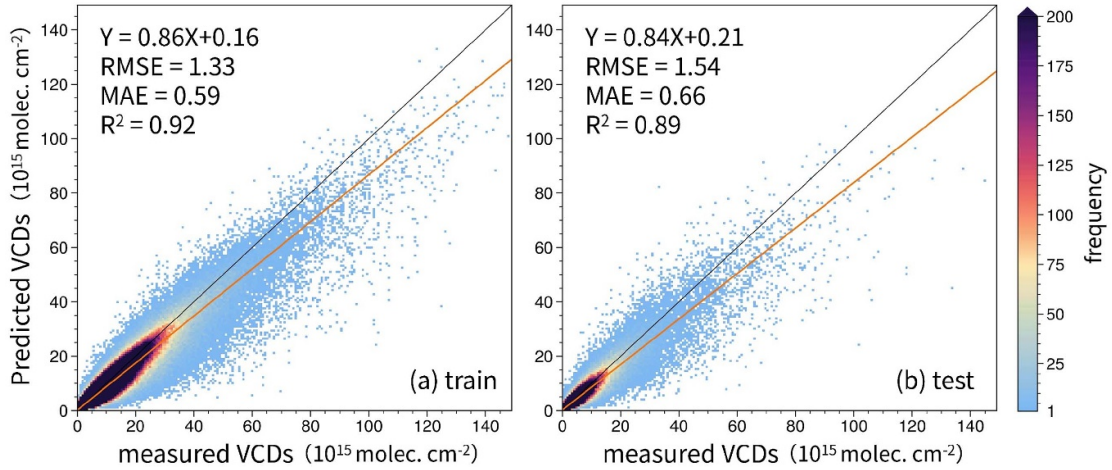

Figure 2. Scatterplots of model fitting cross-validation results of all daily vertical column density measurements (*x*-axis) and predictions (*y*-axis) over Mainland China. The black and orange lines represent the 1:1 and best-fit linear regression lines, respectively.

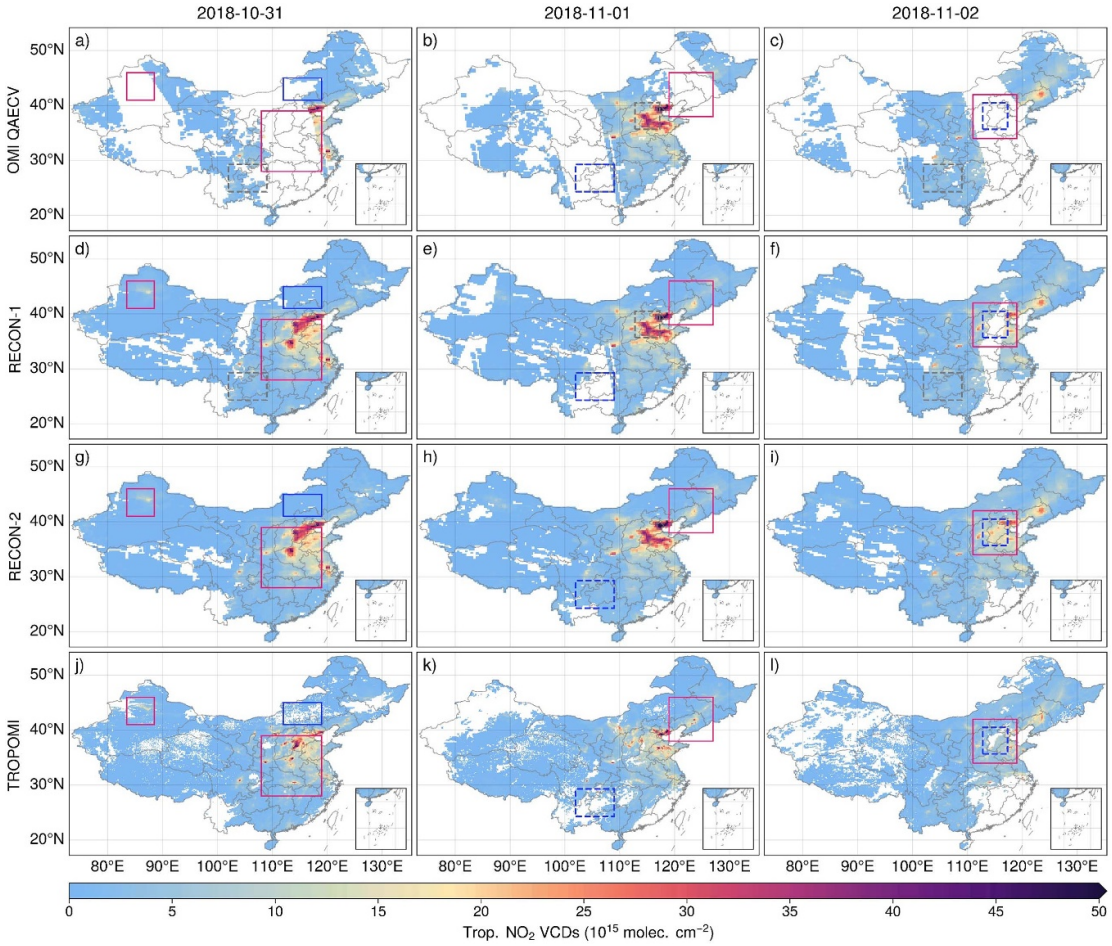

Figure 3. 31 October, 1 November, and 2 November 2018 OMI tropospheric NO_2 vertical columns generated by the reconstruction framework and measured by TROPOMI. The three columns represent the three dates respectively, with the first row (a–c) showing the original OMI measurements, the second row (d–f) showing RECON-1, the third row (g–i) showing RECON-2, and the final row (j–l) showing the measured TROPOMI NO_2 vertical columns. The red grid boxes represent the data missing due to OMI row anomaly, while the blue due to cloud cover. The grid boxes with solid lines are reconstructed in the first step, while those with dashed lines in the second step. The grey dashed box is just a reference for showing the same area over the consecutive days.

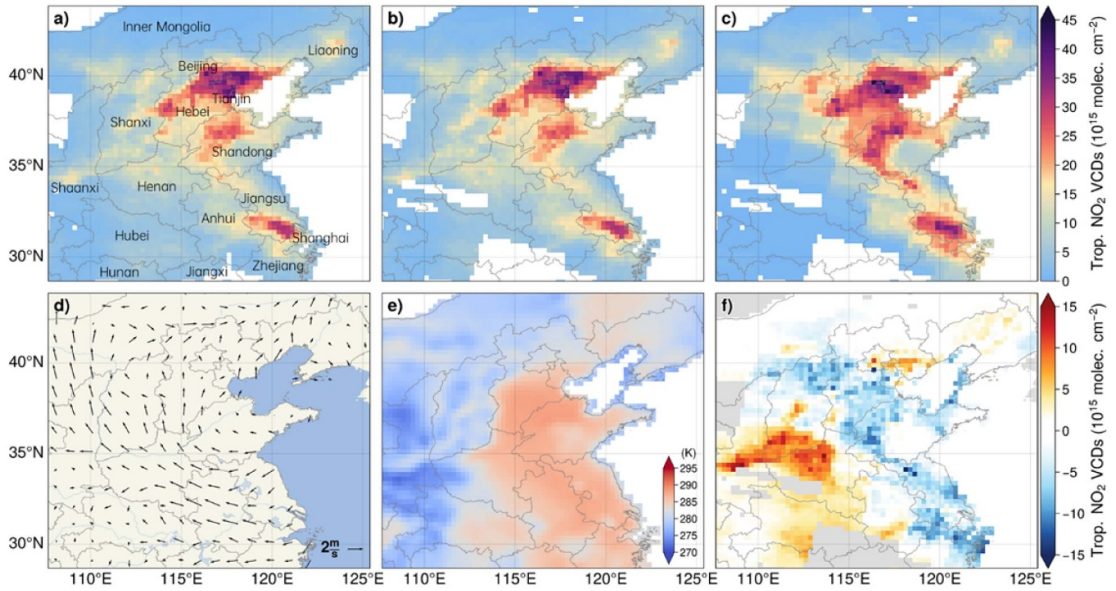

Figure 4. The spatial distribution of the tropospheric NO_2 vertical column over Central Eastern China on March 15 2017 from: (a) RECON-2, (b) RECON-1, (c) GOME-2B, and (f) the difference between RECON-2 and GOME-2B. ERA5 wind and temperature conditions during the satellite overpass times are shown in (d–e), respectively.

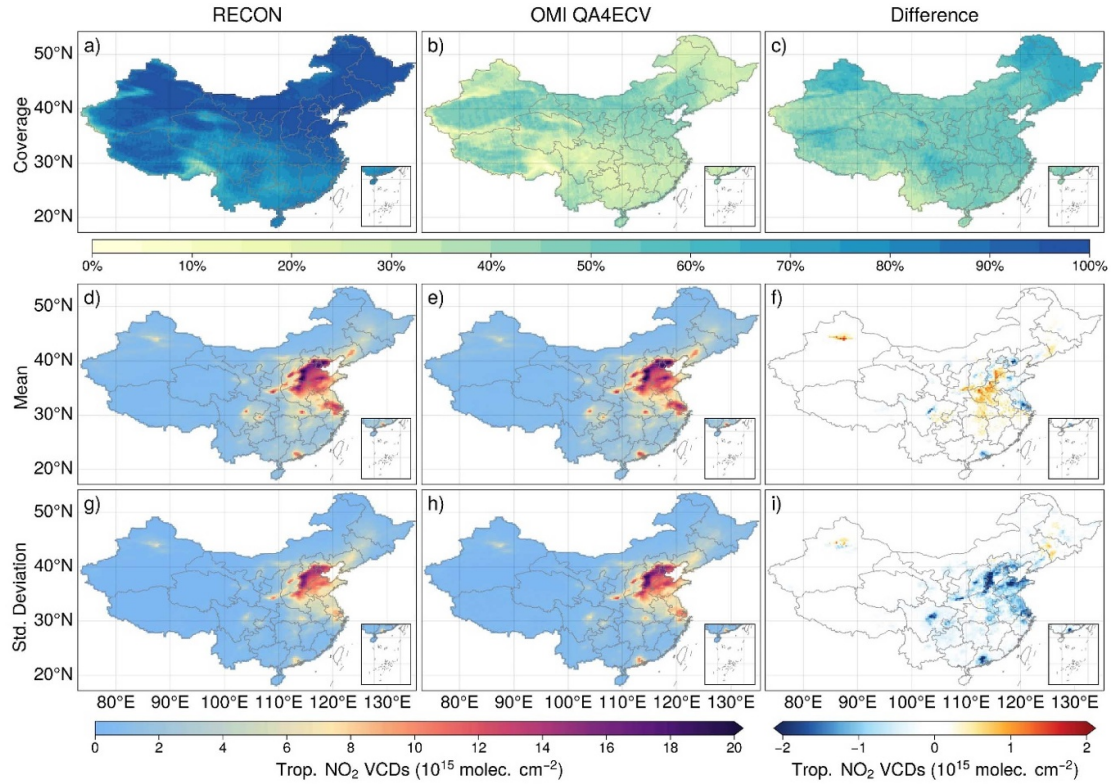

Figure 5. Statistics of daily tropospheric NO_2 VCDs from 2015 to 2018: (a) reconstructed coverage, (b) OMI coverage, (c) difference between reconstructed coverage and OMI coverage, (d) reconstructed mean, (e) OMI mean, (f) difference between reconstructed mean and OMI mean, (g) reconstructed standard deviation, (h) OMI standard deviation, (i) difference between reconstructed standard deviation and OMI standard deviation.

render any scientific analysis impossible on these days. However, the reconstructed result provides a significant amount of data for analysis, allowing for a scientific result to be achieved on the 15, where previously no possibility existed. First of all, the known spatial distribution of high NO_2 column loading over the Beijing-Tianjin-Hebei region is still present, but it is shown to also extend into the northern and western parts of Shandong as well as through the western and southern part of Jiangsu.

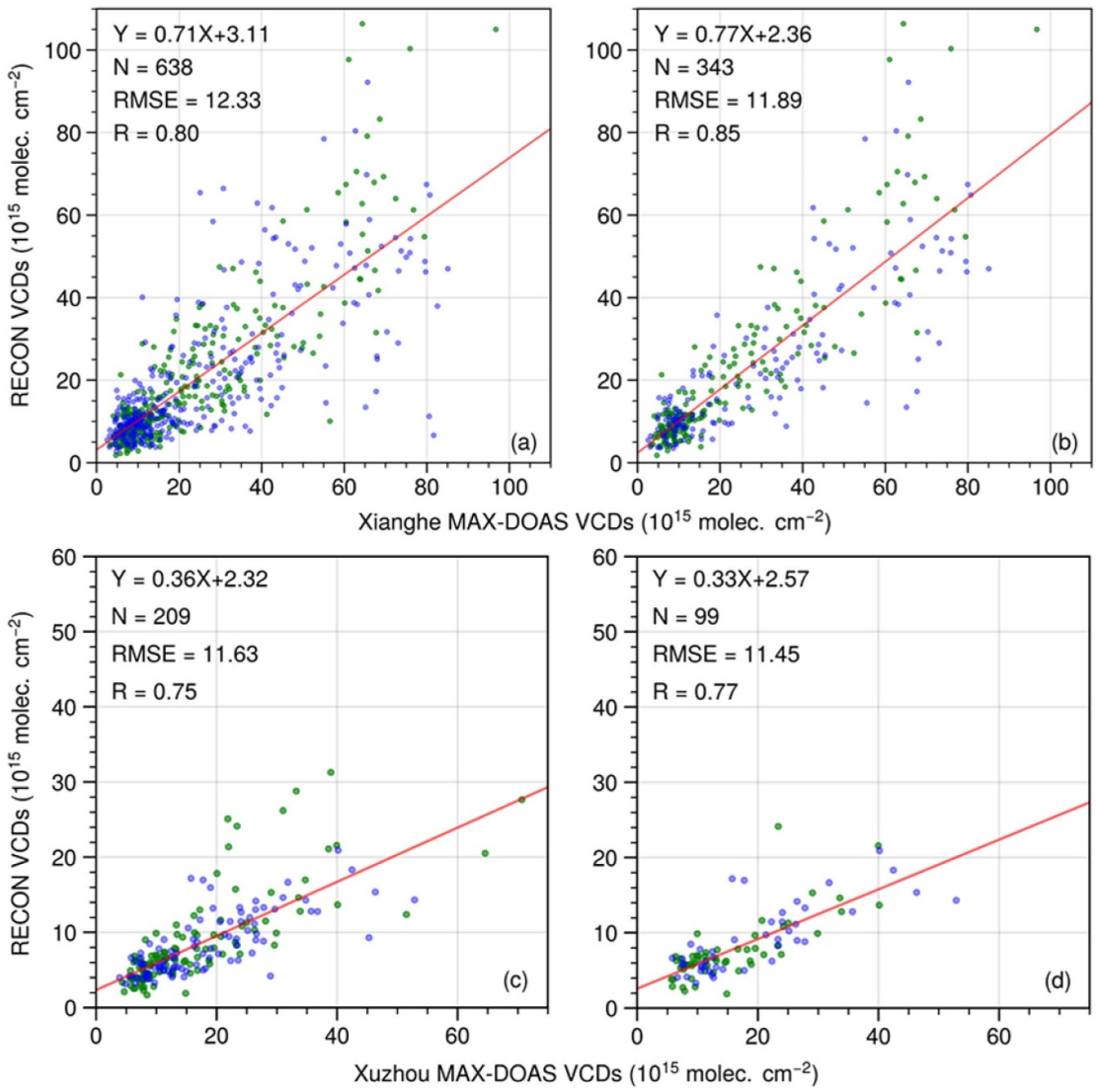

Figure 6. Scatter plots for tropospheric NO_2 VCDs between the RECONstructed data and MAX-DOAS in Xianghe (a–b) and Xuzhou (c–d). MAX-DOAS data are not filtered in the first column, while they are filtered in the second. The green dots represent the data from OMI QA4ECV product itself and the blue dots represent the newly added data relative to QA4ECV.

Secondly, over this region as a whole (except for Beijing itself) the difference between RECON-2 and GOME-2B shows a reduction on average, which is in fact completely consistent with the extremely high temperatures observed during this time (figure 4(e)). Thirdly, the elevated NO_2 vertical column loadings over Henan, Shanxi, Shaanxi, and Hubei is not normally observed, but is found to be completely consistent with the strong winds and subsequent long-range transport observed during this time (figure 4(d)). This indicates that there is not only a practical increase in data coverage using this technique but also an increase in the ability to make important scientific discoveries, frequently which have important scientific differences under difficult conditions of very cloudy or heavily polluted (and usually treated as clouds by remotely sensed measurements).

3.2. Climatological analysis

The spatial distribution of the OMI data coverage (the ratio of the number of days that each grid cell has valid data to the total 1461 d) from 2015 to 2018 is compared in figures 5(a)–(c), and the amount of reconstructed data is twice as much as that of the original data. Overall, the data coverage is improved by more than 40%, with the percentage of grid cells where OMI has no value in mainland China dropping from 63% to 22% in general.

The statistics of the mean NO_2 vertical columns during 2015 and 2018 from the reconstructed results and original measurements show a significant amount of similarity overall, including a high level of consistency in terms of the spatial distribution, as demonstrated in figures 5 (d)–(f). It should be noted that some of the most polluted regions (the Beijing-Tianjin-Hebei region, the Yangtze River Delta

region, the Pearl River Delta region, and Chengdu) show a moderately lower mean (about −2.3%), while some of the regions (Henan, Hubei, Hunan, Jiangxi, Anhui, Liaoning, Jilin, and Heilongjiang provinces) show a slightly higher mean (about +1.1%). Due to the inclusion of the filling of missing data, the drop in the standard deviation of the reconstructed OMI NO_2 VCDs has been shown over almost all of China (roughly −10.4%), as shown in figures 5(g)–(i). The only two minor exceptions are both found in small portions of rapidly developing energy regions in Northern Xinjiang and Jilin (+16.2%), and the reconstructed result has a larger mean (up to +12.9%) at the same time, which is possibly due to a rapidly changing emissions source and access to a greater amount of underlying data (Lin *et al* 2020).

3.3. Data evaluation

A comparison is made between the reconstructed NO_2 columns and measurements from MAX-DOAS as shown in figure 6. Following previous work (Lin *et al* 2014, Liu *et al* 2019, 2020a) a filter is applied to remove all MAX-DOAS data whose standard deviation exceeds 20% of their mean value, as a method to exclude local pollution events near the MAX-DOAS site. The results both without filter (figure 6(a), (c)) and with filter (figure 6(b), (d)) are provided. When the OMI QAECV data has a valid value, the reconstruction result still uses the original value, so match-up points are separated using blue and green colors, which represent the original OMI value and the newly added data relative to QA4ECV, separately. A good agreement of the Pearson correlation coefficient (R) between 0.75 and 0.85 is found between them in the two sites.

The filtered comparison between them has a better *R* and a lower RMSE, while the number of match-up points drops by a half. The detailed results of each item are listed in tables S2–S5. It seems that the reconstructed product has a slightly inferior performance. After calculating the mean value and standard deviation of the MAX-DOAS in the corresponding dates, no matter whether the filter is applied or not, the NO_2 columns from MAX-DOAS have a higher NO_2 loading when comparing to the reconstructed NO_2 columns, as compared to the OMI QA4ECV only. There is a bias in the satellite NO_2, which possibly comes from both the retrievals and the emissions. In reality, the inclusion of more data points has allowed us to both simulate the higher values of the NO_2 column loadings (consistent with the measured conditions) as well as to reduce the variance around these higher values. It should be noted that some regional data products partly correct the retrieval bias via improving air mass factor, such as POMINO (Lin *et al* 2014, 2015, Liu *et al* 2019) over China, HKOMI (Kuhlmann *et al* 2015) and BEHR-HK (Mak *et al* 2018) over the Pearl River Delta region, which deserve further study and comparison.

The time-series changes of daily tropospheric NO_2 columns derived from the reconstructed product and MAX-DOAS without a filter are shown in figures S3–S4, demonstrating that the reconstructed results capture the change of NO_2 columns.

Considering the number of MAX-DOAS stations is limited, the monthly and weekly mean of tropospheric NO_2 VCDs in 2018 from the reconstructed, original OMI QA4ECV and TROPOMI data are compared in figures S5–S6. They have consistency in temporal and spatial changes; however, the original OMI QA4ECV weekly average data also have missing areas in autumn and winter, and even existing in the monthly average data, while TROPOMI only appears at weekly scales.

4. Conclusions

This study successfully produces a tropospheric NO_2 vertical column loading with few gaps and values that are smooth and consistent in terms of both space and time, by applying a multi-step machine learning approach to merge two satellite measurements of NO_2 vertical column loadings and surface measurements (where available). This result, a roughly 40% improvement in data availability over mainland China, includes predicting data missing from row anomalies, cloud cover, and biases based on different overpass time from various platforms. The reconstructed NO_2 vertical columns are consistent spatio-temporally and yield a result that is superior compared with OMI, GOME-2B and TROPOMI in terms of data coverage at daily to weekly scales. Similar to the original OMI dataset, the reconstructed result agrees well with MAX-DOAS measurements (R: 0.75–0.85). A limitation of this approach is that the input measurements have different spatial resolutions and uncertainty values, which may lead to inconsistent uncertainty analysis. Additionally, this approach is limited to regions that have access to ground-based NO_2 measurements. Combining more precise quantification of uncertainties and more data will further reduce the uncertainty of the final reconstructed NO_2 vertical columns.

Importantly, the reconstructed NO_2 vertical column loadings are found to be scientifically useful in terms of analyzing events which would otherwise have no data. A specific period analyzed yielded two important observations in what was otherwise individually cloud-covered data from both the OMI and GOME-2B instruments. On this day, first, long-range transport of NO_2 occurred from Coastal to Central China. Second, there was a decrease in NO_2 due to faster chemical decay in Coastal China under abnormally high local temperatures.

Furthermore, the long-term statistics of the reconstructed NO_2 vertical column loadings yield a

climatological mean and standard deviation which are consistent with measurements and emissions sources on the ground. Some of the most polluted regions show a moderately lower mean (−2.3%), while some of the regions show a slightly higher mean (+1.1%). Almost all regions of China have been found to have a lower climatological standard deviation (−10.4%), consistent with the new approach having measurements that are more spatially and temporally consistent, as well as based on a larger number of data points.

Therefore, the new machine learning approach with multiple remotely sensed measurements of NO_2 from various platforms has been shown to provide access to considerably more high-quality data, as well as the ability to make new discoveries. For these reasons, we urge the community to carefully consider this approach for their future remotely sensed product releases, and both use and improve upon this data for their own scientific experiments.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 41975041 and No. 41871260), Xuzhou Key R&D Program (Grant No. KC18225), and the Ministry of Science and Technology of China (Grant No. 2016YFC0200506).

The OMI QA4ECV NO_2 data presented in this publication can be downloaded from KNMI http://temis.nl/airpollution/no2col/no2regioomi_qa.php. The GOME-2B offline products can be downloaded from EUMETSAT/DLR AC-SAF https://acsaf.org/products/oto_no2.html. The TROPOMI offline or reprocessing NO_2 products can be downloaded from EU/ESA https://s5phub.copernicus.eu/dhus/#/home. The MAX-DOAS measurements in Xianghe station can be downloaded from http://uv-vis.aeronomie.be/groundbased/QA4ECV_MAXDOAS/index.php.The ground level NO_2 concentrations were crawled from http://106.37.208.233:20035/. The ERA5 reanalysis data can be downloaded from EU/ECMWF https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels. The FROM-GLC land cover data can be found in http://data.ess.tsinghua.edu.cn/fromglc2017v1.html. The HARP tool is available at http://stcorp.github.io/harp/doc/html/index.html. The XGBoost model is available at https://xgboost.readthedocs.io. We would like to acknowledge them for making the data, tools and models publicly accessible.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.13126847.

ORCID iDs

Qin He https://orcid.org/0000-0002-6087-5293
Kai Qin https://orcid.org/0000-0002-1280-6330
Jason Blake Cohen https://orcid.org/0000-0002-9889-8175

References

Adams C *et al* 2019 Satellite-derived emissions of carbon monoxide, ammonia, and nitrogen dioxide from the 2016. Horse River wildfire in the Fort McMurray area *Atmos. Chem. Phys.* **19** 2577–99

Atkinson Richard W, Butland Barbara K, Ross A H and Maynard Robert L 2018 Long-term concentrations of nitrogen dioxide and mortality: a meta-analysis of cohort studies *Epidemiology* **29** 460–72

Boersma K F *et al* 2018 Improving algorithms and uncertainty estimates for satellite NO_2 retrievals: results from the Quality Assurance for the Essential Climate Variables (QA4ECV) project *Atmos. Meas. Tech.* **11** 6651–78

Boersma K F, Eskes H J and Brinksma E J 2004 Error analysis for tropospheric NO_2 retrieval from space *J. Geophys. Res. Atmos.* **109** D04311

Boersma K F, Eskes H, Richter A, De Smedt I, Lorente A, Beirle S, Van Geffen J, Peters E, Van Roozendael M and Wagner T 2017 QA4ECV NO_2 tropospheric and stratospheric column data from OMI (www.qa4ecv.eu/ecv/no2-pre/data)

Bovensmann H, Burrows J P, Buchwitz M, Frerick J, Noël S, Rozanov V V, Chance K V and Goede A P H 1999 SCIAMACHY: mission objectives and measurement modes *J. Atmos. Sci.* **56** 127–50

Burrows J P *et al* 1999 The Global Ozone Monitoring Experiment (GOME): mission concept and first scientific results *J. Atmos. Sci.* **56** 151–75

Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM) pp 785–94

Clémer K, Van Roozendael M, Fayt C, Hendrick F, Hermans C, Pinardi G, Spurr R, Wang P and De Mazière M 2010 Multiple wavelength retrieval of tropospheric aerosol optical properties from MAXDOAS measurements in Beijing *Atmos. Meas. Tech.* **3** 863–78

Cohen J B 2014 Quantifying the occurrence and magnitude of the Southeast Asian fire climatology *Environ. Res. Lett.* **9** 114018

Cohen J B, Ng D H L, Lim A W L and Chua X R 2018 Vertical distribution of aerosols over the Maritime Continent during El Niño *Atmos. Chem. Phys.* **18** 7095–108

Cohen J B and Prinn R G 2011 Development of a fast, urban chemistry metamodel for inclusion in global models *Atmos. Chem. Phys.* **11** 7629–56

Compernolle S *et al* 2020 Validation of Aura-OMI QA4ECV NO_2 climate data records with ground-based DOAS networks: the role of measurement and comparison uncertainties *Atmos. Chem. Phys.* **20** 8017–45

Cui Y, Zhang W, Bao H, Wang C, Cai W, Yu J and Streets D G 2019 Spatiotemporal dynamics of nitrogen dioxide pollution and urban development: satellite observations over China, 2005–2016 *Resour. Conserv. Recycl.* **142** 59–68

de Foy B, Lu Z and Streets D G 2016 Satellite NO_2 retrievals suggest China has exceeded its NO_x reduction goals from the twelfth five-year plan *Sci. Rep.* **6** 35912

de Hoogh K, Saucy A, Shtein A, Schwartz J, West E A, Strassmann A, Puhan M, Röösli M, Stafoggia M and Kloog I 2019 Predicting fine-scale daily NO_2 for 2005–2016 incorporating OMI satellite data across Switzerland *Environ. Sci. Technol.* **53** 10279–87

Georgoulias A K, van der A R J, Stammes P, Boersma K F and Eskes H J 2019 Trends and trend reversal detection in 2

decades of tropospheric NO_2 satellite observations *Atmos. Chem. Phys.* **19** 6269–94
Gong P *et al* 2013 Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data *Int. J. Remote Sens.* **34** 2607–54
Gu D, Wang Y, Smeltzer C and Liu Z 2013 Reduction in NO_x emission trends over China: regional and seasonal variations *Environ. Sci. Technol.* **47** 12912–9
Guo J *et al* 2017 Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: implications for PM2.5 remote sensing *Environ. Pollut.* **221** 94–104
Guo J, Li Y, Cohen J B, Li J, Chen D, Xu H, Liu L, Yin J, Hu K and Zhai P 2019 Shift in the temporal trend of boundary layer height in China using long-term (1979–2016) radiosonde data *Geophys. Res. Lett.* **46** 6080–9
Hengl T *et al* 2017 SoilGrids250m: global gridded soil information based on machine learning ed B Bond-Lamberty *Plos One* **12** e0169748
Just A, De Carli M, Shtein A, Dorman M, Lyapustin A and Kloog I 2018 Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2.5 in the Northeastern USA *Remote Sens.* **10** 803
Kang D, Mathur R, Pouliot G A, Gilliam R C and Wong D C 2020 Significant ground-level ozone attributed to lightning-induced nitrogen oxides during summertime over the Mountain West States *Npj Clim. Atmos. Sci.* **3** 6
Kim J *et al* 2020 New era of air quality monitoring from space: geostationary environment monitoring spectrometer (GEMS) *Bull. Am. Meteorol. Soc.* **101** E1–22
Kong H, Lin J, Zhang R, Liu M, Weng H, Ni R, Chen L, Wang J, Yan Y and Zhang Q 2019 High-resolution ($0.05^\circ \times 0.05^\circ$) NO_x emissions in the Yangtze river delta inferred from OMI *Atmos. Chem. Phys.* **19** 12835–56
Krotkov N A, Lamsal L N, Celarier E A, Swartz W H, Marchenko S V, Bucsela E J, Chan K L, Wenig M and Zara M 2017 The version 3 OMI NO_2 standard product *Atmos. Meas. Tech.* **10** 3133–49
Kuhlmann G, Lam Y F, Cheung H M, Hartl A, Fung J C H, Chan P W and Wenig M O 2015 Development of a custom OMI NO_2 data product for evaluating biases in a regional chemistry transport model *Atmos. Chem. Phys.* **15** 5627–44
Lamsal L N, Martin R V, van Donkelaar A, Steinbacher M, Celarier E A, Bucsela E, Dunlea E J and Pinto J P 2008 Ground-level nitrogen dioxide concentrations inferred from the satellite-borne ozone monitoring instrument *J. Geophys. Res.* **113** D16308
Lee H J and Koutrakis P 2014 Daily ambient NO_2 concentration predictions using satellite ozone monitoring instrument NO_2 data and land use regression *Environ. Sci. Technol.* **48** 2305–11
Levelt P F *et al* 2018 The ozone monitoring instrument: overview of 14 years in space *Atmos. Chem. Phys.* **18** 5699–745
Levelt P F, van den Oord G H J, Dobber M R, Malkki A, Visser H, de Vries J, Stammes P, Lundell J O V and Saari H 2006 The ozone monitoring instrument *IEEE Trans. Geosci. Remote Sens.* **44** 1093–101
Li R, Cui L, Fu H, Meng Y, Li J and Guo J 2020 Estimating high-resolution PM1 concentration from Himawari-8 combining extreme gradient boosting-geographically and temporally weighted regression (XGBoost-GTWR) *Atmos. Environ.* **229** 117434
Lin C, Cohen J B, Wang S and Lan R 2020 Application of a combined standard deviation and mean based approach to MOPITT CO column data, and resulting improved representation of biomass burning and urban air pollution sources *Remote Sens. Environ.* **241** 111720
Lin J-T, Liu M-Y, Xin J-Y, Boersma K F, Spurr R, Martin R and Zhang Q 2015 Influence of aerosols and surface reflectance on satellite NO_2 retrieval: seasonal and spatial characteristics and implications for NO_x emission constraints *Atmos. Chem. Phys.* **15** 11217–41
Lin J-T, Martin R V, Boersma K F, Sneep M, Stammes P, Spurr R, Wang P, Van Roozendael M, Clémer K and Irie H 2014 Retrieving tropospheric nitrogen dioxide from the ozone monitoring instrument: effects of aerosols, surface reflectance anisotropy, and vertical profile of nitrogen dioxide *Atmos. Chem. Phys.* **14** 1441–61
Lin J, McElroy M B and Boersma K F 2010 Constraint of anthropogenic NO_x emissions in China from different sectors: a new methodology using multiple satellite retrievals *Atmos. Chem. Phys.* **10** 63–78
Liu F, Beirle S, Zhang Q, Dörner S, He K and Wagner T 2016 NO_x lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations *Atmos. Chem. Phys.* **16** 5283–98
Liu F, Beirle S, Zhang Q, Van der A R J, Zheng B, Tong D and He K 2017a NO_x emission trends over Chinese cities estimated from OMI observations during 2005–2015 *Atmos. Chem. Phys* **17** 9261–75
Liu M *et al* 2019 Improved aerosol correction for OMI tropospheric NO_2 retrieval over East Asia: constraint from CALIOP aerosol vertical profile *Atmos. Meas. Tech.* **12** 1–21
Liu M *et al* 2020a A new TROPOMI product for tropospheric NO_2 columns over East Asia with explicit aerosol corrections *Atmos. Meas. Tech.* **13** 4247–59
Liu S, Valks P, Pinardi G, Xu J, Argyrouli A, Lutz R, Tilstra L G, Huijnen V, Hendrick F and Van Roozendael M 2020b An improved air mass factor calculation for nitrogen dioxide measurements from the global ozone monitoring experiment-2 (GOME-2) *Atmos. Meas. Tech.* **13** 755–87
Liu X, Mizzi A P, Anderson J L, Fung I Y and Cohen R C 2017b Assimilation of satellite NO_2 observations at high spatial resolution using OSSEs *Atmos. Chem. Phys.* **17** 7067–81
Lorente A *et al* 2017 Structural uncertainty in air mass factor calculation for NO_2 and HCHO satellite retrievals *Atmos. Meas. Tech.* **10** 759–82
Lorente A, Boersma K F, Stammes P, Tilstra L G, Richter A, Yu H, Kharbouche S and Muller J-P 2018 The importance of surface reflectance anisotropy for cloud and NO_2 retrievals from GOME-2 and OMI *Atmos. Meas. Tech.* **11** 4509–29
Mak H, Laughner J, Fung J, Zhu Q and Cohen R 2018 Improved satellite retrieval of tropospheric NO_2 column density via updating of air mass factor (AMF): case study of Southern China *Remote Sens.* **10** 1789
Mitchell R and Frank E 2017 Accelerating the XGBoost algorithm using GPU computing *PeerJ Comput. Sci.* **3** e127
Munro R *et al* 2016 The GOME-2 instrument on the Metop series of satellites: instrument design, calibration, and level 1 data processing – an overview *Atmos. Meas. Tech.* **9** 1279–301
Núñez-Alonso D, Pérez-Arribas L V, Manzoor S and Cáceres J O 2019 Statistical tools for air pollution assessment: multivariate and spatial analysis studies in the Madrid Region *J. Anal. Methods Chem.* **2019** 1–9
Peng X, Shen H, Zhang L, Zeng C, Yang G and He Z 2016 Spatially continuous mapping of daily global ozone distribution (2004–2014) with the Aura OMI sensor *J. Geophys. Res. Atmos.* **121** 12,702–22
Penn E and Holloway T 2020 Evaluating current satellite capability to observe diurnal change in nitrogen oxides in preparation for geostationary satellite missions *Environ. Res. Lett.* **15** 034038
Qin K, Han X, Li D, Xu J, Loyola D, Xue Y, Zhou X, Li D, Zhang K and Yuan L 2020 Satellite-based estimation of surface NO_2 concentrations over east-central China: a comparison of POMINO and OMNO2d data *Atmos. Environ.* **224** 117322
Qin K, Rao L, Xu J, Bai Y, Zou J, Hao N, Li S and Yu C 2017 Estimating ground level NO_2 concentrations over Central-Eastern China using a satellite-based geographically and temporally weighted regression model *Remote Sens.* **9** 950
Richter A, Burrows J P, Nüß H, Granier C and Niemeier U 2005 Increase in tropospheric nitrogen dioxide over China observed from space *Nature* **437** 129–32

Shen H, Li X, Cheng Q, Zeng C, Yang G, Li H and Zhang L 2015 Missing information reconstruction of remote sensing data: a technical review *IEEE Geosci. Remote Sens. Mag.* **3** 61–85

Shtein A *et al* 2019 Estimating daily PM2.5 and PM10 over Italy using an ensemble model *Environ. Sci. Technol.* **54** 120–8

Si Y, Wang H, Cai K, Chen L, Zhou Z and Li S 2019 Long-term (2006–2015) variations and relations of multiple atmospheric pollutants based on multi-remote sensing data over the North China Plain *Environ. Pollut.* **255** 113323

Silver J D, Brandt J, Hvidberg M, Frydendall J and Christensen J H 2013 Assimilation of OMI NO_2 retrievals into the limited-area chemistry-transport model DEHM (V2009.0) with a 3-D OI algorithm *Geosci. Model Dev.* **6** 1–16

Silvern R F *et al* 2019 Using satellite observations of tropospheric NO_2 columns to infer long-term trends in US NO_x emissions: the importance of accounting for the free tropospheric NO_2 background *Atmos. Chem. Phys.* **19** 8863–78

Tiwari S and Agrawal M 2018 Tropospheric ozone budget: formation, depletion and climate change *Tropospheric Ozone and Its Impacts on Crop Plants: A Threat to Future Global Food Security* S Tiwari and M Agrawal ed (Berlin: Springer) pp 31–64

van der A R J, Mijling B, Ding J, Koukouli M E, Liu F, Li Q, Mao H and Theys N 2017 Cleaning up the air: effectiveness of air quality policy for SO_2 and NO_x emissions in China *Atmos. Chem. Phys.* **17** 1775–89

Veefkind J P *et al* 2012 TROPOMI on the ESA Sentinel-5 precursor: a GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications *Remote Sens. Environ.* **120** 70–83

Verhoelst T *et al* 2020. Ground-based validation of the Copernicus Sentinel-5p TROPOMI NO_2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks *Atmos. Meas. Tech. Discuss.* (https://doi.org/10.5194/amt-2020-119)

Wang C, Wang T and Wang P 2019 The spatial–temporal variation of tropospheric NO_2 over China during 2005–2018 *Atmosphere* **10** 444

Wang S, Cohen J B, Lin C and Deng W 2020 Constraining the relationships between aerosol height, aerosol optical depth and total column trace gas measurements using remote sensing and models *Atmos. Chem. Phys. Discuss.* (https://doi.org/10.5194/acp-2019-1017)

Wang S, Xing J, Jang C, Zhu Y, Fu J S and Hao J 2011a Impact assessment of ammonia emissions on inorganic aerosols in East China using response surface modeling technique *Environ. Sci. Technol.* **45** 9293–300

Wang X, Mallet V, Berroir J-P and Herlin I 2011b Assimilation of OMI NO_2 retrievals into a regional chemistry-transport model for improving air quality forecasts over Europe *Atmos. Environ.* **45** 485–92

Wang Y *et al* 2017a Validation of OMI, GOME-2A and GOME-2B tropospheric NO_2, SO_2 and HCHO products using MAX-DOAS observations from 2011 to 2014 in Wuxi, China: investigation of the effects of priori profiles and aerosols on the satellite products *Atmos. Chem. Phys.* **17** 5007–33

Wang Y, Lampel J, Xie P, Beirle S, Li A, Wu D and Wagner T 2017b Ground-based MAX-DOAS observations of tropospheric aerosols, NO_2, SO_2 and HCHO in Wuxi, China, from 2011 to 2014 *Atmos. Chem. Phys.* **17** 2189–215

Weng H, Lin J, Martin R, Millet D B, Jaeglé L, Ridley D, Keller C, Li C, Du M and Meng J 2020 Global high-resolution emissions of soil NO_x, sea salt aerosols, and biogenic volatile organic compounds *Sci. Data* **7** 148

Yang K, Carn S A, Ge C, Wang J and Dickerson R R 2014 Advancing measurements of tropospheric NO_2 from space: new algorithm and first global results from OMPS *Geophys. Res. Lett.* **41** 4777–86

Zara M *et al* 2018 Improved slant column density retrieval of nitrogen dioxide and formaldehyde for OMI and GOME-2A from QA4ECV: intercomparison, uncertainty characterisation, and trends *Atmos. Meas. Tech.* **11** 4033–58

Zhan Y, Luo Y, Deng X, Zhang K, Zhang M, Grieneisen M L and Di B 2018 Satellite-based estimates of daily NO_2 exposure in China using hybrid random forest and spatiotemporal kriging model *Environ. Sci. Technol.* **52** 4180–9

Zhang Q, Geng G, Wang S, Richter A and He K 2012a Satellite remote sensing of changes in NO_x emissions over China during 1996–2010 *Chin. Sci. Bull.* **57** 2857–64

Zhang Q, He K and Huo H 2012b Cleaning China's air *Nature* **484** 161–2

Zhao B, Wang S X, Liu H, Xu J Y, Fu K, Klimont Z, Hao J M, He K B, Cofala J and Amann M 2013 NO_x emissions in China: historical trends and future perspectives *Atmos. Chem. Phys.* **13** 9869–97