A global daily High Spatial-Temporal Coverage Merged tropospheric NO_2 dataset (HSTCM-NO_2) from 2007 to 2022 based on OMI and GOME-2

Kai Qin[1], Hongrui Gao[1], Xuancen Liu[1], Qin He[1], Pravash Tiwari[1], Jason Blake Cohen[1*]

[1]School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, 221116, China

Correspondence to: Jason Blake Cohen (jasonbc@alum.mit.edu)

Abstract. Remote sensing based on satellites can provide long-term, consistent, and global coverage of NO_2 (an important atmospheric air pollutant) as well as other trace gases. However, satellites often miss data due to factors including but not limited to clouds, surface features, and aerosols. Moreover, as one of the longest continuous observational platforms of NO_2, OMI has suffered from missing data over certain rows since 2007, significantly reducing spatial coverage. This work uses the OMI-based OMNO2 product, as well as a NO_2 product from GOME-2 in combination with machine learning (XGBoost) and spatial interpolation (DINEOF) method to produce a 16-year global daily High Spatial-Temporal Coverage Merged tropospheric NO_2 dataset (HSTCM-NO_2, https://doi.org/10.5281/zenodo.10968462, Qin et al., 2024), which increases the average global spatial coverage of NO_2 from 39.5% to 99.1%. The HSTCM-NO_2 dataset is validated using upward-looking observations of NO_2 (MAX-DOAS), other satellites (TROPOMI), and reanalysis products. The comparisons show that HSTCM-NO_2 maintains a good correlation with the magnitude of other observational datasets, except for under heavily polluted conditions (>6×10^{15} molec.cm^{-2}). This work also introduces a new validation technique to validate coherent spatial and temporal signals (EOF) and validates that the HSTCM-NO_2 is not only consistent with the original OMNO2 data, but in some parts of the world effectively fills in missing gaps and yields a superior result when analyzing long-range atmospheric transport of NO_2. The few differences are also reported to be related to areas in which the original OMNO2 signal was very low, which has been shown elsewhere, but not from this perspective, further validating that applying a minimum cutoff to retrieved NO_2 data is essential. The reconstructed data product can effectively extend the utilization value of the original OMNO2 data, and the data quality of HSTCM-NO_2 can meet the needs of scientific research.

1 Introduction

The sum of nitrogen dioxide (NO_2) and nitrogen oxide, hereafter referred to as nitrogen oxides (NO_x), plays several important roles in tropospheric chemistry (Eriksson, 1952; Levy, 1972; Crutzen, 1973; Fishman et al., 1979; Crutzen, 1979; Logan et al., 1981), specifically with respect to tropospheric ozone (Sillman et al., 1990), nitrate aerosol (Lu et al., 2021), which indirectly influences radiative forcing both through scattering downward propagating visible light (Richter et al., 2005), as well as through enhancing absorption of black carbon aerosols (Tiwari et al., 2023), and the concentration of tropospheric OH, which indirectly influences both methane and carbon monoxide (Lu and Khalil, 1993; Spivakovsky et al., 2000). During the daytime, under low pollution and low cloud conditions, the photochemical cycle of NO_x can be scaled somewhat stably to NO_2, allowing observations of NO_2 to be an indicator of NO_x concentration (D. Schaub et al., 2006). Under more heavily polluted conditions, such a relationship can also be established, although it is found to vary in space and month-by-month (Qin et al., 2023; Li et al., 2023). Due to its rapid reactivity with water vapor, NO_x forms into nitric acid, contributes directly to acid rain (Wang et al., 2024). Additionally, NO_x has been shown to have adverse effects on human health (Liu et al., 2016), specifically, as an irritant of the respiratory system and via impacts on respiratory diseases when inhaled at high levels (Manisalidis et al., 2020). The Differential Optical Absorption Spectroscopy (DOAS) method is used extensively to retrieve total column amounts of trace gases such as NO_2 and others based on UV-visible measurements of satellite spectrometers (Eskes and Boersma, 2003).

The DOAS technique is based on the wavelength-dependent absorption of light over a specified light path, and it leads to the application of continuous monitoring of tropospheric pollution levels from space (Platt and Stutz, 2008). Initially applied to ground-based upward-looking instruments (i.e. MAX-DOAS, Wagner et al., 2004), nowadays, satellite-based measurements have been proven to offer reliable inversions of column NO_2 when compared with ground-based measurements (Bauer et al., 2012; Wang et al., 2017; Ialongo et al., 2020), with the errors commonly within a 20% bound and nearly always within a 40% bound (Boersma et al., 2004; Irie et al., 2012; Wang et al., 2017; Compernolle et al., 2020; Pinardi et al., 2020; Wang et al., 2020; Verhoelst et al., 2021).

Satellite observations offer the advantages of wide spatial and long-term temporal coverage (Streets et al., 2013), which can help fill spatial gaps between ground-based observations, and do so using a single platform without the need for calibrating multiple individual machines (Kolle et al., 2021). Starting nearly two decades ago, and continuing today, an array of different satellites has been monitoring global tropospheric NO_2 distributions including GOME (from 1995 to 2003) aboard ERS-2, SCIAMACHY (from 2002 to 2012) aboard Envisat, OMI (from 2004) aboard EOS-AURA, GOME-2 (from 2006) aboard Metop and TROPOMI (from 2017) aboard Sentinel-5P (Bovensmann et al., 1999; Laan et al., 2001; Richter and Burrows, 2002; Veefkind et al., 2012; Munro et al., 2016). As a result, there have been useful products relating to estimating surface or near-surface NO_2 emissions (Wang et al., 2012; Li et al., 2021) and detecting the long-term or short-term change of NO_2 (van der A et al., 2006; Cooper et al., 2022).

NO_x is emitted any time there is a high temperature reaction that occurs within the air (Echterhof and Pfeifer, 2012). For this reason, most sources are related to anthropogenic combustion of fossil fuels, biomass, and even forests, as well as a small amount from natural sources induced by lightning. (Sun et al., 2018; Lu et al., 2021; Li et al., 2022). Emissions are frequently computed using a bottom-up approach, where economic, population, and other factors are merged with an activity coefficient associated with each parameter, and applied on average over space and time (Li et al., 2017, Xu et al., 2023). Recent work has looked at using the satellite observations of NO_2 above and applying them on a grid-by-grid and day-by-day basis to attribute emissions to different types of industrial sources, population centers, power generation, transportation, residential uses, agriculture, and natural sources (Li et al., 2023, Qin et al., 2023). Current best estimates vary by considerable amounts from each other in space and time (Wang et al., 2021), and account for both natural (Deng et al., 2021) and human-based factors (according to EDGAR and MEIC). There is controversy about the amounts that lightning and microbial activity may or may not contribute (Logan, 1983).

Vertical column densities (VCDs) of tropospheric NO_2 retrieved from satellite-based instruments provide plentiful data under relatively clean and clear atmospheric conditions, but have many missing pixels in both time and space due to a variety of factors including very bright surfaces, clouds, and aerosols (Lin et al., 2014; Xia and Jia, 2022). One of the underlying sources of error is related to the air mass factor (AMF), which allows conversion from a slant column to a vertical column, which is highly sensitive to cloud and aerosol layer height (Leitão et al., 2010), aerosol absorption (Lin et al., 2014; Cooper et al., 2019), and the spatial and temporal distribution of NO_x emissions (Qin et al., 2023; Li et al., 2023), which can lead to both uncertainties and biases in the retrieval (Bousserez, 2014). For these reasons, pixels known to be impacted by clouds are usually filtered before analysis, however, other impacted pixels may not be properly filtered, leading to other issues. Similarly, for some older satellites, due to the orbit and swath width, it respectively requires 3 days, 6 days and 1.5 days for GOME, SCIAMACHY and GOME-2 to cover the whole globe, with additional missing pixels on a day-by-day basis. OMI, which is carried on a near-polar, sun-synchronous satellite, is the world's first sensor with daily global coverage of NO_2 since 2004. However, in 2007, a reduction in OMI's spatial coverage occurred due to an equipment malfunction, called the row anomaly (RA), which began affecting just two rows of data in June 2007 but has gradually worsened over time (Torres et al., 2018). The absence of data presently affects both short-term estimation of air quality as well as long-term quantitative analysis (Duncan et al., 2013; van Geffen et al., 2020), although OMI is still useful for the detection of extreme events (Wang et al., 2020; Wang et al., 2021; Deng et al., 2021). Due to 19 years of continuous observations, OMI is a very widely used sensor in

the field of atmospheric trace gas research, and finding ways to comprehensively and reasonably fill these missing pixels would allow its usefulness to be extended into other fields (de Hoogh et al., 2019; He et al., 2020; Wu et al., 2021; Wei et al., 2022; Shao et al., 2023; Liu et al., 2024).

There are many existing approaches to fill missing data from satellite-based platforms including interpolation techniques: geostatistical (e.g., kriging), deterministic (e.g., inverse distance weighted, thin plate splines), and hybrid (e.g., regression kriging) methods (Abdulmanov et al., 2021; Achite et al., 2024), as well as machine learning techniques include random forests (Sanabria et al., 2013). As there is a strong correlation in terms of both geospatial relationships as well as retrieval approaches used to determine the VCDs between tropospheric NO_2 obtained by different sensors (Wang et al., 2016; Park et al., 2020), issues of spatial-temporal correlation need to be carefully taken into consideration, something that these previous approaches may not have fully considered. In this work, the machine learning and Data Interpolating Empirical Orthogonal Functions (DINEOF) methods are selected to carry out the reconstruction, which takes advantage of both machine learning and pattern recognition in tandem, as demonstrated by previous studies reconstructing satellite chlorophyll-a data (Wang and Liu, 2013; Chang et al., 2017; Hilborn and Costa, 2018; Park et al., 2020), filling in missing part of both sea and land surface temperature data (Alvera-Azcárate et al., 2009; Zhou et al., 2017), analyzing sea surface salinity data (Alvera-Azcárate et al., 2016; Chen et al., 2022), and Jiang et al. (2022) used DINEOF to reconstruct the XCO_2 data of OCO-2 and OCO-3 by fusing the two, effectively improving the spatiotemporal coverage of XCO_2 products.

This research aims to accurately and precisely reconstruct the tropospheric NO_2 VCD at daily time resolution and grid-by-grid spatial resolution using OMI 2007-2022. Under the support of global daily High Spatial-Temporal Coverage Merged tropospheric NO_2 dataset (HSTCM-NO_2), model validation, spatial distribution analysis and temporal change monitoring can be carried out. Also, HSTCM-NO_2 can be an ideal tool for improving numerical prediction of air quality and AMF, contributing to a better understanding of typical chemical and dynamic processes in the atmosphere, and future remote sensing retrieval improvements.

2 Materials and methods

Table 1: Summary of the parameters used in this research.

Data type	Parameter	Abbreviation	Unit
OMI	Daily tropospheric NO_2 vertical column density	OMI	molec.cm^{-2}
GOME-2	Daily tropospheric NO_2 vertical column density	GOME-2_NO_2	molec.cm^{-2}
	Daily cloud cover	cloud_fraction	%
Land cover types	Water bodies	wb	-
	Evergreen needleleaf vegetation	env	-
	Evergreen broadleaf vegetation	ebv	-
	Deciduous needleleaf vegetation	dnv	-
	Deciduous broadleaf vegetation	dbv	-
	Annual broadleaf vegetation	abv	-
	Annual grass vegetation	agv	-
	Non-vegetated lands	nvl	-

	Urban and built-up lands	ubl	-
ERA5 single levels	Surface pressure	sp	Pa
	Mean surface downward UV radiation flux	msdwuvrf	$W.m^{-2}$
	Total column ozone	tco3	$kg.m^{-2}$
	UV visible albedo for diffuse radiation	aluvd	-
	UV visible albedo for direct radiation	aluvp	-
ERA5 multi-levels	Specific rain water content	crwc	$kg.kg^{-2}$
	Ozone mass mixing ratio	o3	$kg.kg^{-2}$
	Relative humidity	r	%
	Temperature	t	K
	U-component of wind	u	$m.s^{-1}$
	V-component of wind	v	$m.s^{-1}$
	Vertical velocity	w	$Pa.s^{-1}$
Others	Latitude	-	-
	Longitude	-	-
	Day of year	doy	-

2.1 Tropospheric NO_2 products

2.1.1 OMI tropospheric NO_2 (OMNO2)

OMI is a UV/VIS charge-coupled device (CCD) spectrometer aboard Aura satellite, which was launched on 15 July 2004 into a Sun-synchronous orbit with a local equator crossing time of approximately 13:45. OMI covers a spectrum of 270–500 nm with a spectral resolution between 0.42 nm and 0.63 nm and a nominal spatial resolution of 13 km×24 km at nadir (Boersma et al., 2008; Foret et al., 2014), providing coverage over 740 wavelength bands along the satellite track and global coverage via 14 orbits per day.

OMI data are processed and archived at NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC). This work specifically uses the daily Level 3 global gridded data product that corresponds to the OMI NO_2 standard product (OMNO2), and the adopted L3 grid is a 0.25-degree by 0.25-degree grid in longitude and latitude.

2.1.2 GOME-2 tropospheric NO_2

The Global Ozone Monitoring Experiment-2 (GOME-2) is an optical spectrometer aboard the MetOp satellites. MetOp-A was launched on 19 October 2006, MetOp-B was launched on 17 September 2012, and MetOp-C was launched on 7 November 2018. GOME-2 senses backscattered and reflected radiance in the ultraviolet and visible part of the spectrum from 240 nm-790 nm, with a high spectral resolution between 0.26 nm and 0.51 nm covering 4096 spectral points from four detector channels (Fioletov et al., 2013). The spatial resolution varies from 80 km×40 km to 40 km×40 km, and provides daily near global coverage at the equator (Liu et al., 2019).

The GOME Data Processor version 4.8 is used for MetOp-A and -B, while version 4.9 is used for -C. Datasets were resampled

at a uniform gridding of 0.25×0.25 degree using HARP tool.

2.1.3 TROPOMI tropospheric NO_2

The Tropospheric Monitoring Instrument (TROPOMI) was launched on October 13, 2017, aboard the polar-orbiting Sentinel-5 Precursor satellite. It measures solar radiation reflected by and emitted from Earth, and provides measurements of atmospheric trace including NO_2, O_3, SO_2, HCHO, CH_4, and CO, as well as cloud and aerosol properties. NO_2 retrieval is performed using the visible band (400-496 nm), which has spectral resolution and sampling of 0.54 and 0.20 nm. The instrument operates in a push-broom configuration with a swath width of approximately 2,600 km, yielding on Earth's surface. The typical pixel size (near nadir) for NO_2 is 7 km×3.5 km which was reduced to 5.5 km×3.5 km in 2019 (Ialongo et al., 2020; Ludewig et al., 2020). This work specifically uses the level 2 NO_2 data products based on version 1.4, and an applied quality filter of qa_value>0.75 (van Geffen et al., 2019). The TROPOMI data products are also resampled to a spatial resolution of 0.25°×0.25° by HARP.

2.2 Auxiliary data

2.2.1 Land cover type data

The Moderate Resolution Imaging Spectroradiometer (MODIS) land cover type (MCD12Q1) provides data that maps global land cover at 500-meter spatial resolution annually derived from six different classification schemes. The maps were created from classifications of spectra-temporal features derived using the BIOME-Biogeochemical Cycles approach described by Running et al. (1993).

2.2.2 MAX-DOAS data

Multi-AXis Differential Optical Absorption Spectroscopy (MAX-DOAS) is a passive DOAS ground-based remote sensing observation technology using solar scattering as the light source. MAX-DOAS technology can be used to detect trace gases in the troposphere and has been widely applied in related fields. This instrument can observe scattered sunlight from different perspectives, thus having high sensitivity to trace gases in the troposphere, specifically using low elevation observations as the measurement intensity and zenith measurements as the reference intensity. The Beer-Lambert Law can be used to determine the total molecular amount of specific gas categories along the optical path (subtracting zenith concentration from non-zenith measurements), which is known as differential slant column concentration. The tropospheric vertical column density is inverted using a radiative transfer model. This work specifically adopts the QA4ECV NO_2 MAX-DOAS reference datasets, which includes 10 sites. The sites are categorized into three types (Sub-urban, Urban and Rural) based on their location, and 3 sites of different types are used here. The information of the 3 sites is listed in Table 2.

Table 2: Information of MAX-DOAS sites.

Station	Latitude	Longitude	Range of NO_2 Observations (molec.cm^{-2})	Time Zone	Data Used
Uccle	50.8°N	4.4°E	0-26×10^{15}	0	2011.04-2015.06
OHP	43.9°N	5.7°E	0-7×10^{15}	0	2007.01-2016.12
Xianghe	39.8°N	117.0°E	0-59×10^{15}	UTC+8	2010.04-2017.01

2.2.3 Reanalysis meteorological data

Reanalysis combines model data with observations from across the world into a globally complete and consistent dataset using a model of the atmosphere based on the laws of physics and chemistry. For this reason, this work uses the fifth generation ECMWF reanalysis (ERA5) for 12 specific meteorological parameters as given in Table 1. The dataset used has an hourly

temporal resolution and a 0.25°×0.25° spatial resolution. Those meteorological products in this work are used at the following pressure levels: 100 hPa, 200 hPa, 500 hPa, 700 hPa, 850 hPa, 925 hPa, and 1000 hPa. The actual weightings used in this work are computed using principal component analysis (PCA), and rely nearly fully on the data from 850hPa and below, at roughly equal weights.

This study also uses the fourth generation ECMWF reanalysis (EAC4) specifically for its modeled NO_2 column values, which are used as a means of comparison against the NO_2 fields generated within this work. EAC4 data has a spatial resolution of 0.25°×0.25°, and a temporal resolution of 3 hours. In this work, a vertical column density of tropospheric NO_2 is derived from EAC4 and is used for comparison.

2.3 XGBoost (eXtreme Gradient Boosting) algorithm

A gradient boosting framework is used by the decision-tree based ensemble machine learning approach known as XGBoost (Chen and Guestrin, 2016). This method employs a more regularized model formalization than other techniques (Cisty and Soldanova, 2018; Zhang et al., 2018), having greater control against overfitting compared with gradient boosting decision tree (GBDT) approaches (Dong et al., 2022). Similar to the random forest algorithm, XGBoost needs its hyperparameters tuned (Kapoor and Perrone, 2021). It has a more intricate structure and adds regularization components to the loss function to prevent overfitting so that it can handle complicated data better. Therefore, XGBoost is a better option for working with vast volumes of data and multidimensional affecting factors like NO_2 gap filling. Additionally, XGBoost has been used to estimate pollutants, and its results outperform those of certain other statistical and machine learning methods (Reid et al., 2015; Just et al., 2018; Zhai and Chen, 2018; Fan et al., 2020). Table 1 shows the data used in this research, which are input into the machine learning model.

2.4 SHAP (SHapley Additive exPlanation) values

SHAP values quantitatively represent the conditional expected value function of the machine learning model, implying the average contribution of a feature to a prediction (Lloyd Shapley, 1952). The use of a black box model, such as XGBoost in this work, necessitates an explanatory model in contrast to interpretable algorithms (i.e. Cohen and Prinn, 2011). According to each feature's marginal contribution, SHAP distributes the overall gain, in terms of both negative and positive contribution. In this work, SHAP values are used to quantify the importance of features, as shown in Figure 2.

2.5 DINEOF method

DINEOF is used in this work to reconstruct the missing points in the spatio-temporal field of NO_2. This method relies on an empirical orthogonal function (EOF) decomposition in space and a principle component (PC) decomposition in time that identifies spatial-temporal domains of maximal variation following (Cohen, 2014). The method allows the assignment of a prediction under conditions in time and/or space that are missing observational data. By using the weighted EOFs and PCs in an iterative manner, missing data points can be re-synthesized based on a weighting of the various underlying orthogonal basis functions. The number of iterations which minimizes the cross-validation error is used to obtain the best-reconstructed data. For a more detailed description of the overall approach, see Beckers and Rixen (2003) and Alvera-Azcarate et al. (2005). In this work, the amount of data filled using this approach ranges from 27% to 35% on a year-by-year basis, as given in Table 4.

2.6 Validation strategy

In order to analyze the performance of the reconstructed dataset, this work not only uses cross-validation based on the original data itself, but also refers to the observations from TROPOMI, MAX-DOAS, and the EAC4 reanalysis product mentioned above. The root mean square error (RMSE), Pearson correlation coefficient (R), normalized mean bias (NMB) and mean absolute error (MAE) are all used in the validation process.

Also, as an important and innovative approach, EOF is performed on the three-dimensional observed and HSTCM-NO_2 fields. And these values are compared to ensure that the maximum changes in spatial and temporal signal are consistent with the original observations. EOF is an exploratory technique for multivariate data, which is in essence an eigenvalue problem, aiming at explaining and interpreting the variability in the data. Till now, EOF has been introduced into data analysis of satellite-based remote sensing to estimate the spatiotemporal distribution characteristics of pollutants such as HCHO (Kim et al., 2014), CO (Baek & Kim, 2011), aerosols (Cohen et al., 2017) and NO_2 (Li et al., 2023).

2.7 Method selection

The goal of this work is to use all available day-by-day and pixel-by-pixel NO_2 column data from both OMI and GOME-2 in tandem to reconstruct a consistent global NO_2 column product with the highest possible coverage. Machine learning used in this work can only predict OMNO2 data which also has GOME-2 data at corresponding position in space and time. For this reason, this work introduces DINEOF to reconstruct data in locations where both of OMI and GOME-2 do not have values, but where data exists at other times or nearby locations in space.

Since DINEOF and machine learning have not previously been used in tandem for this type of issue, a critical component of the methodology is to quantify the impact of using the two approaches individually, in tandem, and if in tandem in what order. To first determine which sets of methods are best suited for this work, a subset of data from 2007 is selected. Furthermore, due to the issue of the row anomaly, a second comparison dataset from 2013 is used as a mask. In this way, data from 2007 which are masked by data from 2013 will be separated for validation, and the missing data will be the major difference assuming the changes in the climatology are not significant. Therefore, the following methods are applied, as displayed in Fig. 1:

I. First XGBoost is used to predict OMNO2 data based on GOME-2 data. Subsequently, DINEOF is applied to fill the remaining gaps.
II. DINEOF is first used to fill the gaps in GOME-2 data. This is then followed by XGBoost prediction based on the reconstructed GOME-2 dataset.
III. DINEOF is used solely to fill in gaps in OMNO2.

The reconstructed dataset is evaluated based on comparison between the masked data from 2007 and the results. Additionally, in order to verify whether and how the absence of GOME-2 values affects prediction accuracy, further partitioning of the dataset based on the presence or absence of GOME-2 values is performed. All results are given in Fig. 1, where the row is the method and the column is the amount of GOME-2 data used.

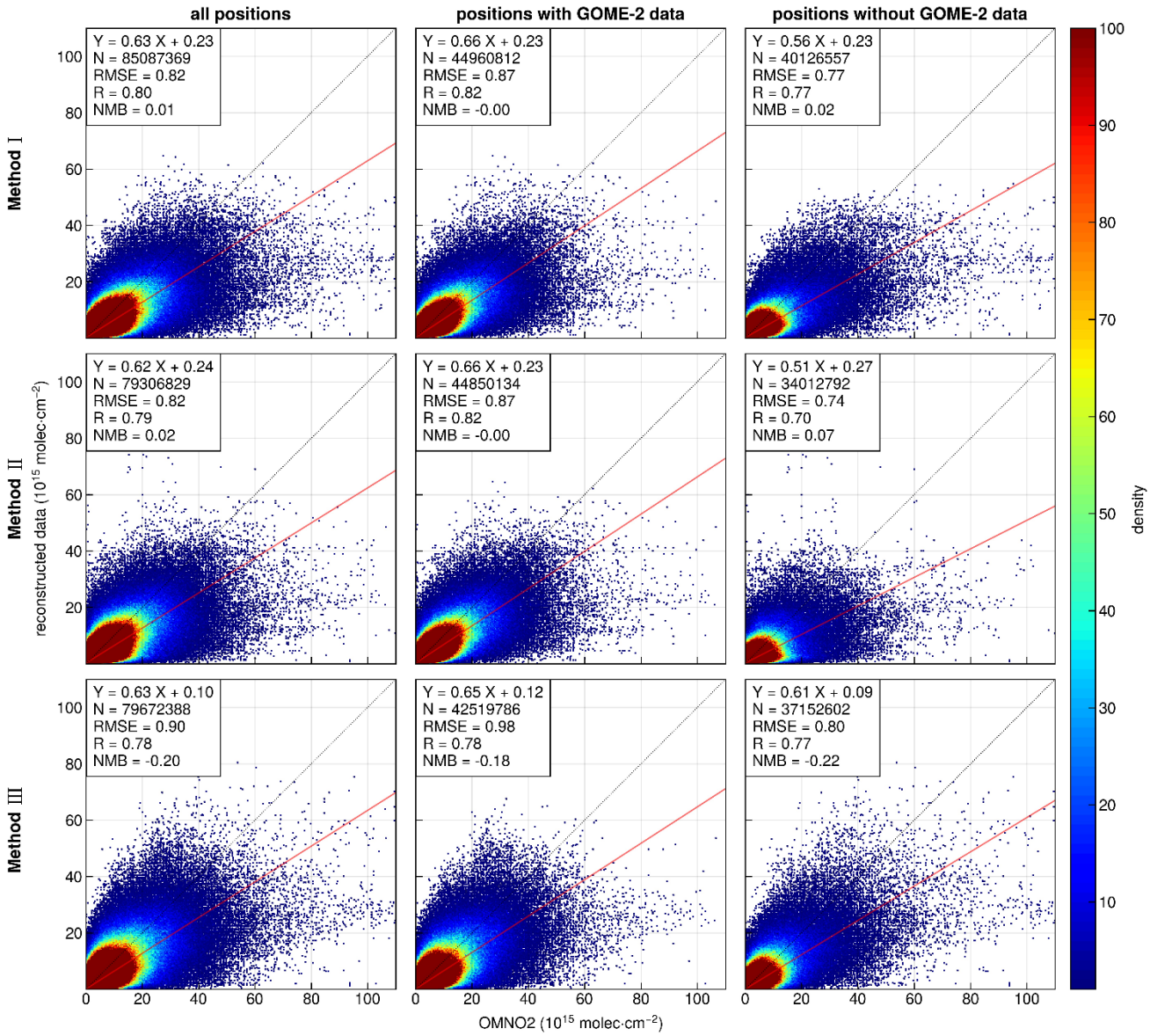


Figure 1: Cross-validation of 3 methods between reconstructed data and masked OMNO2 in 2007.

As shown in Fig. 1, the complete or partial datasets reconstructed by Method I all have the maximum R value and the minimum RMSE and NMB in the same scenario. Meanwhile, by comparing Column 2 and Column 3, it is obvious that the presence of GOME-2 observations can greatly improve the accuracy of reconstruction and have an impact on the fitted slopes (especially in the cases of methods I and II). From Row 3, it can be found that DINEOF has universality, but does not have outstanding performance. Therefore, it is necessary to use machine learning for prediction in positions with values obtained from GOME-2 and DINEOF only used for filling in positions that do not contain GOME-2 data. In conclusion, in order to obtain the optimal results, Method I will be chosen as the reconstruction scheme in this work, which is consistent with the idea that using the most amount of actual observational data possible best supports the machine learning approach.

3 Results

3.1 Reconstruction process and model evaluation

3.1.1 Quantifying the importance of individual features

The results of the SHAP value and its statistics are given in Fig. 2 based on global training data from February 2019 as an example.

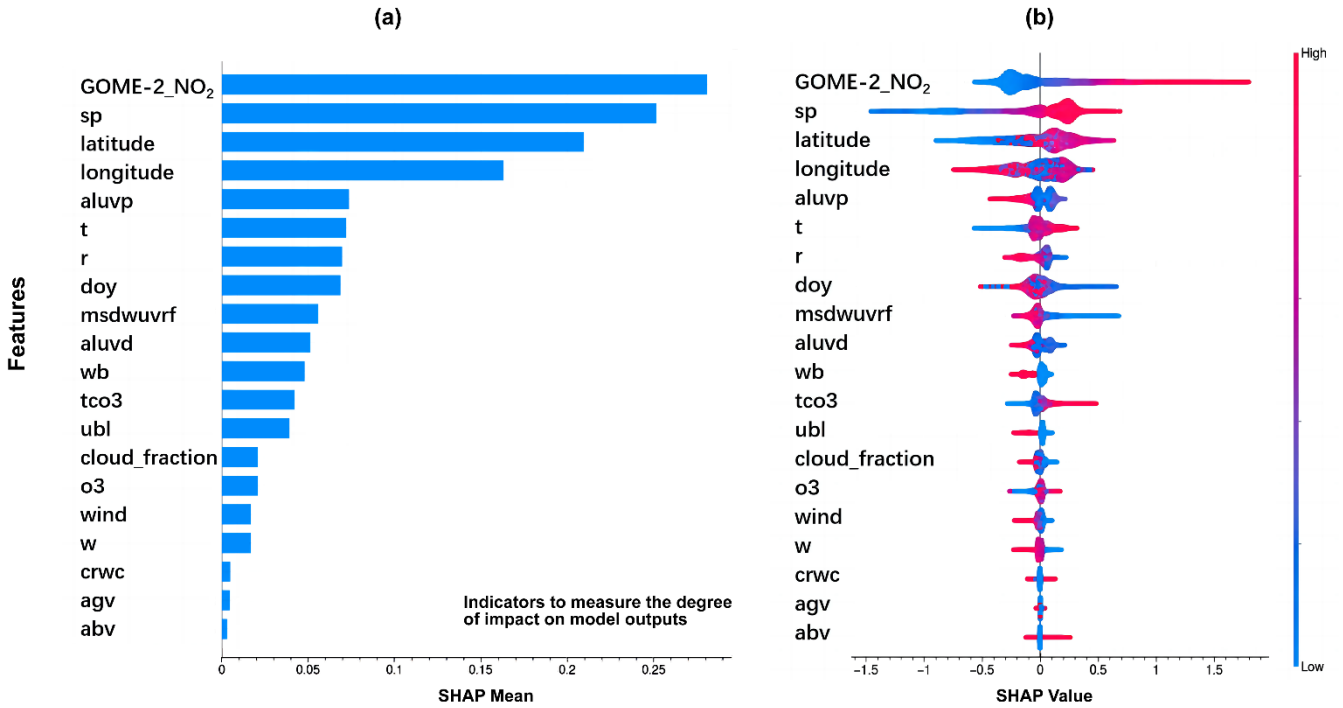


Figure 2: Feature importance ranking **(a)**, and scatter plot of feature density for each parameter of XGBoost **(b)**, represented as a beeswarm. In specific, each row represents a feature, and the order of arrangement is determined by the importance of the feature calculated in the previous step. The horizontal coordinate is the SHAP value, where the sign of the value indicates the direction of the contribution of that feature. Each point in each row represents a single sample, and the color of the point indicates the magnitude of the feature value (high in red and low in blue).

The 20 features with the highest contribution are provided. Data from GOME-2_NO_2 has both the highest overall mean contribution, as well as the largest absolute contribution (up to 1.8), which is larger than the absolute values of all other contributing factors, as well as the only significant source in terms of positive contribution (greater than 0.6). This result is consistent with the fact that GOME-2_NO_2 is the base observation upon which the machine learning is acting. The second most significant driving feature is the surface pressure, which has both the second highest mean and the second largest absolute contribution (down to -1.5) of any factor. This is consistent with the fact that human settlements tend to occur at lower elevations in general and that changes in pressure tend to accompany changes in the rates of transport and chemical activity of NO_2 in situ (Wang et al., 2020; Li et al., 2023). Below this there are some interesting patterns in which some species contribute more to the mean SHAP, but not necessarily to the extreme SHAP values, meaning that the global and local contribution factors are different in different locations. As expected, latitude, longitude, day of year, and downwelling UV radiation are all relatively important in different areas, which is consistent with the highly heterogenous nature of NO_2 emissions, different driving forces which impact the ratio of NO to NO_2 emissions within NO_x, issues of geospatial change, and processing once the NO_2 is in the atmosphere, among other factors. These factors are sufficient to capture the presence of pollution sources within specific pixels, and therefore it is required to not only be able to predict the long-term signal, but also account for short-term changes of a sudden nature as well.

3.1.2 Separation of models over ocean and land

Globally, the distribution of NO_2 observed by satellites is not balanced, due to the fact that NO_2 has a relatively short lifetime, and the vast majority of its emissions occur over land in and around areas of anthropogenic disturbance. Furthermore, if major shipping lanes and areas of significant downwind transport are excluded, NO_2 generally has lower values over the sea compared to land. On top of this, the surface absorption profile over the oceans is different from land, which may further contribute to differences in the column interpretation. This section quantitatively explores the impact of separating those pixels over the sea from those over the land in terms of training the machine learning model, and works to quantify any reduction in the overall error rate of the models between the separated and unified approaches.

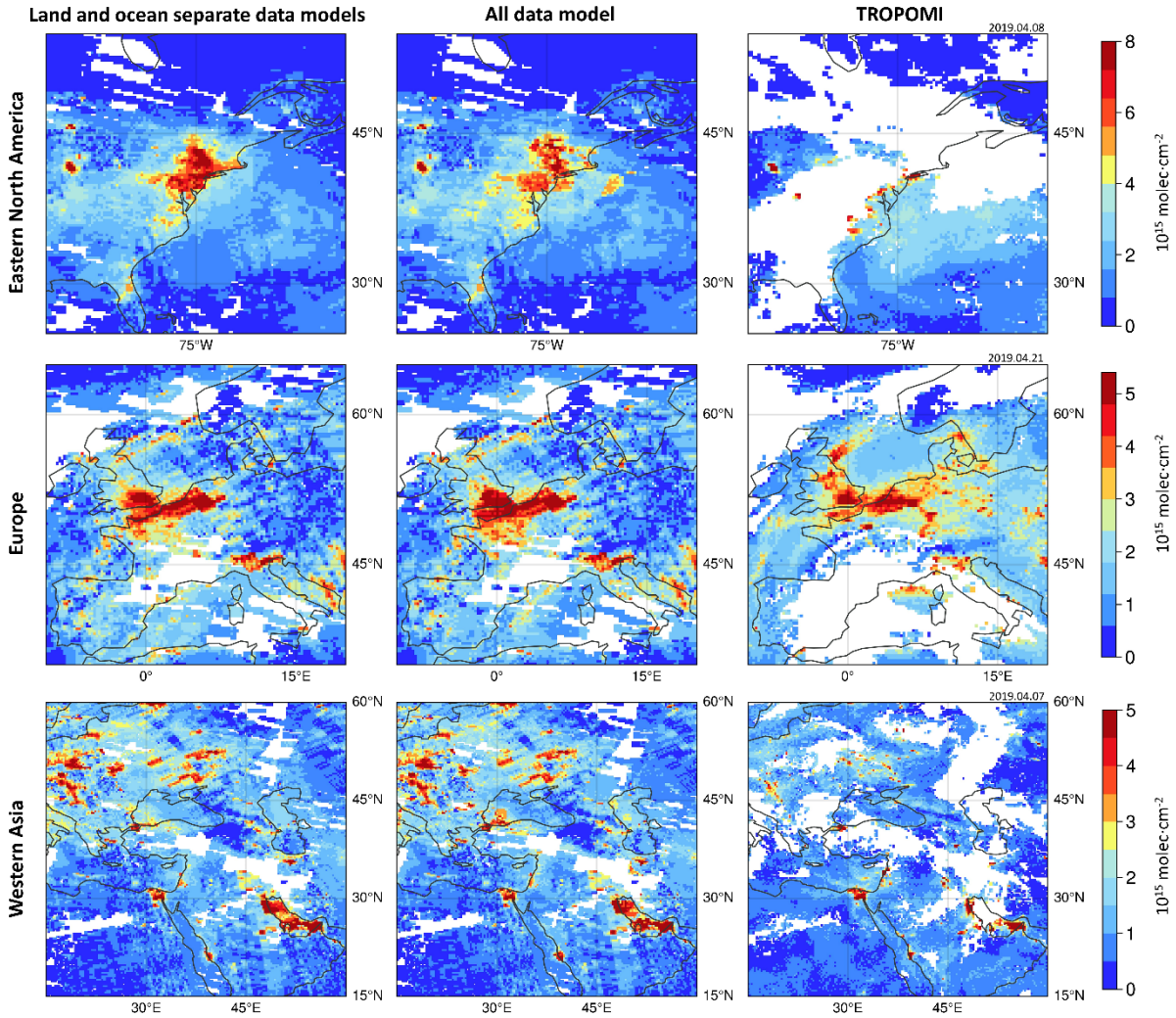


Figure 3: Prediction results with and without prior knowledge, versus TROPOMI observations over 3 regions (Eastern North America, Europe and Western Asia).

The effects of separating the land from the ocean models are demonstrated clearly over April 2019 in Fig. 3. First, the high values of NO_2 observed over the Western Atlantic Ocean found in the all data model are no longer observed in the land and ocean separate data models, which is consistent with TROPOMI NO_2 observations. Over Western Europe, the high values off of Scotland as confirmed by TROPOMI still remain in the land and ocean separate data models' case, while the unusually high values in the all data model case are reduced to more reasonable values compared to the observations from TROPOMI over the areas between Spain and France and between the UK and France. Even with the separation, there are still erroneously high values between the UK and Ireland, and in the Eastern Atlantic which are not resolved. The third row shows the distribution of NO_2 concentrations in Western Asia. In the TROPOMI observations, high values are observed on the southwest side of the Arabian Peninsula only over land, and mostly on land over Northern Turkey except for the Bosphorus Straits, which is consistent with what is understood, and which the separate land and ocean data models are able to capture, while the all data model misrepresents these values as being higher than the observations support. Overall, there is a considerable improvement observed over near-sea areas, in terms of both retaining enhancement where it is justified and reducing enhancement where it is not justified by using the separately trained models. However, there are still inconsistencies which are not resolved.

3.1.3 Evaluation of machine learning

After applying XGBoost and prior knowledge mentioned above, Fig. 4 demonstrates the reconstructed results and compares them with the original data on a pixel-by-pixel basis in 2007.

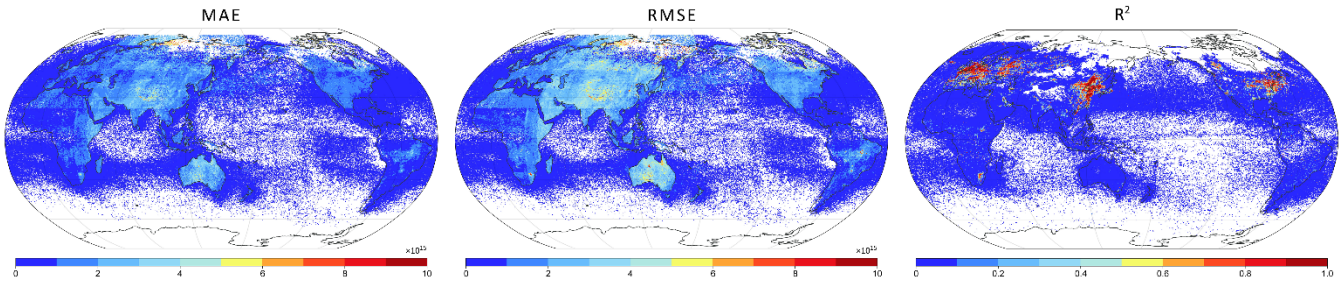


Figure 4: Accuracy validation of XGBoost prediction results (MAE, RMSE, and R^2).

Among the results predicted by XGBoost, the MAE and RMSE of the results located over water are slightly lower than those located over land. The predicted results over Eastern Asia, Europe and Eastern North America show a higher correlation with the observations, indicating that the variability is captured better over regions where the vertical column density of NO_2 is larger. For these reasons, the machine learning model is trained separately over both land and over the ocean, with training done on a month-by-month basis. The results of this fitting are given in Fig. 5, demonstrating the time series of the statistics from 2013 to 2015.

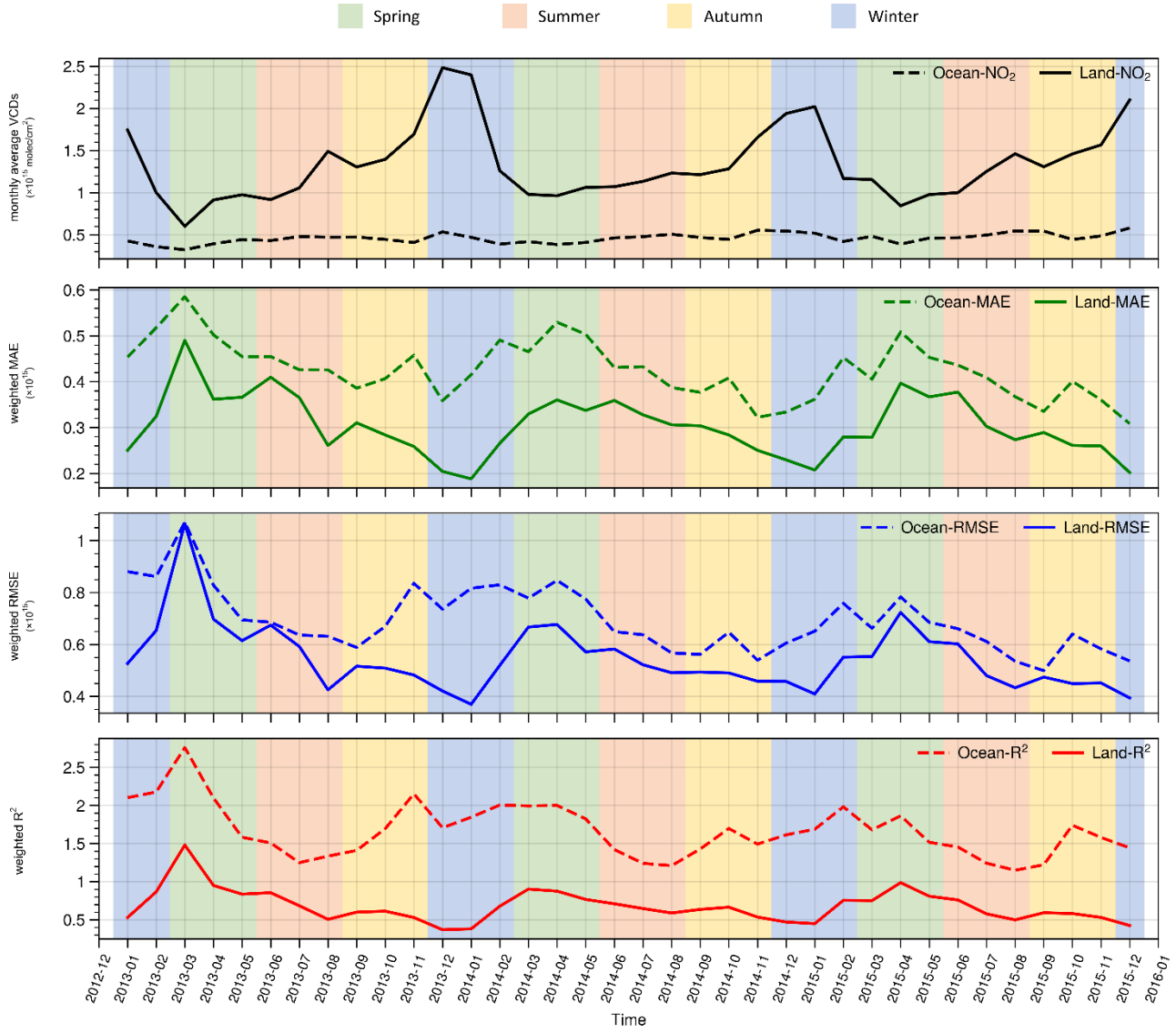


Figure 5: Land and ocean model quality of XGBoost from 2013 to 2015.

As demonstrated, the RMSE and MAE of the ocean model are both always higher than, and less temporally variable than those of the land model, also, unlike the land model, which shows improved performance during the winter, the ocean model does

not experience this seasonal improvement. This indicates that these errors scale both in terms of the magnitude, which is higher over land, but also in response to the retrieval algorithms themselves, which have a different amount of error over bright and dark surfaces. Additionally, the abundance of surface-based measurements over land has initially enhanced the accuracy of these retrievals. The correlation over time of the land model is slightly higher than that of the ocean model, indicating that the data predicted by the land model may have a lower uncertainty, possibly due to a better apriori data, a better-defined AMF over land, or due to the overall retrieval being better over land as compared to water (Richter et al., 2011, Streets et al., 2013, Lamsal et al., 2021).

The quality of the land model fit shows a strong decrease in quality over a period of 1 to 3 months every year, experiencing both inter-annual and intra-annual variation, while the ocean model shows a weaker decrease in the fit for a few months in two of the years, and no change in the other years. This indicates that there must be a few different forces acting upon the fits, including some of which are clearly seasonal in nature with only small variation (air temperature), while others are more variable (UV radiation and AAOD), consistent with the results from Wang et al. (2020), and Li et al. (2023). On the one hand, the UV intensity is generally lowest in December and January, leading to an increase in the residence time of NO_2 in the atmosphere, and generally highest in May and June, leading to a decrease in the residence time. However, the UV itself is also modified by the effects of both clouds and absorbing aerosols. Cloud coverage tends to affect a larger percentage of the ocean surface compared to land. However, absorbing aerosols have a more significant impact over land, which contributes to the findings mentioned earlier. The effects of temperature tend to peak differently from those of UV radiation, but these effects tend to be climatologically more similar year to year, given that the years analyzed do not contain any El Niño or La Niña types of patterns. In addition to this, the vertical column density of NO_2 itself also changes from month to month with the peak values over land occurring in December and January, with both the magnitude of the peak and the peak month varying from year to year. This allows for a greater amount of differentiation between the heavily polluted and more clean regions during this time, especially so over land. As discussed previously, such high variability may lead to additional machine learning fitting issues. On the other hand, there is generally less cloud during the winter, meaning more observations on a day-by-day basis, as well as more atmospheric stability in the winter, leading to less vertical and long-range transport of pollutants away from their source regions. The combination of all of which enables the model to achieve more accurate predictions.

3.1.4 Reconstruction process and accuracy analysis of DINEOF

The EOF separates the data into its primary basis functions, of which there are spatial and temporal components. To test the efficacy of the EOF procedure as a function of the time length of data used, this work has run the procedure over different time periods from a minimum of 1 month of data to a maximum of 3 years of data. The annual data, as shown in Table 4, yields the lowest overall standard deviation. This is consistent with the above results showing that there is a clear annual peak in the NO_2 columns occurring each winter, and indicates that this amount of variability drives the model more than the smaller year-to-year changes in the peak or overall characteristics of NO_2. This result is consistent with a year (intra-annual variability) tending to be smaller than year-to-year variability unless a very long time series is considered (minimum of 20-30 years) (Chowdhury, 2022), unless capturing a known extreme such as El Niño, La Niña, etc. (Deng et al., 2021). Based on the timing chosen and the results below, this work will rely upon applying the DINEOF reconstruction of the dataset on a year-by-year basis.

Table 4: Reconstruction results of different time lengths of DINEOF.

Time Length	1 month	3 months	6 months	1 year	3 years
Start Time	2008.01.01	2008.01.01	2008.01.01	2008.01.01	2008.01.01
End Time	2008.01.31	2008.03.31	2008.06.30	2008.12.31	2010.12.31
Image Number	31	91	182	366	1090
Missing Rate	34.2%	27.7%	29.1%	29.3%	32.0%
Mean Value ($\times 10^{15}$)	0.60	0.54	0.54	0.56	0.58

Standard Deviation ($\times 10^{15}$)	1.57	1.35	1.15	1.12	1.13
Iterations Made	112	50	16	12	12
Convergence Achieved	10.0E-4	10.0E-4	9.8E-4	8.61E-3	9.9E-4

Table 5 shows the DINEOF results for each year, with most years achieving convergence after 12 to 29 iterations. The standard deviation is shown to be lowest when analyzing data one year at a time. Interestingly, the year 2009 saw the most data loss, with more than one-third of the total data (34.3%) lost. This indicates that both the geospatial nature of the data and the range of column loading values are important factors, in addition to the absolute amount of data reconstructed.

Table 5: Statistics of DINEOF reconstruction results by year.

Year	Mean Value ($\times 10^{15}$)	Standard Deviation ($\times 10^{15}$)	Iterations Made	Convergence Achieved	Missing Rate
2007	0.56	1.18	12	9.2E-04	31.1%
2008	0.55	1.12	12	8.5E-04	29.3%
2009	0.58	1.07	16	9.7E-04	34.3%
2010	0.59	1.17	16	9.4E-04	32.5%
2011	0.59	1.21	14	8.8E-04	33.1%
2012	0.59	1.21	13	9.4E-04	30.7%
2013	0.59	1.16	12	9.1E-04	23.0%
2014	0.58	1.05	14	9.3E-04	23.5%
2015	0.58	0.98	22	9.6E-04	22.6%
2016	0.60	0.91	18	9.9E-04	23.9%
2017	0.59	0.93	22	9.3E-04	32.2%
2018	0.58	0.92	29	9.9E-04	23.2%
2019	0.59	0.59	25	9.9E-04	21.2%
2020	0.58	0.86	42	9.8E-04	21.4%
2021	0.64	0.93	64	9.9E-04	22.9%
2022	0.68	0.88	23	9.6E-04	21.3%

3.1.5 Overall analysis

The performance of reconstruction can be tested by simulating the known "row anomaly" issue, wherein OMI started to lose access to specific camera angles on a swath-by-swath basis starting in 2007, leading to the appearance of missing lines of data. Since the data is otherwise in good order, a well-conditioned filling method should be able to produce data to cover these well-known and geometrically simple gaps. Five regions (East Asia, Europe, Eastern North America, South Africa, and South Asia) are used to demonstrate the effectiveness of the procedure to fill these gaps on different days in 2007.

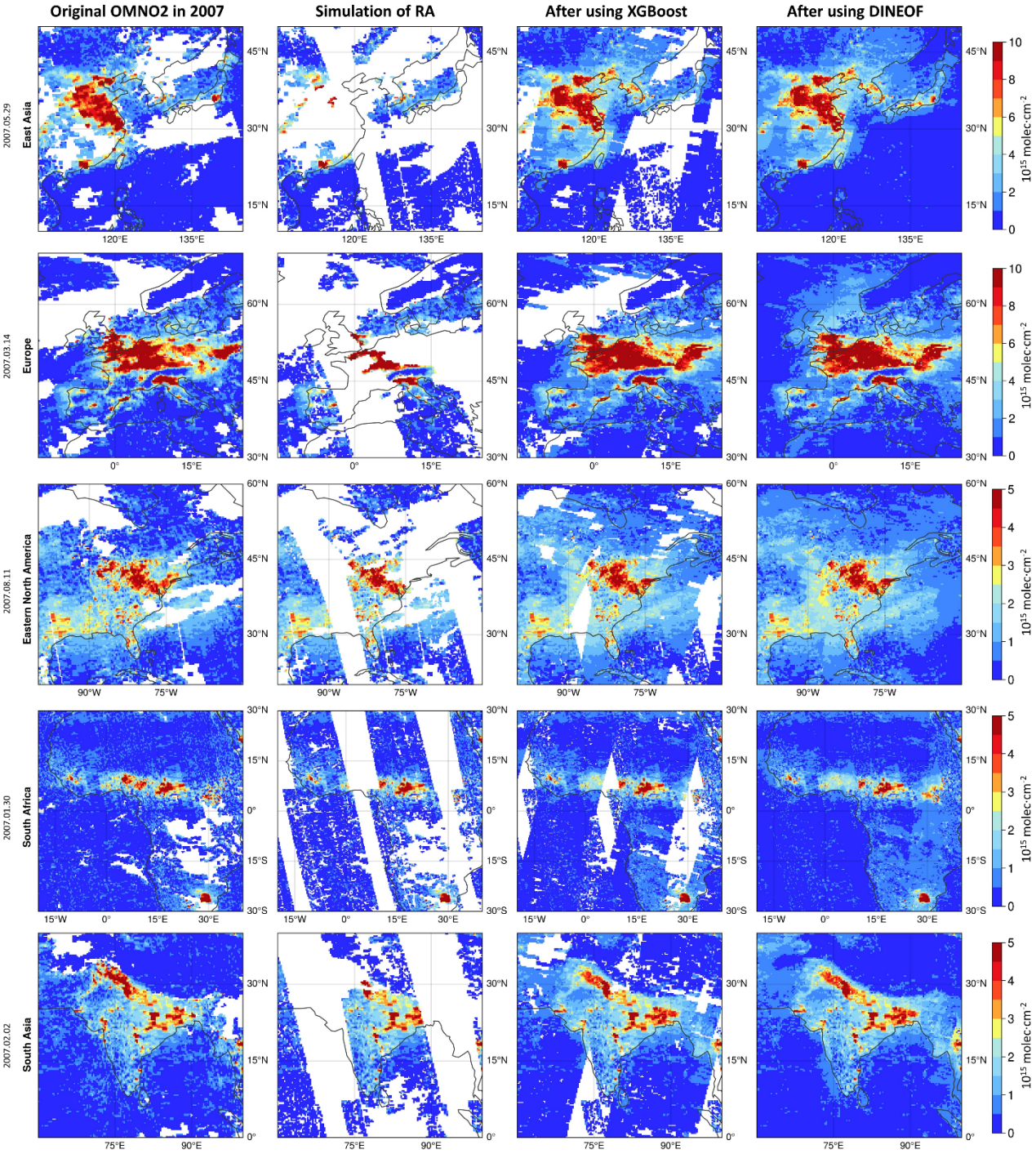


Figure 6: Results of stepwise reconstruction of masked data over 5 regions (East Asia, Europe, Eastern North America, South Africa, and South Asia).

As shown in Fig. 6, the first column shows the original OMNO2, and the second column shows the data distribution after simulating the effect of "row anomaly". The machine learning reconstructs the image of GOME-2 at positions with value, keeps the original observations, and only reconstructs the missing parts. After reconstruction by XGBoost, the image elements that are still missing are reconstructed using DINEOF to obtain a dataset with more than 99% coverage. Comparing the reconstructed data with the original data, it can be found that the reconstructed results are basically consistent with the distribution of the original OMNO2, with the following two exceptions: some very high pixels observed in the EU and USA have been removed and replaced with lower value pixels in the reconstruction, while some moderate and low pixels in China and India have been replaced with high value pixels in the reconstruction. In general, the overall shapes are reasonably similar and the transition from high to low values seems to make sense based on the values from the original OMNO2.

3.1.6 Coverage statistics

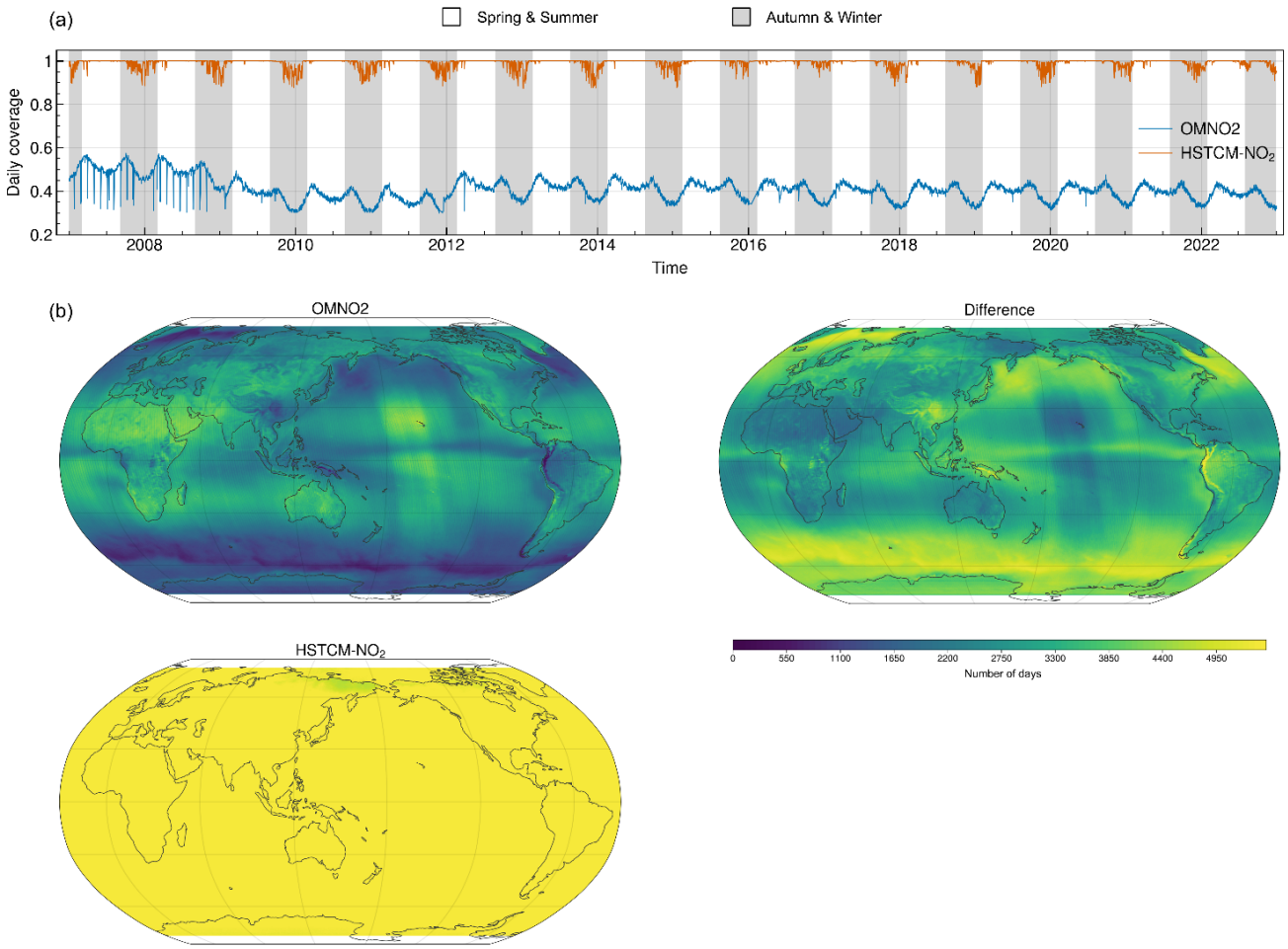


Figure 7: Coverage statistics of HSTCM-NO_2 from 2007 to 2022, daily coverage (0.3 is used as a cut-off) is shown in **(a)**, number of days with data for each pixel is shown in **(b)**.

The spatial coverage of the original OMNO2 declined from about 50% in 2007 to 35%-40% after 2009 due to the "row anomaly" phenomenon and cloud occlusion, and improved slightly in late 2012 (although not recovering to previous levels). The reconstructed data however has a daily coverage of over 90%. As shown in Fig. 7(a), the original data has more gaps when the cloud volume is higher and less data when the cloud volume is smaller. The reconstructed data also shows such a trend, although with a much smaller difference between the high cloud and low cloud periods of time, indicating that some fraction of cloud-covered data can be reconstructed successfully, while some other amount has so much data lost that even this technique used in this work cannot fully reconstruct the data.

Fig. 7(b) shows the comparison between the original OMNO2 and the HSTCM-NO_2 in terms of spatial distribution. OMNO2 in the eastern part of North America, northwestern part of South America, Europe and southeastern part of Asia is obviously missing, although after reconstruction all the data in the above locations are reconstructed. The reliability of their HSTCM-NO_2 is verified over such land-based and near-land areas. There are a few exceptions, such as perpetually cloud-covered areas in the North Pacific and along the equator, but in these cases, there is likely no possible solution since they are covered for days in a row over huge spatial areas. Globally on average, the 39.5% coverage of OMNO2 increases to a 99.1% coverage of HSTCM-NO_2.

3.2 Multi-source validation of HSTCM-NO_2

3.2.1 Comparison with MAX-DOAS data

The original OMNO2, HSTCM-NO_2 were validated against MAX-DOAS and the following results were obtained. The 3 sites

used in this work are Xianghe, Uccle and OHP, which are located separately in Sub-urban, Urban and Rural.

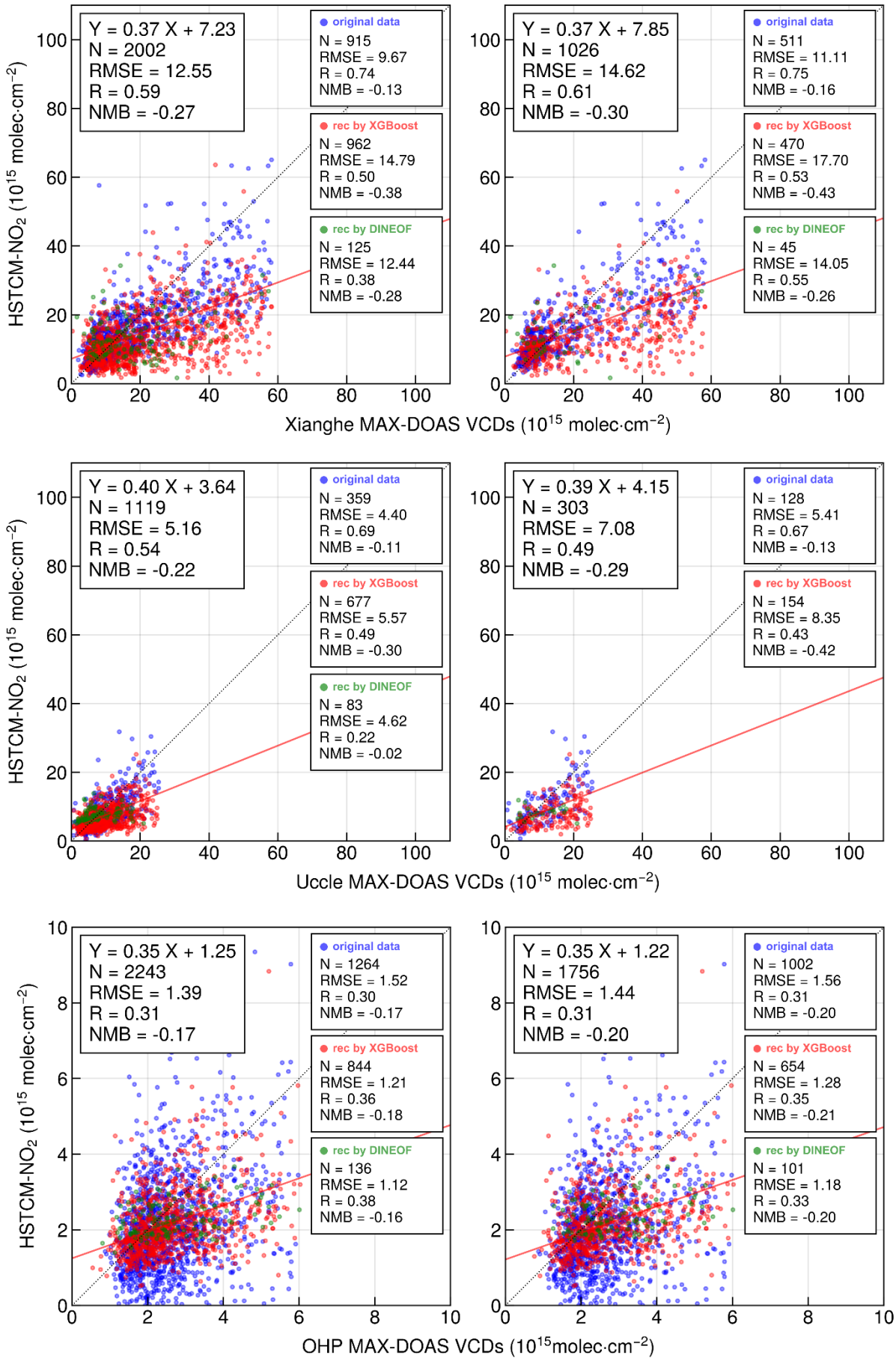


Figure 8: Scatter plots of comparison between MAX-DOAS observations (Xianghe, Uccle and OHP) and HSTCM-NO_2, the figures of the left panel use all observations of MAX-DOAS, while of the right panel are filtered out extreme cases. The boxes in the upper left corner summarize the statistical comparisons, while the boxes to the right of each subfigure represent the statistics of each individual reconstruction step.

Comparisons between the various different products and MAX-DOAS are shown in Fig. 8. Due to the small amount of data, there is a missing box which corresponds to a result that didn't pass the p-test. In all cases, there is a sufficient number of data points to consider the fits under both all data and extreme event filtered data conditions. At the site with very high NO_2 column loading (Xianghe) and moderately high NO_2 column loading (Uccle), the results using both XGBoost and DINEOF together are still less good than the original OMI data, regardless of whether the data is filtered or not filtered. In Xianghe this difference is even larger than in Uccle, confirming that the approaches employed here do not work very well when a substantial amount of data is located at or above 6×10^{15} molec.cm^{-2}. However, it is clearly shown even at these high sites that using both XGBoost and DINEOF together yields a final product which is more representative of OMI than using only one method independently. In the case of Xianghe, using all data with XGBoost alone, or using filtered data with either XGBoost or DINEOF alone yields similar results, which are worse than applying both XGBoost and DINEOF in tandem. At Uccle, applying XGBoost on its own always yields a result with a much higher R coefficient and a lower RMSE coefficient than when DINEOF is applied on its own, consistent across both filtered and unfiltered data. Interestingly under the relatively cleaner conditions found at OHP, the results of applying both XGBoost and DINEOF together yield a result which is better than the result of OMI in terms of RMSE and similar in terms of R. The application of either XGBoost or DINEOF independently at this location yields results which are quite good when compared with OMI. This set of results makes it clear that under cleaner conditions, the use of one or both of XGBoost and/or DINEOF yields benefits and can be considered trustworthy, while their combination yields a large amount of additional data and still works well. Clearly the benefits of the gap filling and prediction are consistent with the observations under these conditions, allowing conclusions observed above under different polluted conditions to be further supported.

3.2.2 Comparison with TROPOMI

Compared with OMI, TROPOMI has a higher spatial resolution and wider swath angle, allowing improved spatial observation of tropospheric NO_2, with the caveat that higher resolution may mean that some pixels are cloud-covered, whereas at lower resolution this may not be the case. For these reasons, TROPOMI NO_2 is used as an external data source to allow comparison with the various products and to serve as a means for ensuring that the derived products are reasonable.

The spatial distribution on 4 specific days in 2019 and annual average coverage over East Asia, South Asia, Europe and North America are used to compare OMNO2, HSTCM-NO_2 and TROPOMI NO_2, as displayed in Fig. 9.

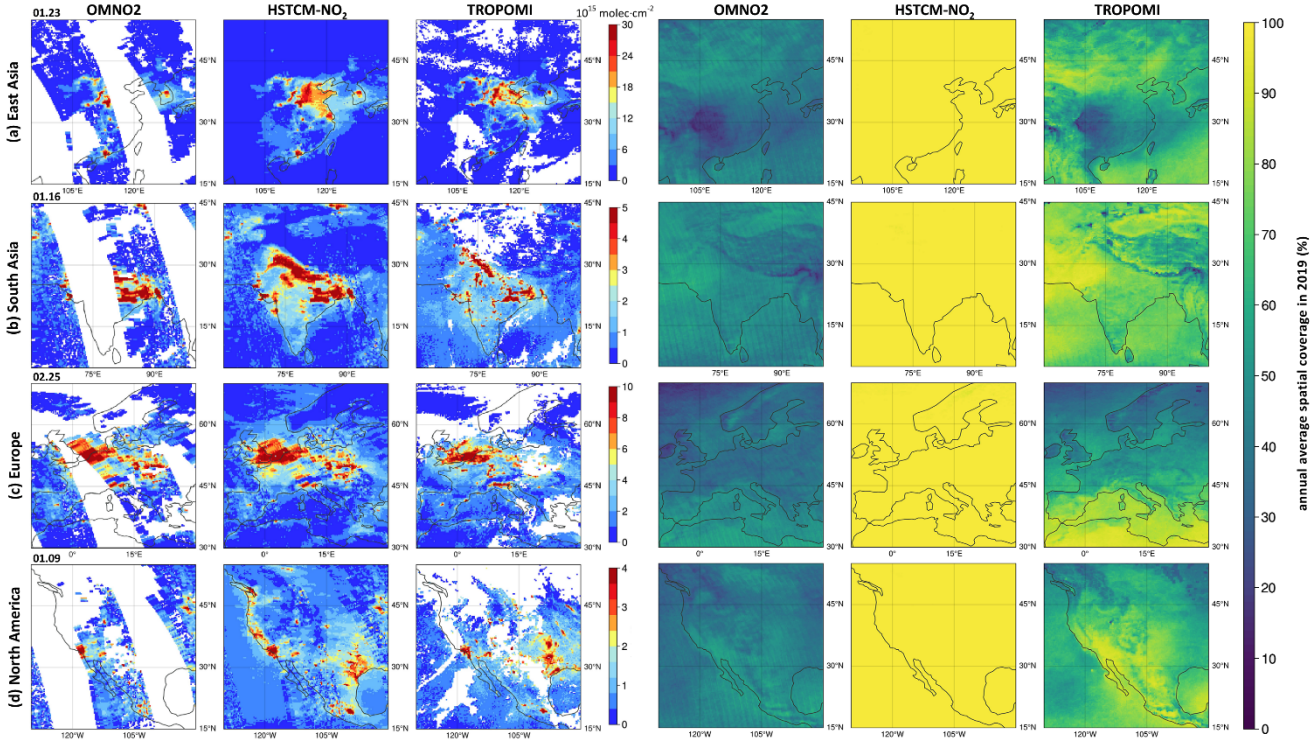


Figure 9: Distribution and coverage statistics of OMI, HSTCM-NO_2 and TROPOMI over East Asia **(a)**, South Asia **(b)**, Europe **(c)** and North America **(d)** in 2019.

As shown in Fig. 9(a), the coastal areas of China are severed affected by the RA, leading to a significant portion of data missing in OMNO2. The reconstructed results of HSTCM-NO_2 are similar to TROPOMI on average, and in particular in Hebei, Henan, Shanxi, Shaanxi, parts of industrial Inner Mongolia, the Pearl River Delta, and even the transport corridors between China and South Korea in the East China Sea. However, there are some regions in Shandong, southern Jiangsu, Wuhan, and Shanghai, where the characteristics on average may be acceptable, but where high and low values are too smoothed over and extremes are not well predicted by HSTCM-NO_2 as compared with TROPOMI. Due to the effects of cloud cover, both OMNO2 and TROPOMI show no data over the megacities of Chongqing and Chengdu, while HSTCM-NO_2 effectively solves this problem in terms of large-scale spatial averaging, with a coverage of almost 100%. However, the fine-scale centers of the two cities are not clear in this case.

In Fig. 9(b), again due to the RA, OMNO2 lacks data over New Delhi, Lahore and other cities in central and western India. The reconstructed HSTCM-NO_2 products fill this part of data well, and the NO_2 distribution shown in HSTCM-NO_2 is similar to that of TROPOMI, with the major issue being that heavily polluted areas are more diffuse than in TROPOMI. In particular, the areas of Northeast India which are known to have seasonal fires this time of the year are reflected well in HSTCM-NO_2 but not in TROPOMI, possibly indicating that the information from the morning provided by GOME-2 identifies information which is missed by TROPOMI in the afternoon. The special geographical environment of the Qinghai-Tibet Plateau has led to both high cloud cover and significant surface reflection in the region. As a result, the coverage of OMI and TROPOMI products in the Qinghai-Tibet Plateau region is relatively low, and HSTCM-NO_2 is able to provide some amount of geospatial information, likely again from GOME-2, while filling the climatological gap.

As shown in Fig. 9(c), due to the influence of the marine climate, high coverage of cloud often occurs in the European region, which causes significant interference to satellite observations. The coverage of both OMNO2 and TROPOMI products in the European region is relatively low on this day. The HSTCM-NO_2 has effectively reconstructed missing data in the UK from Scotland through London, most of central France, and even into Algeria and Tunisia, while greatly increasing data coverage throughout Europe as a whole.

As shown in Fig. 9(d), missing data in areas such as the western coast of North America, Texas, and Oklahoma have been well

reconstructed. Due to the impact of RA, the spatial coverage of OMNO2 is lower than TROPOMI, and the coverage of both is not ideal in both high latitude and high altitude regions. Through the comparison of the four regions, it can be found that the HSTCM-NO_2 solves this problem, and has high consistency with TROPOMI NO_2. It works particularly well along the West Coast from San Francisco up through Vancouver, energy producing areas from Texas through New Mexico, and in general around urban and energy producing areas along the Eastern edge of the Rockies.

As shown in Fig. 10, each day and grid which contains a value of both HSTCM-NO_2 and TROPOMI NO_2 are compared in 2019. The comparison consists of a total of 171297320 pixels, and shows a reasonable fit globally with a RMSE of 0.64, R of 0.75, and NMB of 0.09. As pointed out elsewhere in this work, at values larger than 6×10^{15} molec.cm^{-2}, and especially so at values larger than 20×10^{15} molec.cm^{-2}, there are some small differences in the overall shape.

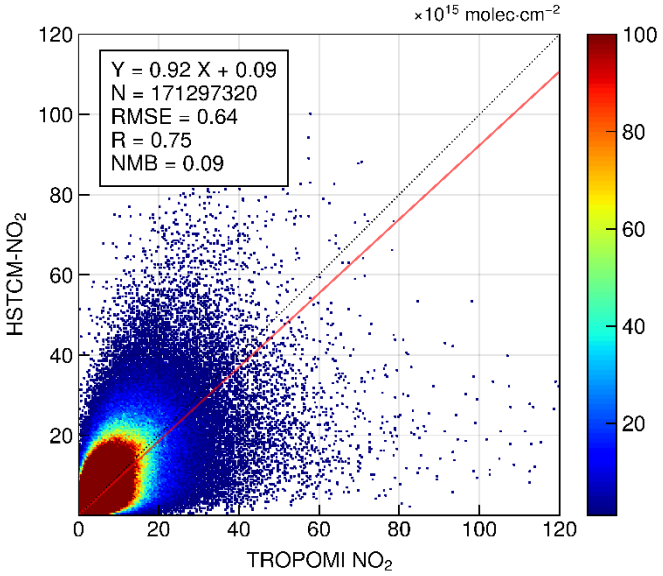


Figure 10: Comparison between global HSTCM-NO_2 and TROPOMI data in 2019.

3.2.3 Comparison with EAC4 data

Global results as well as results over three regions with a sufficient number of pixels with high NO_2 vertical column densities (East Asia, North America, and Europe) were selected to compare the reconstruction results with EAC4 data for February 2008. The reconstructed data at a global scale contains more than 2.6 million points and has an RMSE of 1.02, R of 0.77, and NMB of -0.25. Among the 3 regions, East Asia has the validation results with the highest R and the lowest NMB, followed by North America and Europe, as displayed in Fig. 11.

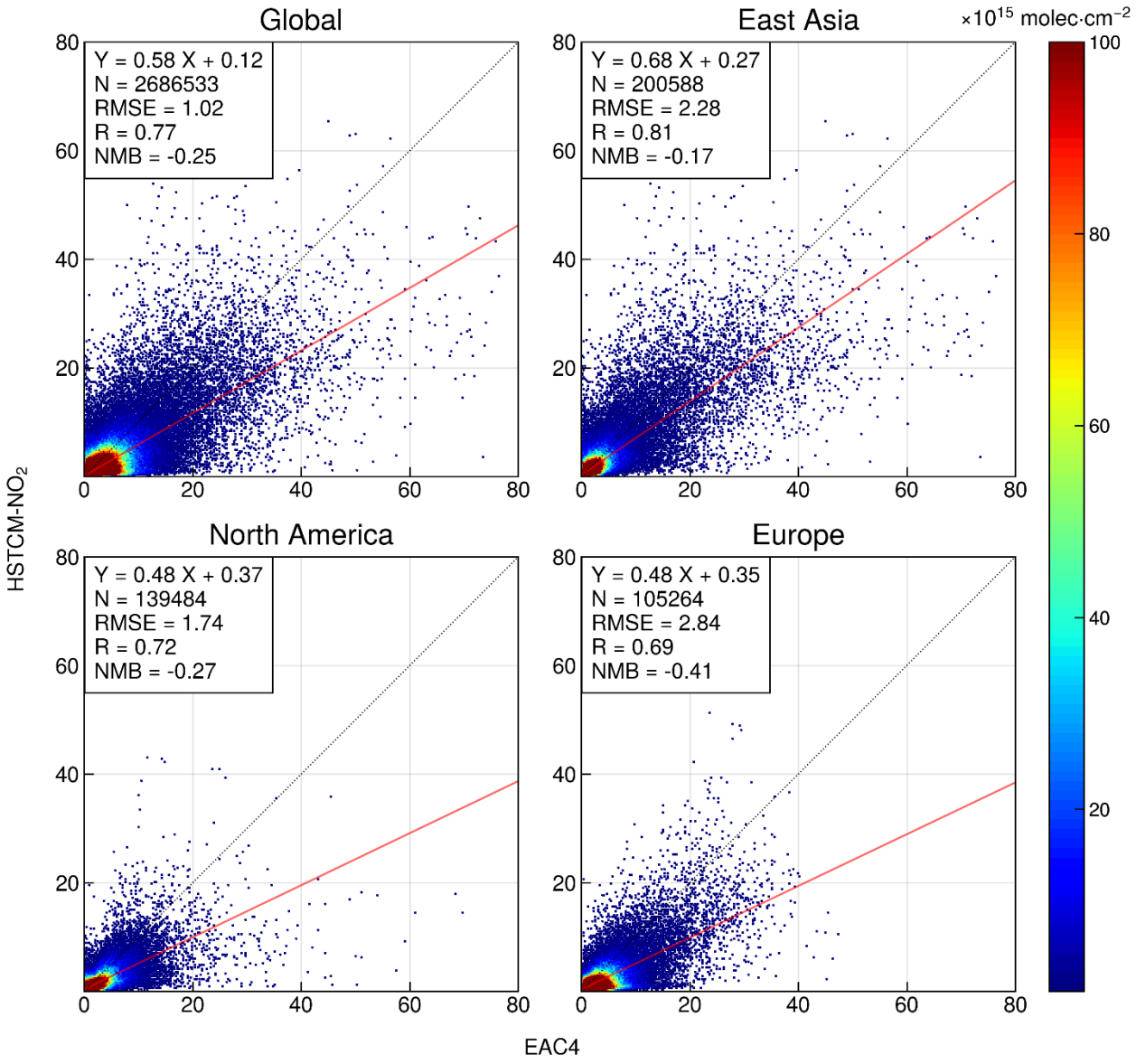
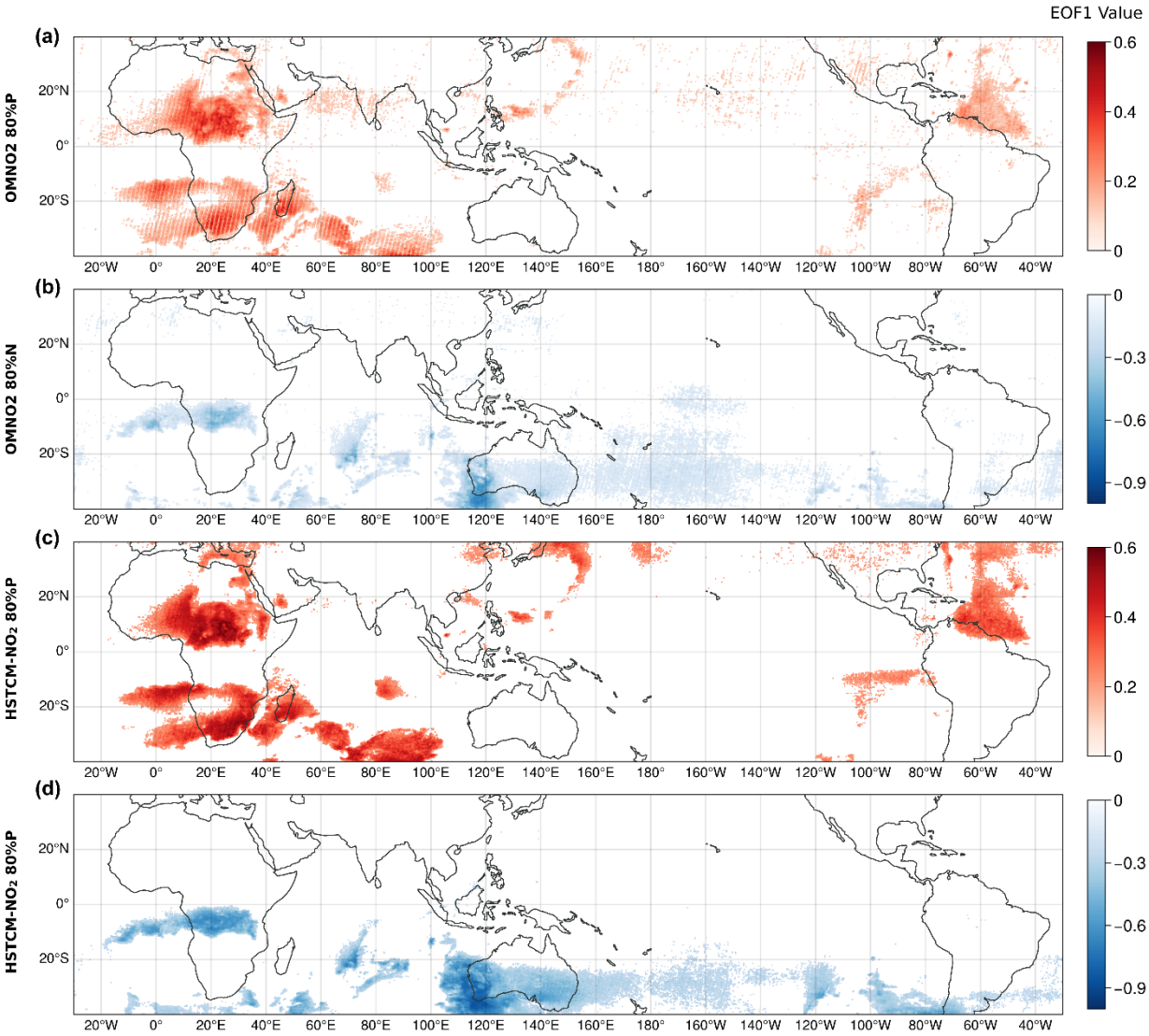


Figure 11: Global and regional (East Asia, Europe and North America) comparison between HSTCM-NO_2 and EAC4 data.

3.3 Results of EOF analysis

In order to verify the performance of HSTCM-NO_2, the temporal and spatial patterns are expected to match the observed variability. In specific, analysis was done over the time period from 2019-2021. The first three modes contribute 7.6%, 2.2%, and 2.0% of the total original OMNO2 respectively, while they contribute 26.1%, 4.0%, and 3.2% respectively for HSTCM-NO_2. This indicates that a spatial and temporal comparison using the first mode is sufficient to demonstrate the ability of HSTCM-NO_2 to reproduce OMNO2, given the fact that they both contribute more than the approximated global background 5% of error associated with the NO_2 retrieval itself. The contribution of HSTCM-NO_2's first mode to the total variance indicates that the reconstructed data is missing many finer modes of variability, however, as demonstrated below, the good spatial and temporal match shows that it is able to reproduce the signal reasonably well in actuality, with the major sources of this difference being regions north of 40°N and south of 40°S, both of which tend to be relatively clean and have the majority of their variability due to noise in the retrievals themselves, which is not explicitly considered by the methods employed herein.


EOF1 Value
(a)
OMNO2 80%P
(b)
OMNO2 80%N
(c)
HSTCM-NO2 80%P
(d)
HSTCM-NO2 80%P
20°N
0°
20°S
20°W 0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E 180° 160°W 140°W 120°W 100°W 80°W 60°W 40°W
0.6
0.4
0.2
0
−0.3
−0.6
−0.9

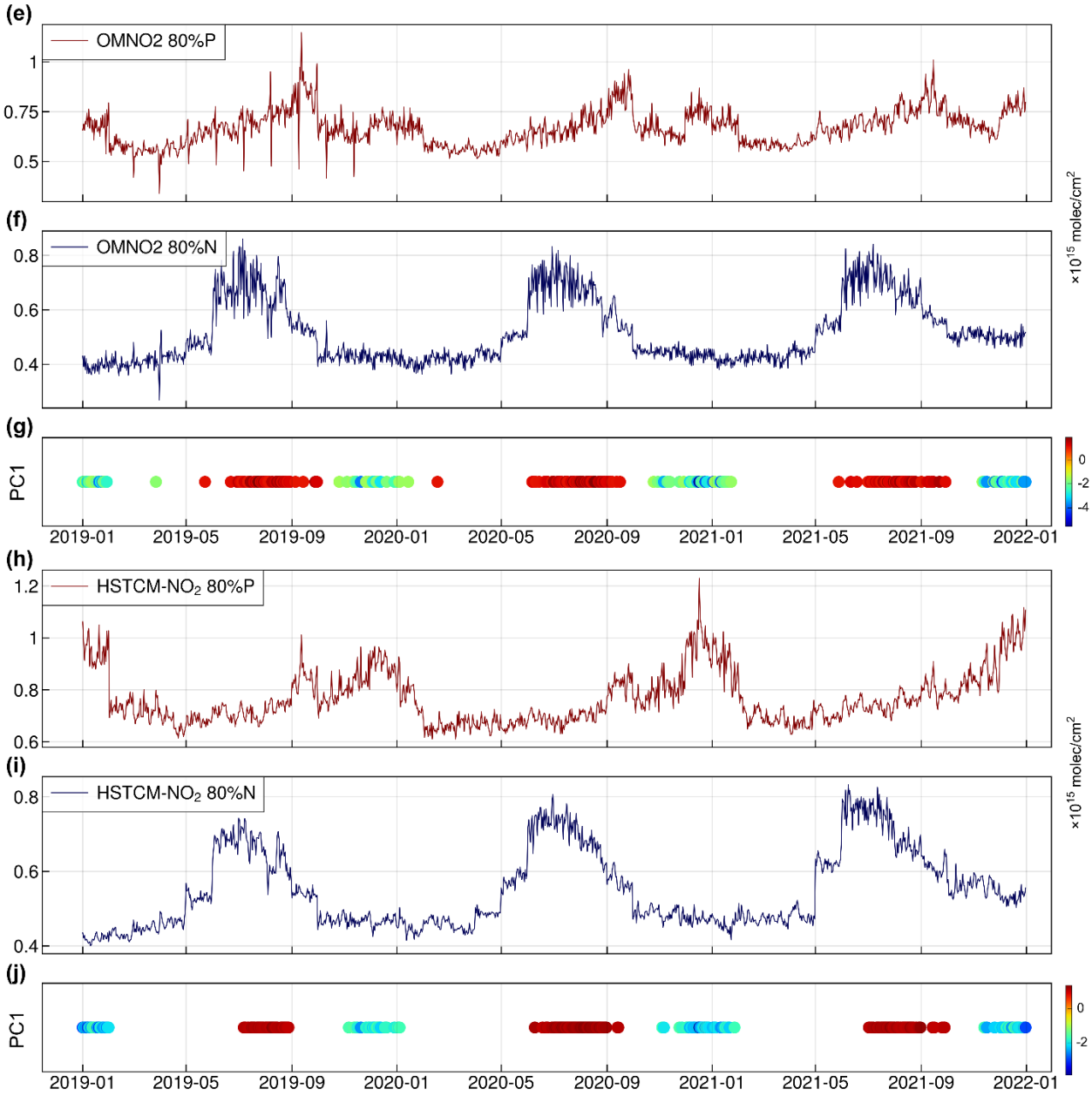


Figure 12: Spatial and temporal patterns after EOF variance maximization is performed on both OMNO2 and HSTCM-NO_2. EOF1 is given for OMNO2 positive **(a)**, OMNO2 negative **(b)**, HSTCM-NO_2 positive **(c)**, and HSTCM-NO_2 negative **(d)**. The temporal mean value of OMNO2 over the EOF1 positive region and EOF1 negative region are respectively given in **(e)** and **(f)**, while PC1 is given in **(g)**, where red and blue represent the peaks in the positive and negative factors respectively. Subfigures **(h)**, **(i)**, and **(j)** are similar to **(e)**, **(f)**, and **(g)** except when applied to HSTCM-NO_2.

Figure 12 shows the spatial and temporal patterns after EOF variance maximization is performed on both OMNO2 and HSTCM-NO_2. First and foremost, the EOFs represent a few general patterns, seeming to capture a combination of biomass burning (across Africa, South America, and Australia), urbanization (across South Africa, Northeastern China and Japan), energy producing regions in the Southern US and northern Mexico, and transport regions from the Mediterranean to the Indian Ocean. This includes the large areas of pollution transported downwind over the various oceans, and uncertainty associated with clouds, sea salt, and low signal strengths near where the Southern Ocean intrudes into the cleaner areas of the Indian and Pacific Oceans respectively. The overall patterns look reasonable in both space and time.

A more detailed analysis clearly demonstrates that three such examples are consistently represented between the original OMNO2 and HSTCM-NO_2. First, the negative mode of EOF1 representing biomass burning over Congo and its subsequent transport over the Southern Atlantic Ocean, and the positive mode of EOF1 representing biomass burning and urbanization over respective parts of Southern Africa are interpolated well and line-filled by the respective negative and positive modes of

HSTCM-NO_2 EOF1 (Du et al., 2020). Second, the wildfires off of Southwestern Australia and subsequent transport into the Southern Ocean are clearly shown by the negative mode of EOF1, while the negative mode of EOF1 of HSTCM-NO_2 expands these observations into the Indian Ocean and all the way to New Zealand, while narrowing the band and reducing the error due to the mixing from the Southern Ocean, consistent with observations (Wenig et al., 2003). Third, the positive region of EOF1 loosely picks up the transported wave-trains from East Asia to North America, while the HSTCM-NO_2 is able to clearly pick up the entire wave-train clearly originating in industrial regions of Japan and spreading part of the time to Luzon and another part of the time to the USA (Wang, et al., 2023). In terms of time, it is clear that the negative EOF1 regions in both plots are well represented by the positive PC1 values. All three peaks demonstrated are clearly observed in the average values of NO_2 over the negative EOF1 regions respectively. There are four large peaks and two small peaks represented in the negative PC1 values, all of which are picked up well in the average values of NO_2 over the positive EOF1 regions respectively. All of the peak times are represented in the time series using different colors.

While the distribution of the HSTCM-NO_2 EOF is more smeared spatially than the OMNO2 product in some regions, this is not unexpected. In some cases, this makes the story consistent, by filling in missing data, especially so in cases of long-range transported plums which are otherwise missing, as well as for the known variation observed over Henan and Shandong. However, some of the smearing is also noise, as identified over the low NO_2 concentration regions near where the Indian and Pacific Oceans intersect with the Southern Ocean.

This analysis shows that the HSTCM-NO_2 product does a decent job at representing the temporal and spatial extremes in the original OMNO2 dataset. While this test is not frequently done in the community (Cohen, 2014; Liu et al., 2023; Liu et al., 2024), it clearly demonstrates in an objective manner a new and additional way to test the goodness of the final product, in that it requires the product to not only match in space and time with observed mean conditions, but also with observed extreme conditions. The fact that there is spatial smearing in some aspects is good, in that it fills in missing long-range transport events that are missed between swaths or due to clouds in situ. In other aspects, it may extend the actual signals too far in space. For these reasons, care must be used when applying the results. We hope that this section sets a gold standard by which future big data products are more carefully compared with and validated against the underlying data.

4 Data availability

The global daily High Spatial-Temporal Coverage Merged tropospheric NO_2 dataset (HSTCM-NO_2) from 2007 to 2022 based on OMI and GOME-2 can be accessed directly through: https://doi.org/10.5281/zenodo.10968462 (Qin et al., 2024).

5 Conclusions and discussion

In order to improve the spatial coverage of OMNO2 due to data loss caused by cloud occlusion, row anomaly, high retrieval noise, and other issues, this study proposes an effective method of reconstruction consisting of machine learning (XGBoost) and gap filling (DINEOF) to produce a new reconstructed product (HSTCM-NO_2).

First, the process of applying XGBoost first followed by DINEOF second yields the highest correlation and lowest RMSE between the OMNO2 and HSTCM-NO_2. One reason for this is that XGBoost requires the presence of GOME-2 data, allowing for additional observational support in the final reconstructed product. This is consistent with the fact that GOME-2 occupies a very high SHAP value. There are a few qualifiers however: first that cases without prior knowledge perform less well than places with priori knowledge; and second that locations with a lower column loading of OMNO2 work better than places with a higher column loading of OMNO2. Since the majority of the data points globally are biased towards lower (i.e. non-polluted) areas, comparison with additional datasets and using different approaches is essential.

Second, external observations from MAX-DOAS and TROPOMI as well as reanalysis data from EAC4 are used to validate HSTCM-NO_2 on a column-by-column, large-area basis. HSTCM-NO_2 shows good correlations with all of the observations

above, especially so when the VCDs are below 6×10^{15} molec.cm^{-2}. Specific issues in terms of spatial distribution mismatches and issues reproducing very high VCDs are explained in detail within the paper. There are a few exceptions to this, specifically over Wuhan and the Yangtze River from Wuhan up to Nanjing, and specific urban parts of India (such as New Delhi) being reasonably well represented.

Third, additional analysis to verify the goodness of HSTCM-NO_2 in terms of being able to capture extreme events observed within the OMNO2 data is also performed. In this case, variance maximization is used to decompose the OMNO2 data into standing spatial (EOF) and temporal (PC) signals. A similar analysis is performed on the HSTCM-NO_2 data, with the resulting signals compared. It is shown that in addition to generally matching in terms of space and time, data after gap filling observed by HSTCM-NO_2 especially downwind from large pollution areas over various oceans (South Atlantic, Indian, South Pacific and North Pacific) are improved. Interestingly, some of the strongest signals, including biomass burning from central and northern Africa including in Algeria, including pixels over the value of 6×10^{15} molec.cm^{-2} are also well represented, in terms of both the magnitude, as well as the spatial and temporal extremes.

This combination of findings indicates that the new HSTCM-NO_2 product works well in terms of representing both the grid-by-grid and climatological mean conditions, as well as extreme events, with the caveats that first there is some apriori knowledge and second that the original OMNO2 data has an VCD below 6×10^{15} molec.cm^{-2} (i.e. is not heavily polluted).In the future, related work will focus on how to enhance the application of datasets in polluted scenes. Separating low and high values for training might be an effective approach, since it is known that there are different retrieval assumptions and impacts that occur under polluted and non-polluted conditions (Boersma et al., 2007; Chimot et al., 2016; Lorente et al., 2018; Liu et al., 2019; Zhou et al., 2024). Presently the criteria for demarcation and the sets of impacting variables are still undergoing discussion by the community and are not yet agreed upon. Whether there are better methods or combinations of methods that can be applied across the full range of scenarios at the same time is also something that needs to be considered.

6 Author contributions

KQ is responsible for conceptualization, funding acquisition, investigation, project administration, resources, supervision, and editing. JBC is responsible for refining the methodology, supervision, validation, reviewing the results, and editing the original draft. HRG is responsible for data curation, formal analysis, software, visualization, and writing the original draft. XL is responsible for data curation and visualization. QH is responsible for editing the original draft and visualization. PT is responsible for editing the draft and additional formal analysis during the response phase. All authors contributed to the final paper.

7 Competing interests

The contact author has declared that none of the authors has any competing interests.

8 Financial support

This research has been supported by the National Natural Science Foundation of China (Grant No. 42375125).

References

Abdulmanov, R., Miftakhov, I., Ishbulatov, M., Galeev, E., and Shafeeva, E.: Comparison of the effectiveness of GIS-based interpolation methods for estimating the spatial distribution of agrochemical soil properties, Environ. Technol. Innov., 24, 101970, https://doi.org/10.1016/j.eti.2021.101970, 2021.

Achite, M., Katipoğlu, Okan Mert, Javari, M., and Caloiero, T.: Hybrid interpolation approach for estimating the spatial variation of annual precipitation in the Macta basin, Algeria, Theor. Appl. Climatol., 155, 1139–1166, https://doi.org/10.1007/s0070402304685w, 2024.

Alvera-Azcárate, A., Barth, A., Sirjacobs, D., and Beckers, J.-M.: Enhancing temporal correlations in EOF expansions for the reconstruction of missing data using DINEOF, Ocean Sci., 5, 475–485, https://doi.org/10.5194/os-5-475-2009, 2009.

Alvera-Azcárate, A., Barth, A., Parard, G., and Beckers, J.-M.: Analysis of SMOS sea surface salinity data using DINEOF, Remote Sens. Environ., 180, 137–145, https://doi.org/10.1016/j.rse.2016.02.044, 2016.

Baek, K.. and Kim, J.: Analysis of Characteristics of Satellite-derived Air Pollutant over Southeast Asia and Evaluation of Tropospheric Ozone using Statistical Methods, J. Korean Soc. Atmos. Environ., 27, 650–662, https://doi.org/10.5572/kosae.2011.27.6.650, 2011.

Bauer, R., A. Rozanov, McLinden, C. A., Gordley, L. L., Lotz, W., Iii, R., Walker, K. A., Zawodny, J. M., A. Ladstätter-Weißenmayer, H. Bovensmann, and Burrows, J. P.: Validation of SCIAMACHY limb NO_2 profiles using solar occultation measurements, Atmos. Meas. Tech., 5, 1059–1084, https://doi.org/10.5194/amt-5-1059-2012, 2012.

Beckers, J.-M. and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets, J. Atmos. Ocean. Technol., 20, 1839–1856, https://doi.org/10.1175/15200426(2003)020%3C1839:ECADFF%3E2.0.CO;2, 2003.

Boersma, K. F., Eskes, H. J., and Brinksma, E. J.: Error analysis for tropospheric NO_2 retrieval from space, J. Geophys. Res., 109, https://doi.org/10.1029/2003JD003962, 2004.

Boersma, K. F., Eskes, H. J., Veefkind, J. P., Brinksma, E. J., van der A, R. J., Sneep, M., van den Oord, G. H. J., Levelt, P. F., Stammes, P., Gleason, J. F., and Bucsela, E. J.: Near-real time retrieval of tropospheric NO_2 from OMI, Atmos. Chem. Phys., 7, 2103–2118, https://doi.org/10.5194/acp721032007, 2007.

Boersma, K. F., Jacob, D. J., Eskes, H. J., Pinder, R. W., Wang, J., and van der A, R. J.: Intercomparison of SCIAMACHY and OMI tropospheric NO_2 columns: Observing the diurnal evolution of chemistry and emissions from space, J. Geophys. Res., 113, https://doi.org/10.1029/2007JD008816, 2008.

Bousserez, N.: Space-based retrieval of NO_2 over biomass burning regions: quantifying and reducing uncertainties, Atmos. Meas. Tech., 7, 3431–3444, https://doi.org/10.5194/amt-7-3431-2014, 2014.

Bovensmann, H., Burrows, J. P., Buchwitz, M., Frerick, J., Noël, S., Rozanov, V. V., Chance, K. V., and Goede, A. P. H.: SCIAMACHY: Mission Objectives and Measurement Modes, J. Atmos. Sci., 56, 127–150, https://doi.org/10.1175/1520-0469(1999)056%3C0127:SMOAMM%3E2.0.CO;2, 1999.

Chang, N., Bai, K., and Chen, C.: Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management, J. Environ. Manage., 201, 227–240, https://doi.org/10.1016/j.jenvman.2017.06.045, 2017.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 2016, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chen, Z., Wang, P., Bao, S., and Zhang, W.: Rapid reconstruction of temperature and salinity fields based on machine learning and the assimilation application, Front. Mar. Sci., 9, https://doi.org/10.3389/fmars.2022.985048, 2022.

Chimot, J., Vlemmix, T., Veefkind, J. P., de Haan, J. F., and Levelt, P. F.: Impact of aerosols on the OMI tropospheric NO_2 retrievals over industrialized regions: how accurate is the aerosol correction of cloud-free scenes via a simple cloud model?, Atmos. Meas. Tech., 9, 359–382, https://doi.org/10.5194/amt-9-359-2016, 2016.

Chowdhury, M. R.: Overview of Weather, ENSO, and Climate Scale, in: Seasonal Flood Forecasts and Warning Response Opportunities: ENSO Applications in Bangladesh, Springer, Cham, 59–74, https://doi.org/10.1007/9783031178252_4, 2022.

Cisty, M. and Soldanova, V.: Flow Prediction Versus Flow Simulation Using Machine Learning Algorithms, in: Machine Learning and Data Mining in Pattern Recognition, vol. 10935, edited by: Perner, P., Springer, Cham, 369–382, 2018.

Cohen, J. B.: Quantifying the occurrence and magnitude of the Southeast Asian fire climatology, Environ. Res. Lett., 9, 114018, https://doi.org/10.1088/1748-9326/9/11/114018, 2014.

Cohen, J. B., Prinn, R. G., and Wang, C.: The impact of detailed urban-scale processing on the composition, distribution, and radiative forcing of anthropogenic aerosols, Geophys. Res. Lett., 38, L10808, https://doi.org/10.1029/2011gl047417, 2011.

Cohen, J. B., Lecoeur, E., and Hui Loong Ng, D.: Decadal-scale relationship between measurements of aerosols, land-use change, and fire over Southeast Asia, Atmos. Chem. Phys., 17, 721–743, https://doi.org/10.5194/acp-17-721-2017, 2017.

Compernolle, S., Verhoelst, T., Pinardi, G., Granville, J., Hubert, D., Keppens, A., Niemeijer, S., Rino, B., Bais, A., Beirle, S., Boersma, F., P, B. J., De Smedt, I, Eskes, H., Goutail, F., Hendrick, F., Lorente, A., Pazmino, A., Piters, A., and Peters, E.: 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, 8017–8045, https://doi.org/10.5194/acp2080172020, 2020.

Cooper, M. J., Martin, R. V., Hammer, M. S., and McLinden, C. A.: An Observation-Based Correction for Aerosol Effects on Nitrogen Dioxide Column Retrievals Using the Absorbing Aerosol Index, Geophys. Res. Lett., 46, 8442–8452, https://doi.org/10.1029/2019GL083673, 2019.

Cooper, M. J., Martin, R. V., Hammer, M. S., Levelt, P. F., Veefkind, P., Lamsal, L. N., Krotkov, N. A., Brook, J. R., and McLinden, C. A.: Global fine-scale changes in ambient NO_2 during COVID-19 lockdowns, Nature, 601, 380–387, https://doi.org/10.1038/s41586-021-04229-0, 2022.

Crutzen, P. J.: Gas-Phase Nitrogen and Methane Chemistry in the Atmosphere, in: Physics and Chemistry of Upper Atmosphere, 110–124, 1973.

Crutzen, P. J.: The Role of NO and NO_2 in the Chemistry of the Troposphere and Stratosphere, Annu. Rev. Earth Planet. Sci., 7, 443–472, https://doi.org/10.1146/annurev.ea.07.050179.002303, 1979.

Cui, Y., Wang, L., Jiang, L., Liu, M., Wang, J., Shi, K., and Duan, X.: Dynamic spatial analysis of NO_2 pollution over China: Satellite observations and spatial convergence models, Atmos. Pollut. Res., 12, 89–99, https://doi.org/10.1016/j.apr.2021.02.003, 2021.

de Hoogh, K., Saucy, A., Shtein, A., Schwartz, J., West, E. A., Strassmann, A., Puhan, M., Röösli, M., Stafoggia, M., and Kloog, I.: Predicting Fine-Scale Daily NO_2 for 2005–2016 Incorporating OMI Satellite Data Across Switzerland, Environ. Sci. Technol., 53, 10279–10287, https://doi.org/10.1021/acs.est.9b03107, 2019.

Deng, W., Cohen, J. B., Wang, S., and Lin, C.: Improving the understanding between climate variability and observed extremes of global NO_2 over the past 15 years, Environ. Res. Lett., 16, 054020, https://doi.org/10.1088/1748-9326/abd502, 2021.

Dong, J., Chen, Y., Yao, B., Zhang, X., and Zeng, N.: A neural network boosting regression model based on XGBoost, Appl. Soft. Comput., 125, 109067, https://doi.org/10.1016/j.asoc.2022.109067, 2022.

Du, M., Chen, L., Lin, J., Liu, Y., Feng, K., Liu, Q., Liu, Y., Wang, J., Ni, R., Zhao, Y., Si, W., Li, Y., Kong, H., Weng, H., Liu, M., and Adeniran, J. A.: Winners and losers of the Sino–US trade war from economic and environmental perspectives, Environ. Res. Lett., 15, 094032, https://doi.org/10.1088/1748-9326/aba3d5, 2020.

Duncan, B. N., Yoshida, Y., de Foy, B., Lamsal, L. N., Streets, D. G., Lu, Z., Pickering, K. E., and Krotkov, N. A.: The observed response of Ozone Monitoring Instrument (OMI) NO_2 columns to NO_x emission controls on power plants in the United States: 2005–2011, Atmos. Environ., 81, 102–111, https://doi.org/10.1016/j.atmosenv.2013.08.068, 2013.

Echterhof, T. and Pfeifer, H.: Nitrogen Oxide Formation in the Electric Arc Furnace—Measurement and Modeling, Metall. Mater. Trans. B-Proc. Metall. Mater. Proc. Sci., 43, 163–172, https://doi.org/10.1007/s11663-011-9564-8, 2011.

Eriksson, E.: Composition of Atmospheric Precipitation, Tellus, 4, 280–303, https://doi.org/10.3402/tellusa.v4i4.8813, 1952.

Eskes, H. J. and Boersma, K. F.: Averaging kernels for DOAS total-column satellite retrievals, Atmos. Chem. Phys., 3, 1285–1291, https://doi.org/10.5194/acp-3-1285-2003, 2003.

European Commission, Joint Research Centre, Crippa, M., Guizzardi, D., Schaaf, E. GHG emissions of all world countries:

Publications Office of the European Union. https://data.europa.eu/doi/10.2760/953322, 2023.

Fan, Z., Zhan, Q., Yang, C., Liu, H., and Bilal, M.: Estimating $PM_{2.5}$ Concentrations Using Spatially Local Xgboost Based on Full-Covered SARA AOD at the Urban Scale, Remote Sens., 12, https://doi.org/10.3390/rs12203368, 2020.

Fioletov, V. E., McLinden, C. A., Krotkov, N., Yang, K., Loyola, D. G., Valks, P., Theys, N., Van Roozendael, M., Nowlan, C. R., Chance, K., Liu, X., Lee, C., and Martin, R. V.: Application of OMI, SCIAMACHY, and GOME-2 satellite SO_2 retrievals for detection of large emission sources, J. Geophys. Res. Atmos., 118, 11, 399–11, 418, https://doi.org/10.1002/jgrd.50826, 2013.

Fishman, J., Solomon, S., and Crutzen, P. J.: Observational and theoretical evidence in support of a significant in-situ photochemical source of tropospheric ozone, Tellus, 31, 432–446, https://doi.org/10.1111/j.21533490.1979.tb00922.x, 1979.

Foret, G., Eremenko, M., Cuesta, J., Sellitto, P., Barré, J., Gaubert, B., Coman, A., Dufour, G., Liu, X., Joly, M., Doche, C., and Beekmann, M.: Ozone pollution: What can we see from space? A case study, J. Geophys. Res. Atmos., 119, 8476–8499, https://doi.org/10.1002/2013JD021340, 2014.

He, Q., Qin, K., Cohen, J. B., Loyola, D., Li, D., Shi, J., and Xue, Y.: Spatially and temporally coherent reconstruction of tropospheric NO_2 over China combining OMI and GOME-2B measurements, Environ. Res. Lett., 15, 125011, https://doi.org/10.1088/17489326/abc7df, 2020.

Hilborn, A. and Costa, M.: Applications of DINEOF to Satellite-Derived Chlorophyll-a from a Productive Coastal Region, Remote Sens., 10, 1449, https://doi.org/10.3390/rs10091449, 2018.

Ialongo, I., Virta, H., Eskes, H., Hovila, J., and Douros, J.: Comparison of TROPOMI/Sentinel-5 Precursor NO_2 observations with ground-based measurements in Helsinki, Atmos. Meas. Tech., 13, 205–218, https://doi.org/10.5194/amt-13-205-2020, 2020.

Irie, H., Boersma, K. F., Kanaya, Y., Takashima, H., Pan, X., and Wang, Z. F.: Quantitative bias estimates for tropospheric NO_2 columns retrieved from SCIAMACHY, OMI, and GOME-2 using a common standard for East Asia, Atmos. Meas. Tech., 5, 2403–2411, https://doi.org/10.5194/amt-5-2403-2012, 2012.

Jiang, Y., Gao, Z., He, J., Wu, J., and Christakos, G.: Application and Analysis of XCO_2 Data from OCO Satellite Using a Synthetic DINEOF–BME Spatiotemporal Interpolation Framework, Remote Sens., 14, 4422, https://doi.org/10.3390/rs14174422, 2022.

Just, A. C., Margherita, C., Shtein, A., Dorman, M., Lyapustin, A., and Kloog, I.: Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling $PM_{2.5}$ in the Northeastern USA, Remote Sens., 10, https://doi.org/10.3390/rs10050803, 2018.

Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, J. Basic Eng, 82, 35–45, https://doi.org/10.1115/1.3662552, 1960.

Kapoor, S. and Perrone, V.: A Simple and Fast Baseline for Tuning Large XGBoost Models, arXiv (Cornell University), https://doi.org/10.48550/arxiv.2111.06924, 2021.

Kim, J., Baek, K., and Kim, S.: Validation of OMI HCHO with EOF and SVD over Tropical Africa, Korean J. Remote Sensing, 30, 417–430, https://doi.org/10.7780/kjrs.2014.30.4.1, 2014.

Kolle, O., Kalthoff, N., Kottmeier, C., and William, M. J.: Ground-Based Platforms, in: Springer Handbook of Atmospheric Measurements, edited by: Foken, T., Springer International Publishing, Cham, 155–182, https://doi.org/10.1007/9783030521714_6, 2021.

Laan, E., D. de Winter, J. de Vries, P. F. Levelt, G. H. van den Oord, A. Malkki, G. W. Leppelmeier, and E. Hilsenrath: Toward the use of the Ozone Monitoring Instrument (OMI), Proc. SPIE, 4540, 270 – 277, 2001.

Lamsal, L. N., Krotkov, N. A., Vasilkov, A., Marchenko, S., Qin, W., Yang, E., Fasnacht, Z., Joiner, J., Choi, S., Haffner, D., Swartz, W. H., Fisher, B., and Bucsela, E.: Ozone Monitoring Instrument (OMI) Aura nitrogen dioxide standard product version 4.0 with improved surface and cloud treatments, Atmos. Meas. Tech., 14, 455–479, https://doi.org/10.5194/amt-14-

455-2021, 2021.

Leitão, J., Richter, A., Vrekoussis, M., Kokhanovsky, A., Zhang, Q. J., Beekmann, M., and Burrows, J. P.: On the improvement of NO_2 satellite retrievals – aerosol impact on the airmass factors, Atmos. Meas. Tech., 3, 475–493, https://doi.org/10.5194/amt-3-475-2010, 2010.

Levy, H.: Photochemistry of the lower troposphere, Planet Space Sci., 20, 919–935, https://doi.org/10.1016/00320633(72)901778, 1972.

Li, D., Qin, K., Cohen, J. B., He, Q., Wang, S., Li, D., Zhou, X., Ling, X., and Xue, Y.: Combing GOME-2B and OMI Satellite Data to Estimate Near-Surface NO_2 of Mainland China, IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 14, 10269–10277, https://doi.org/10.1109/JSTARS.2021.3117396, 2021.

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., and He, K.: Anthropogenic emission inventories in China: a review, Natl. Sci. Rev., 4, 834–866, https://doi.org/10.1093/nsr/nwx150, 2017.

Li, M., Mao, J., Chen, S., Bian, J., Bai, Z., Wang, X., Chen, W., and Yu, P.: Significant contribution of lightning NO_x to summertime surface O_3 on the Tibetan Plateau, Sci. Total Environ., 829, 154639, https://doi.org/10.1016/j.scitotenv.2022.154639, 2022.

Li, X., Cohen, J. B., Qin, K., Geng, H., Wu, X., Wu, L., Yang, C., Zhang, R., and Zhang, L.: Remotely sensed and surface measurement-derived mass-conserving inversion of daily NO_x emissions and inferred combustion technologies in energy-rich northern China, Atmos. Chem. Phys., 23, 8001–8019, https://doi.org/10.5194/acp-23-8001-2023, 2023.

Lin, J., Martin, R. V., Boersma, K. F., Sneep, M., Stammes, P., Spurr, R., Wang, P., Van Roozendael, M., Clémer, K., and Irie, H.: 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, 1441–1461, https://doi.org/10.5194/acp-14-1441-2014, 2014.

Liu, F., Zhang, Q., van der A, R. J., Zheng, B., Tong, D., Yan, L., Zheng, Y., and He, K.: Recent reduction in NO_x missions over China: synthesis of satellite observations and emission inventories, Environ. Res. Lett., 11, 114002, https://doi.org/10.1088/1748-9326/11/11/114002, 2016.

Liu, J., Cohen, J. B., He, Q., Tiwari, P., and Qin, K.: Accounting for NO_x emissions from biomass burning and urbanization doubles existing inventories over South, Southeast and East Asia, Commun. Earth Environ., 5, 255, https://doi.org/10.1038/s43247024014245, 2024.

Liu, J., Cohen, J. B., Yim, S., and Gupta, P.: New Top-Down Estimation of Daily Mass and Number Column Density of Black Carbon Driven by Omi and Aeronet Observations, https://doi.org/10.2139/ssrn.4762410, 2024.

Liu, M., Lin, J., Boersma, K. F., Pinardi, G., Wang, Y., Chimot, J., Wagner, T., Xie, P., Eskes, H., Van Roozendael, M., Hendrick, F., Wang, P., Wang, T., Yan, Y., Chen, L., and Ni, R.: Improved aerosol correction for OMI tropospheric NO_2 retrieval over East Asia: constraint from CALIOP aerosol vertical profile, Atmos. Meas. Tech., 12, 1–21, https://doi.org/10.5194/amt-12-1-2019, 2019.

Liu, S., Valks, P., Pinardi, G., De Smedt, I., Yu, H., Beirle, S., and Richter, A.: An improved total and tropospheric NO_2 column retrieval for GOME-2, Atmos. Meas. Tech., 12, 1029–1057, https://doi.org/10.5194/amt-12-1029-2019, 2019.

Liu, Z., Cohen, J. B., Wang, S., Wang, X., Tiwari, P., and Qin, K.: Remotely sensed BC columns over rapidly changing Western China show significant decreases in mass and inconsistent changes in number, size, and mixing properties due to policy actions, npj Clim. Atmos. Sci., 7, 124, https://doi.org/10.1038/s41612024006639, 2024.

Logan, J. A.: Nitrogen oxides in the troposphere: Global and regional budgets, J. Geophys. Res., 88, 10785, https://doi.org/10.1029/jc088ic15p10785, 1983.

Logan, J. A., Prather, M. J., Wofsy, S. C., and McElroy, M. B.: Tropospheric chemistry: A global perspective, J. Geophys. Res., 86, 7210, https://doi.org/10.1029/jc086ic08p07210, 1981.

Lorente, A., Boersma, K. F., Stammes, P., Tilstra, L. G., Richter, A., Yu, H., Kharbouche, S., and Muller, J.-P.: The importance of surface reflectance anisotropy for cloud and NO_2 retrievals from GOME-2 and OMI, Atmos. Meas. Tech., 11, 4509–4529, https://doi.org/10.5194/amt-11-4509-2018, 2018.

Lu, X., Ye, X., Zhou, M., Zhao, Y., Weng, H., Kong, H., Li, K., Gao, M., Zheng, B., Lin, J., Zhou, F., Zhang, Q., Wu, D., Zhang, L., and Zhang, Y.: The underappreciated role of agricultural soil nitrogen oxide emissions in ozone pollution regulation in North China, Nat. Commun., 12, 5021, https://doi.org/10.1038/s41467021251479, 2021.

Lu, Y. and Khalil, M. A. K.: Methane and carbon monoxide in OH chemistry: The effects of feedbacks and reservoirs generated by the reactive products, Chemosphere, 26, 641–655, https://doi.org/10.1016/0045-6535(93)90450-J, 1993.

Lu, Z., Liu, X., Zaveri, R. A., Easter, R. C., Tilmes, S., Emmons, L. K., Vitt, F., Singh, B., Wang, H., Zhang, R., and Rasch, P. J.: Radiative Forcing of Nitrate Aerosols From 1975 to 2010 as Simulated by MOSAIC Module in CESM2-MAM4, J. Geophys. Res.-Atmos., 126, https://doi.org/10.1029/2021jd034809, 2021.

Ludewig, A., Kleipool, Q., Bartstra, R., Landzaat, R., Leloux, J., Loots, E., Meijering, P., van der Plas, E., Rozemeijer, N., Vonk, F., and Veefkind, P.: In-flight calibration results of the TROPOMI payload on board the Sentinel-5 Precursor satellite, Atmos. Meas. Tech., 13, 3561–3580, https://doi.org/10.5194/amt-13-3561-2020, 2020.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E.: Environmental and Health Impacts of Air Pollution: A Review, Front. Public Health, 8, 1–13, https://doi.org/10.3389/fpubh.2020.00014, 2020.

Munro, R., Lang, R., Klaes, D., Poli, G., Retscher, C., Lindstrot, R., Huckle, R., Lacan, A., Grzegorski, M., Holdak, A., Kokhanovsky, A., Livschitz, J., and Eisinger, M.: The GOME-2 instrument on the Metop series of satellites: instrument design, calibration, and level 1 data processing – an overview, Atmos. Meas. Tech., 9, 1279–1301, https://doi.org/10.5194/amt-9-1279-2016, 2016.

Park, J., Kim, H., Bae, D., and Jo, Y.: Data Reconstruction for Remotely Sensed Chlorophyll-a Concentration in the Ross Sea Using Ensemble-Based Machine Learning, Remote Sens., 12, 1898, https://doi.org/10.3390/rs12111898, 2020.

Pinardi, G., Roozendael, V., Hendrick, F., Theys, N., Abuhassan, N., Bais, A., Boersma, F., Cede, A., Chong, J., Donner, S., Drosoglou, T., Dzhola, A., Eskes, H., Frieß, U., Granville, J., Herman, J., Holla, R., Hovila, J., Irie, H., and Wittrock, F.: Validation of tropospheric NO_2 column measurements of GOME-2A and OMI using MAX-DOAS and direct sun network observations, Atmos. Meas. Tech., 13, 6141–6174, https://doi.org/10.5194/amt1361412020, 2020.

Platt, U. and Stutz, J.: Differential Optical Absorption Spectroscopy: Principles and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Qin, K., Lu, L., Liu, J., He, Q., Shi, J., Deng, W., Wang, S., and Cohen, J. B.: Model-free daily inversion of NO_x emissions using TROPOMI (MCMFE-NO_x) and its uncertainty: Declining regulated emissions and growth of new sources, Remote Sens. Environ., 295, 113720, https://doi.org/10.1016/j.rse.2023.113720, 2023.

Qin, K., Gao, H., Liu, X., He, Q., and Cohen, J. B.: HSTCM-NO_2 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10968462, 2024.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., and Balmes, J. R.: Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning, Environ. Sci. Technol., 49, 3887–3896, https://doi.org/10.1021/es505846r, 2015.

Richter, A. and Burrows, J. P.: Tropospheric NO_2 from GOME measurements, Adv. Space Res., 29, 1673–1683, https://doi.org/10.1016/S02731177(02)00100X, 2002.

Richter, A., Burrows, J. P., Nüß, H., Granier, C., and Niemeier, U.: Increase in tropospheric nitrogen dioxide over China observed from space, Nature, 437, 129–132, https://doi.org/10.1038/nature04092, 2005.

Richter, A., Begoin, M., Hilboll, A., and Burrows, J. P.: An improved NO_2 retrieval for the GOME-2 satellite instrument, Atmos. Meas. Tech., 4, 1147–1159, https://doi.org/10.5194/amt411472011, 2011.

Running, S. W.: Generalization of a forest ecosystem process model for other biomes, BIOME-BGC, and an application for

global-scale models, Scaling Physiological Processes Leaf to Globe, 1993.

Sanabria, L. A., Qin, X., Li, J., Cechet, R. P., and Lucas, C.: Spatial interpolation of McArthur's Forest Fire Danger Index across Australia: Observational study, Environ. Modell. Softw., 50, 37–50, https://doi.org/10.1016/j.envsoft.2013.08.012, 2013.

Schaub, D., Boersma, K. F., Kaiser, J. W., Weiss, A. K., Folini, D., Eskes, H. J., and Buchmann, B.: Comparison of GOME tropospheric NO_2 columns with NO_2 profiles deduced from ground-based in situ measurements, Atmos. Chem. Phys., 6, 3211–3229, https://doi.org/10.5194/acp-6-3211-2006, 2006.

Shao, Y., Zhao, W., Liu, R., Yang, J., Liu, M., Fang, W., Hu, L., Adams, M., Bi, J., and Ma, Z.: Estimation of daily NO_2 with explainable machine learning model in China, 2007–2020, Atmos. Environ., 314, 120111, https://doi.org/10.1016/j.atmosenv.2023.120111, 2023.

Shapley, L. S.: A Value for N-Person Games, RAND Corporation, Santa Monica, CA, https://doi.org/10.7249/P0295, 1952.

Sillman, S., Logan, J. A., and Wofsy, S. C.: The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes, J. Geophys. Res., 95, 1837, https://doi.org/10.1029/jd095id02p01837, 1990.

Sorenson, H. W.: Least-squares estimation: from Gauss to Kalman, IEEE Spectr., 7, 63–68, https://doi.org/10.1109/MSPEC.1970.5213471, 1970.

Spivakovsky, C. M., Logan, J. A., Montzka, S. A., Balkanski, Y. J., Foreman-Fowler, M., Jones, D. B. A., Horowitz, L. W., Fusco, A. C., Brenninkmeijer, C. A. M., Prather, M. J., Wofsy, S. C., and McElroy, M. B.: Three-dimensional climatological distribution of tropospheric OH: Update and evaluation, J. Geophys. Res., 105, 8931–8980, https://doi.org/10.1029/1999jd901006, 2000.

Streets, D. G., Canty, T., Carmichael, G. R., de Foy, B., Dickerson, R. R., Duncan, B. N., Edwards, D. P., Haynes, J. A., Henze, D. K., Houyoux, M. R., Jacob, D. J., Krotkov, N. A., Lamsal, L. N., Liu, Y., Lu, Z., Martin, R. V., Pfister, G. G., Pinder, R. W., Salawitch, R. J., and Wecht, K. J.: Emissions estimation from satellite retrievals: A review of current capability, Atmos. Environ., 77, 1011–1042, https://doi.org/10.1016/j.atmosenv.2013.05.051, 2013.

Sun, W., Shao, M., Granier, C., Liu, Y., Ye, C., and Zheng, J.: Long-Term Trends of Anthropogenic SO_2, NO_x, CO, and NMVOCs Emissions in China, Earth's Future, 6, 1112–1133, https://doi.org/10.1029/2018ef000822, 2018.

Timlin, A., Hastings, A., and Hardiman, M.: Workbased facilitators as drivers for the development of person-centred cultures: a shared reflection from novice facilitators of person-centred practice, International Practice Development Journal, 8, 1–9, https://doi.org/10.19043/ipdj81.008, 2018.

Tiwari, P., Cohen, J. B., Wang, X., Wang, S., and Qin, K.: Radiative forcing bias calculation based on COSMO (CoreShell Mie model Optimization) and AERONET data, npj Clim. Atmos. Sci., 6, 193, https://doi.org/10.1038/s41612023005201, 2023.

Torres, O., Bhartia, P. K., Jethva, H., and Ahn, C.: Impact of the ozone monitoring instrument row anomaly on the long-term record of aerosol products, Atmos. Meas. Tech., 11, 2701–2715, https://doi.org/10.5194/amt-11-2701-2018, 2018.

van der A, R. J., Peters, D. H. M. U., Eskes, H., Boersma, K. F., Van Roozendael, M., De Smedt, I., and Kelder, H. M.: Detection of the trend and seasonal variation in tropospheric NO_2 over China, J. Geophys. Res., 111, 1–10, https://doi.org/10.1029/2005JD006594, 2006.

van Geffen, J., Eskes, H. J., Boersma, K. F., Maasakkers, J. D., and Veefkind, J. P.: TROPOMI ATBD of the total and tropospheric NO_2 data products (Royal Netherlands Meteorological Institute), 2019.

van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., ter Linden, M., Zara, M., and Veefkind, J. P.: S5P TROPOMI NO_2 slant column retrieval: method, stability, uncertainties and comparisons with OMI, Atmos. Meas. Tech., 13, 1315–1335, https://doi.org/10.5194/amt-13-1315-2020, 2020.

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J., de Haan, J. F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., and

Vink, R.: 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, https://doi.org/10.1016/j.rse.2011.09.027, 2012.

Verhoelst, T., Compernolle, S., Pinardi, G., Lambert, J.-C., Eskes, H., Eichmann, K.-U., Fjaeraa, A. M., Granville, J., Niemeijer, S., Cede, A., Tiefengraber, M., Hendrick, F., Pazmiño, A., Bais, A., Bazureau, A., Boersma, K. F., Bognar, K., Dehn, A., Donner, S., and Zehner, C.: 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., 14, 481–510, https://doi.org/10.5194/amt144812021, 2021.

Wagner, T., Dix, B., Friedeburg, C. v., Frieß, U., Sanghavi, S., Sinreich, R., and Platt, U.: MAX-DOAS O_4 measurements: A new technique to derive information on atmospheric aerosols—Principles and information content, J. Geophys. Res.-Atmos., 109, D22205, https://doi.org/10.1029/2004jd004904, 2004.

Wang, C., Wang, T., Wang, P., and Rakitin, V.: Comparison and Validation of TROPOMI and OMI NO_2 Observations over China, Atmosphere, 11, https://doi.org/10.3390/atmos11060636, 2020.

Wang, S., Cohen, J. B., Lin, C., and Deng, W.: Constraining the relationships between aerosol height, aerosol optical depth and total column trace gas measurements using remote sensing and models, Atmos. Chem. Phys., 20, 15401–15426, https://doi.org/10.5194/acp-20-15401-2020, 2020.

Wang, S., Cohen, J. B., Deng, W., Qin, K., and Guo, J.: Using a New Top-Down Constrained Emissions Inventory to Attribute the Previously Unknown Source of Extreme Aerosol Loadings Observed Annually in the Monsoon Asia Free Troposphere, Earth's Future, 9, https://doi.org/10.1029/2021ef002167, 2021.

Wang, S., Ma, X., Zhou, S., Wu, L., Wang, H., Tang, Z., Xu, G., Jing, Z., Chen, Z., and Gan, B.: Extreme atmospheric rivers in a warming climate, Nat. Commun., 14, 3219, https://doi.org/10.1038/s4146702338980x, 2023.

Wang, S., Cohen, J. B., Guan, L., Tiwari, P., and Qin, K.: Classifying and quantifying decadal changes in wet deposition over Southeast and East Asia using EANET, OMI, and GPCP, Atmos. Res., 304, 107400, https://doi.org/10.1016/j.atmosres.2024.107400, 2024.

Wang, S., Zhang, Q., Streets, D. G., He, K., Martin, R. V., Lamsal, L. N., Chen, D., Lei, Y., and Lu, Z.: Growth in NO_x emissions from power plants in China: bottom-up estimates and satellite observations, Atmos. Chem. Phys., 12, 4429–4447, https://doi.org/10.5194/acp-12-4429-2012, 2012.

Wang, Y. and Liu, D.: Reconstruction of satellite chlorophyll-a data using a modified DINEOF method: a case study in the Bohai and Yellow seas, China, Int. J. Remote Sens., 35, 204–217, https://doi.org/10.1080/01431161.2013.866290, 2013.

Wang, Y., Beirle, S., Lampel, J., Koukouli, M., De Smedt, I., Theys, N., Li, A., Wu, D., Xie, P., Liu, C., Van Roozendael, M., Stavrakou, T., Müller, J.-F., and Wagner, T.: 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, 5007–5033, https://doi.org/10.5194/acp-17-5007-2017, 2017.

Wang, Z., Wang, G., Guo, X., Bai, Y., Xu, Y., and Dai, M.: Spatial reconstruction of long-term (2003–2020) sea surface pCO_2 in the South China Sea using a machine-learning-based regression method aided by empirical orthogonal function analysis, Earth Syst. Sci. Data, 15, 1711–1731, https://doi.org/10.5194/essd-15-1711-2023, 2023.

Wei, J., Li, Z., Li, K., Dickerson, R. R., Pinker, R. T., Wang, J., Liu, X., Sun, L., Xue, W., and Cribb, M.: Full-coverage mapping and spatiotemporal variations of ground-level ozone (O_3) pollution from 2013 to 2020 across China, Remote Sens. Environ., 270, 112775, https://doi.org/10.1016/j.rse.2021.112775, 2022.

Wenig, M., Spichtinger, N., Stohl, A., Held, G., Beirle, S., Wagner, T., Jähne, B., and Platt, U.: Intercontinental transport of nitrogen oxide pollution plumes, Atmos. Chem. Phys., 3, https://doi.org/10.5194/acp-3-387-2003, 2003.

Wu, Y., Di, B., Luo, Y., Grieneisen, M. L., Zeng, W., Zhang, S., Deng, X., Tang, Y., Shi, G., Yang, F., and Zhan, Y.: A robust

approach to deriving long-term daily surface NO_2 levels across China: Correction to substantial estimation bias in back-extrapolation, Environ. Int., 154, 106576, https://doi.org/10.1016/j.envint.2021.106576, 2021.

Xia, M. and Jia, K.: Reconstructing Missing Information of Remote Sensing Data Contaminated by Large and Thick Clouds Based on an Improved Multitemporal Dictionary Learning Method, IEEE Trans. Geosci. Remote Sensing, 60, 1–14, https://doi.org/10.1109/TGRS.2021.3095067, 2022.

Xu, R., Tong, D., Xiao, Q., Qin, X., Chen, C., Yan, L., Cheng, J., Cui, C., Hu, H., Liu, W., Yan, X., Wang, H., Liu, X., Geng, G., Lei, Y., Guan, D., He, K., and Zhang, Q.: MEIC-global-CO_2: A new global CO_2 emission inventory with highly-resolved source category and sub-country information, Sci. China Earth Sci., 67, 450–465, https://doi.org/10.1007/s1143002312303, 2024.

Zhai, B. and Chen, J.: Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China, Sci. Total Environ., 635, 644–658, https://doi.org/10.1016/j.scitotenv.2018.04.040, 2018.

Zhang, W., Quan, H., and Srinivasan, D.: Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination, Energy, 160, 810–819, https://doi.org/10.1016/j.energy.2018.07.019, 2018.

Zhou, W., Peng, B., and Shi, J.: Reconstructing spatial-temporal continuous MODIS land surface temperature using the DINEOF method, J. Appl. Remote Sens., 11, 1–15, https://doi.org/10.1117/1.JRS.11.046016, 2017.

Zhou, W., Qin, K., He, Q., Wang, L., Luo, J., Xie, W.: Comparison and Optimization of Ground-Level NO_2 Concentration Estimation in China Based on TROPOMI and OMI. Acta Opt. Sin., 44(6): 0601010, https://dx.doi.org/10.3788/AOS231013, 2024.