Sub-city Scale Hourly Air Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements

C. Malings[1,2,3], K. E. Knowland[2,3], C. A. Keller[2,3], and S. E. Cohn[2]

[1]NASA Postdoctoral Program Fellow, Goddard Space Flight Center, Greenbelt, MD, USA.

[2]Global Modeling and Assimilation Office, Goddard Space Flight Center, Greenbelt, MD, USA.

[3]Universities Space Research Association, Columbia, MD, USA.

Corresponding Author: Carl Malings (carl.a.malings@nasa.gov)

Key Points:

- Multiple air quality data sources (GOES-CF model, TROPOMI satellite, EPA monitors) are combined to improve city-scale NO2 forecasts.
- Forecasts using combined data outperform forecasts using ground-based measurements only.
- Updating of forecasts based on residuals against the most recent ground measurements further improves short-term forecasting.

Abstract

While multiple information sources exist concerning surface-level air pollution, no individual source simultaneously provides large-scale spatial coverage, fine spatial and temporal resolution, and high accuracy. It is therefore necessary to integrate multiple data sources, using the strengths of each source to compensate for the weaknesses of others. In this paper, we propose a method incorporating outputs of NASA's GEOS Composition Forecasting model system with satellite information from the TROPOMI instrument and ground measurement data on surface concentrations. Although we use ground monitoring data from the EPA network in the continental United States (US), the model and satellite data sources used have the potential to allow for global application. This method is demonstrated using surface measurements of nitrogen dioxide as a test case in regions surrounding five major US cities. The proposed method is assessed through cross-validation against withheld ground monitoring sites. In these assessments, the proposed method demonstrates major improvements over two baseline approaches which use ground-based measurements only. Results also indicate the potential for near-term updating of forecasts based on recent ground measurements.

Plain Language Summary

Air quality is a major health concern worldwide, leading to millions of premature deaths annually. In order to better understand this risk and mitigate its impacts, there are numerous sources of information about air quality. These include ground-based measurement stations, satellites, and global air quality models. By combining these data sources together, we can use the strengths of each source to compensate for the weaknesses of others. This paper presents one method of combining these data sources and uses it to make air quality forecasts over five US cities up to 24 hours in advance. These forecasts are compared to pollution estimates made using ground-based measurement data only to see how integrating additional data sources improves the forecast. Overall, we find that there are large increases in accuracy of forecasting using the proposed method, and that further improvements can be made by comparing the forecasts to the most recent ground-based measurements and making some more final adjustments. Methods like this, which use a combination of globally available satellite and model data together with some local measurements, can be applied to different types of air pollution in all regions of the world, thereby improving our understanding of air pollution globally.

1 Introduction

Air pollution is recognized as one of the leading risk factors of human mortality worldwide, and its relative impact has been increasing in recent years (Brauer et al., 2012, 2016; Forouzanfar et al., 2015; Cohen et al., 2017). Monitoring of air quality has traditionally been conducted by government regulatory agencies, such as the United States (US) Environmental Protection Agency (EPA), operating networks of fixed monitoring stations (Snyder et al., 2013). These networks typically focus on assessing the background levels in urban areas, along with some known major emission sources such as industrial facilities and highways (Chow, 1995). The relatively high setup and operating cost of these networks limits the number of monitoring stations which can feasibly be deployed. On the other hand, air pollutant concentrations can vary greatly in space, especially in urban areas with a large number and variety of pollutant sources in close proximity (Marshall et al., 2008; Karner et al., 2010; Tan et al., 2014). This variability means that air quality estimates based on these traditional monitoring stations may underrepresent the variability or extremes in local air pollution (Jerrett, Burnett, et al., 2005). There are several techniques to extend the data collected by these limited monitoring sites to better represent air quality over a region. These include proximity-based or statistically interpolated methods, e.g., “kriging”, and land use regression approaches using assumed relationships between land use characteristics and pollutant concentrations and extrapolating these beyond the measurement sites (see Jerrett, Arain, et al., 2005 for an overview). However, these approaches have limitations: simpler statistical models of air pollutants, such as land use regression, tend to generalize poorly to regions different from those where they were developed (Hoek et al., 2008; Liu et al., 2012) and cannot account for transient pollution events, e.g., from fires or dust storms.

Besides relying on ground-based regulatory monitors alone, alternative sources of air quality data can be considered. These include low-cost sensors, which can be deployed in greater numbers for a comparable cost, thereby increasing the spatial density and coverage for data collection (Snyder et al., 2013; Loh et al., 2017; Turner et al., 2017). However, low-cost sensors require careful calibration against available regulatory-grade instruments to ensure sufficient data quality (Popoola et al., 2016; Malings et al., 2019). The use of satellite data to inform local air quality estimation is also a promising area of work (e.g., Engel-Cox et al., 2004; Han et al., 2018; Lyapustin & Wang, 2018; Cooper et al., 2020). Satellite instruments are limited, however, to observe only during cloud-free daylight conditions, and typically measure pollutant concentrations integrated over the atmospheric column (see Duncan et al., 2014 for an overview). Finally, there are modeling approaches ranging from gaussian plume dispersion models to full atmospheric chemistry simulation models (Jerrett, Arain, et al., 2005). These sophisticated approaches require knowledge of emission sources and rates, and are typically computationally intensive, especially at high spatial resolutions (C. A. Keller et al., 2014; Hu et al., 2018). This can preclude their use in certain areas where the necessary input information and computational resources do not yet exist. However, they have the potential to produce global concentration estimates as well as forecasts of near-future conditions.

There are many possibilities to combine these different sources of information to improve the spatial and/or temporal resolution of air quality estimates. For example, information on NO_2 vertical profiles from the GEOS-Chem global atmospheric chemistry model (Bey et al., 2001) was combined with tropospheric column NO_2 concentration data from the Ozone Monitoring Instrument (OMI) aboard the US National Aeronautics and Space Administration (NASA) Aura satellite, resulting in a better correlation to measured daily-average surface concentrations, despite some remaining bias (Lamsal et al., 2008). Machine learning approaches have been used to refine the predictions of global atmospheric chemistry models to better match the records of specific measurement stations, improving location-specific forecasts (Christoph A. Keller et al., 2020). Low-cost sensors have been used together with aerosol optical depth (AOD) data from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS) satellite instruments to produce estimates of surface-level fine particulate matter mass ($PM_{2.5}$), with the low-cost sensor networks functioning nearly as well as the sparser regulatory-grade networks when used for this purpose (Gupta et al., 2018; Malings et al., 2020). These approaches still only provided information about the situation at the satellite overpass times, however. Estimates of the “typical” air quality in a region, derived from fine-resolution pollutant dispersion models, have been updated with low-cost sensor data for near-real-time air quality mapping (Schneider et al., 2017; Ahangar et al., 2019). Regional-scale atmospheric chemistry models have also been used together with MODIS AOD, surface-level EPA monitoring data, and other information such as land usage and meteorology to produce daily-average surface $PM_{2.5}$ estimates at one-kilometer spatial resolution over the southeastern and eastern US (Friberg et al., 2016; Goldberg et al., 2019; Murray et al., 2019; Just et al., 2020). These estimates were highly correlated with the EPA measurements during cross-validation. Similar approaches have been applied at a global scale for estimating annual-average $PM_{2.5}$ concentration, although accuracy of the method was regionally-dependent (van Donkelaar et al., 2010; Shaddick et al., 2018). Much recent research has focused on one-kilometer daily-average surface $PM_{2.5}$ estimation combining similar data sources (Cleland et al., 2020; Danesh Yazdi et al., 2020; Just et al., 2020; Mhawish et al., 2020), with some research into hourly-average concentration estimation (Jiang et al., 2021) and into forecasting daily averages (Zhang et al., 2020). Similar efforts include regional forecasting of coarse particulate matter (Michaelides et al., 2017) and global estimation of 8-hour maximum surface ozone concentrations (Chang et al., 2019) by combining model, satellite, and/or ground data.

Building on this previous work, this paper proposes and demonstrates an approach for using globally-available atmospheric composition historical estimates and forecasts and satellite information together with localized surface measurements for generating sub-city-scale and hourly resolution estimates and near-term forecasts up to 24 hours in advance of surface-level pollutant concentrations relevant for air quality. We make use of the Global Earth Observing System Composition Forecasting (GEOS-CF) atmospheric chemistry model system and satellite data from the TROPOspheric Monitoring Instrument, TROPOMI. Although this paper focuses on surface NO_2 across several US cities as a case study, the data sources and methods are broadly applicable to different pollutants of interest and for any location worldwide with surface-level monitoring. While being generally applicable, the proposed methods are intended for targeted application to limited spatial and temporal domains, since previous results indicate that the relationships between ground concentrations and model outputs or satellite retrievals vary in space and time, which limits the generalizability of any specific derived relationship. Finally, the proposed approach does not combine data sources to improve retrospective air quality analyses, as has been the focus of much previous work, but instead examines how these combined data can better inform near-term forecasting of air quality at fine spatial and temporal resolutions. The data sources used are discussed in Section 2, and the methods of their integration are discussed in Section 3. The performance of these methods is evaluated as outlined in Section 4, with the results presented in Section 5. Section 6 presents some general conclusions and discussion of areas for future work.

2 Data Sources

2.1 GEOS-CF surface Nitrogen Dioxide concentration

The GEOS-CF system couples the GEOS model with GEOS-Chem chemistry module (Bey et al., 2001; Eastham et al., 2014; C. A. Keller et al., 2014; Long et al., 2015). It uses the increments from an assimilated meteorological product from a near-real time GEOS numerical weather prediction system (Orbe et al., 2017) in order to produce global estimates and five-day forecasts of concentrations for several chemicals of interest for atmospheric chemistry and air quality (Hu et al., 2018; Knowland et al., 2020). Outputs are gridded to $0.25^{\circ} \times 0.25^{\circ}$, roughly 25×25 km^2. For this project, the hourly-average surface concentrations of NO_2 are used. GEOS-CF global estimates are available since 1 January 2018 and forecasts since 1 January 2019 (see Knowland et al., 2020). It should also be noted that, in its current configuration, there is no direct chemical data assimilation within GEOS-CF. Instead, the system simulates the emission, transportation, chemical evolution and deposition of atmospheric pollutants, taking the state of the atmosphere from the GEOS outputs.

2.2 TROPOMI tropospheric Nitrogen Dioxide concentration

Data from TROPOMI aboard the European Space Agency's Copernicus Sentinel-5 Precursor satellite are used to provide remote-sensing estimates of tropospheric column NO_2 concentrations ("TROPOMI Level 2 Nitrogen Dioxide," n.d.). TROPOMI is considered to be a successor to OMI, with a finer spatial resolution, nominally 7×3.5 km^2 (Veefkind et al., 2012). Values for NO_2 are available as part of the Level-2 data product since July 2018, with the current operational version in service since June 2019. Satellite overpasses occur at approximately 13:30 local solar time. The resulting observed patterns are therefore likely to be representative of daytime concentrations, but may not capture the heavily traffic-influenced conditions of the morning and afternoon rush-hours. For data quality assurance (QA), pixels with provided QA values above 0.5 are used, as recommended for "good" quality data. For this application, data are re-gridded to a $0.05^o \times 0.05^o$ grid by averaging together all valid pixels falling within each grid cell for each satellite overpass.

2.3 EPA ground Nitrogen Dioxide monitoring data

The "ground truth" for NO_2 concentrations in this project is provided by regulatory-grade air quality monitoring stations in the US, with data collected by the EPA. These stations are usually sited in or near urban areas and major pollutant sources to monitor compliance with the Clean Air Act. Chemiluminescent analyzers remain the recommended method for quantifying ambient $NO_{2,}$ despite some known interference from other reactive nitrogen compounds (US EPA, 2017). A measurement accuracy within 15% is recommended for all regulatory-grade monitor data (Williams et al., 2014); for typical US ambient NO_2 concentrations, this would correspond to an accuracy on the order of 1 ppb.

The application areas considered for this paper are listed in Table 1; maps of these areas are provided in the supplemental information, Figure S1. These areas are $2^o \times 2^o$ domains, representing several large US cities and their surrounding metropolitan areas. Application to such restricted domains is important to minimize the impact of spatial variability in surface-to-satellite concentration relationships and to limit the effect of multiple time zones which would "spread out" diurnal signals such as rush-hour traffic emissions. For the current work, analysis is focused on Las Vegas, New Orleans, New York City, Salt Lake City, and San Francisco. These areas were chosen to represent different regions across the country with relatively large numbers of EPA NO_2 monitoring sites, which facilitate evaluation of the urban-scale air quality estimation and forecasting abilities of the proposed methods. EPA ground data collected during the calendar month of September 2019 in each area are used. This month is considered as a candidate for a "typical" month of the year since it is usually neither a minimum nor maximum for NO_2 in the US (Lamsal et al., 2010). This year is selected due to the availability of GEOS-CF forecasts and the current operational version of the TROPOMI data product.

Table 1. Designated analysis areas considered in this paper.

Area Name	Lower-Left Corner	Upper-Right Corner	EPA NO_2 sites Active Sept. 2019
Las Vegas	35°N, 116°W	37°N, 114°W	5
New Orleans	29°N, 92°W	31°N, 90°W	8
New York City	40°N, 75°W	42°N, 73°W	14
Salt Lake City	40°N, 113°W	42°N, 111°W	15
San Francisco	37°N, 123°W	39°N, 121°W	27

3 Data Fusion Methodology

A representation of the proposed scheme for surface concentration estimation and forecasting is presented in Figure 1. The idea is to use outputs from a global atmospheric chemistry model to drive estimates and forecasting at a coarse spatial resolution. Information from other data sources, especially satellites, is then incorporated to help resolve finer spatial variabilities. Ultimately, ground-based measurement data are used to establish a relationship between the model and satellite-derived spatial patterns and observed surface concentration levels during a specified calibration period $T_{calibration}$ leading up to the current time, $t_{current}$. A seven-day calibration period is used in this paper. This length was chosen as a compromise between having too short a period, during which there might be too few satellite passes to extract a robust pattern, and too long a period, during which the extracted typical pattern might be subject to change and important temporary spatial patterns smoothed out. Relationships established for $T_{calibration}$ are extrapolated forward in time to support predictions of surface concentrations at t_{target} in the near future, e.g., within a day of $t_{current}$.

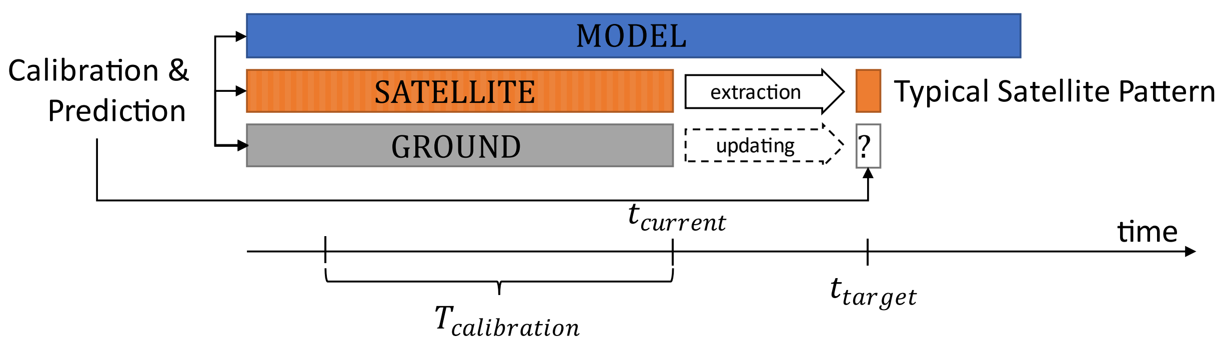

Figure 1. Overview of the proposed approach to integrate model, satellite, and ground data sources for short-term forecasting of surface concentrations.

The implementation of this general approach is divided into several stages, described in the following sections. First, in Section 3.1, the GEOS-CF model outputs are downscaled to a finer target resolution, i.e., that of the satellite data. Second, in Section 3.2, the TROPOMI satellite data collected during the calibration period are compared to the model's estimates for the same period. This comparison identifies a "typical pattern" in these data, which is assumed to remain valid until at least t_{target}. This pattern is combined with the model's estimates during the calibration period in Section 3.3, and these combined estimates are compared with EPA ground measurement data. This comparison establishes model-to-ground-truth and/or satellite-to-ground-truth relationships via linear regression in Sections 3.4 and 3.5. These relationships are assumed to persist into the near future. The model's forecast for t_{target} is combined with the typical satellite pattern and adjusted using established relationships to ground data, providing surface concentration predictions for that time in Section 3.6. An optional final update to these predictions, based on correlations with the most recent ground measurement data, is discussed in Section 3.7.

3.1 Downscaling of model outputs

Initial downscaling of the spatially coarse model is done in one of two ways. These are demonstrated in Figure 2 using the monthly average GEOS-CF surface NO_2 concentration fields for September 2019 over the New York City area. First, a naïve or "nearest-grid-point" interpolation method assigns the value at any fine-resolution grid point to be the same as the value at the nearest grid point at the original resolution. This results in a field sub-divided into squares around the model grid points, with abrupt changes at the boundaries, as in Figure 2a. An advantage of this method is that it preserves spatial averages, i.e., an area-averaged value will be the same before and after interpolation. However, since these values are to be rescaled anyway as part of the proposed approach, this may not be a useful feature here. Second, a bi-linear interpolation can be applied to the two-dimensional surface grid to produce linearly interpolated concentration estimates over the new grid. The smoothed field resulting from linear interpolation features more physically realistic gradual changes in surface concentration, as in Figure 2b.

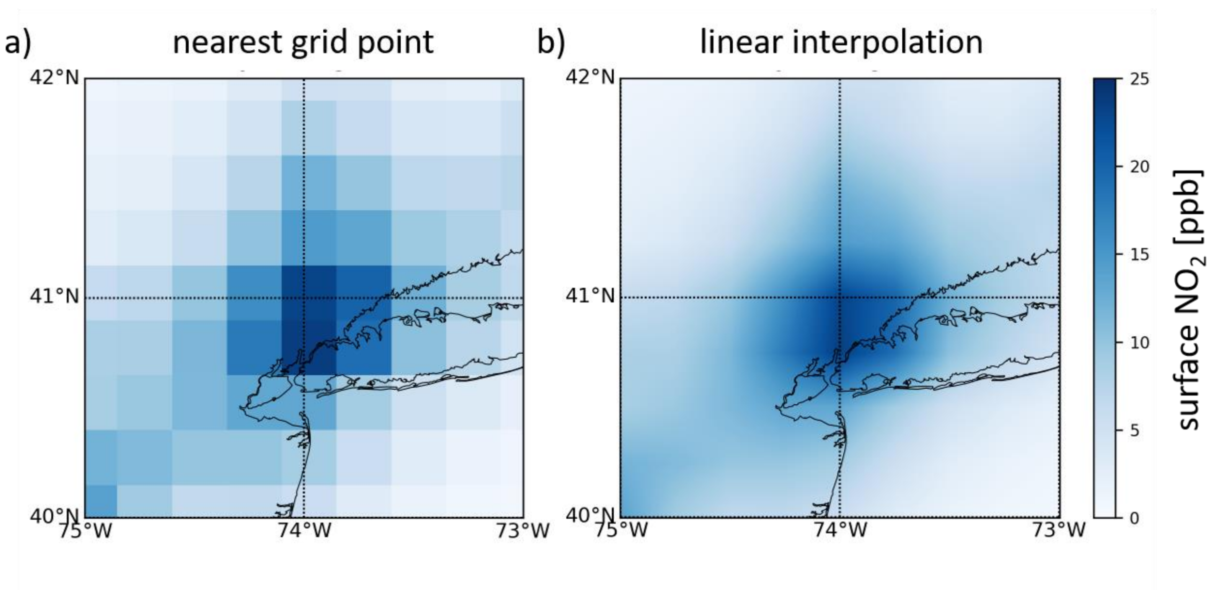

Figure 2. Comparison of representative model-predicted GEOS-CF surface NO_2 concentrations for September 2019 downscaled to higher resolution using either a nearest-grid-point method (**a**) or linear interpolation (**b**) for New York City.

During preliminary testing, it was found that in most cases, with all other factors being equal, linear interpolation outperforms nearest-point interpolation by a slight margin in terms of the ultimate quality of the surface concentration forecasts; see the supplemental information Section S2.1 for details. This method for downscaling is therefore preferred and used for the results presented in Section 5. The resolution to which the model is downscaled depends on the ultimate desired resolution. At a minimum, the resolution must be increased to match the highest resolution data source being used. Interpolation should also be performed to all locations at which ground measurement data are available. Let $f_{MODEL}(x,t)$ denote the model's interpolated estimate of the ground-level hourly-average NO_2 concentration at spatial location x and time t.

3.2 Extraction of typical patterns

Let $f_{SAT}(x,t)$ denote a satellite-retrieved quantity at location x and time t. In the proposed method, data collected for a specified calibration interval, $T_{calibration}$, are used to define a typical satellite pattern map, $\overline{f_{SAT}}(x)$. This is done by averaging:

$$\overline{f_{SAT}}(x) = \frac{1}{n_{SAT}} \sum_{t \in T_{calibration}} f_{SAT}(x,t) \tag{1}$$

where n_{SAT} is the number of timesteps during $T_{calibration}$ over which $f_{SAT}(x,t)$ data are available, i.e., the number of satellite overpasses which occur during the calibration interval. Note that missing satellite data, e.g., due to cloud cover, are ignored.

The data source $f_{SAT}(x,t)$ may represent a different quantity of interest than $f_{MODEL}(x,t)$, e.g., tropospheric column versus ground-level NO_2 concentrations. Instead of using $\overline{f_{SAT}}(x)$ directly, it is re-scaled to best match the values of $f_{MODEL}(x,t)$ for the same period of time. To do this, a typical model pattern $\overline{f_{MODEL}}(x)$ is first extracted. This can be done in two different ways. One method obtains the "full" average of the calibration time period:

$$\overline{f_{MODEL}}_{\text{full}}(x) = \frac{1}{n_{MODEL}} \sum_{t \in T_{calibration}} f_{MODEL}(x,t) \tag{2}$$

where n_{MODEL} denotes the number of model timesteps during the calibration period. Alternatively, the model average can be "restricted" to only those times and locations where data are available from both sources. This is evaluated as:

$$\overline{f_{MODEL}}_{\text{restricted}}(x) = \frac{\sum_{t \in T_{calibration}} f_{MODEL}(x,t)\mathbb{I}(\exists f_{SAT}(x,t))}{\sum_{t \in T_{calibration}} \mathbb{I}(\exists f_{SAT}(x,t))} \tag{3}$$

where $\mathbb{I}(\cdot)$ takes value 1 when the argument is true and 0 otherwise. Its argument, $\exists f_{SAT}(x,t)$, is used to determine whether there exists ($\exists$) a valid datapoint from $f_{SAT}(x,t)$ at location x and time t, i.e., whether a satellite pass occurs during that timestep of the model and whether there are valid cloud-free data from the satellite for that timestep.

Regardless of which method is used to obtain $\overline{f_{MODEL}}(x)$, the final step is to re-scale $\overline{f_{SAT}}(x)$ to better match $\overline{f_{MODEL}}(x)$. This is done via ordinary least squares linear regression across the spatial domain of the calibration, $X_{calibration}$, with $\overline{f_{SAT}}(x)$ as the independent variables and $\overline{f_{MODEL}}(x)$ as the dependent variables. The resulting regression is denoted $\theta_{SAT \to MODEL}$, and the process of regression is denoted:

$$\theta_{SAT \to MODEL} \leftarrow \text{regress } \overline{f_{SAT}}(x) \text{ to } \overline{f_{MODEL}}(x)\ \forall\ x \in X_{calibration} \tag{4}$$

The final extracted pattern $\overline{\overline{f_{SAT}}}(x)$, representing the difference between the re-scaled $\overline{f_{SAT}}(x)$ and $\overline{f_{MODEL}}(x)$, is:

$$\overline{\overline{f_{SAT}}}(x) = \theta_{SAT \to MODEL}\left(\overline{f_{SAT}}(x)\right) - \overline{f_{MODEL}}(x) \tag{5}$$

Passes of TROPOMI occur at certain local times of day, and therefore any information captured can only represent spatial patterns present at those times. This can introduce a bias with respect to the true average spatial pattern if these are compared directly. Using the "full" approach of Equation 2, the systematic bias between the spatial pattern as determined by the satellite at the overpass times and the spatial pattern as determined from the model throughout the entire calibration period is incorporated into $\overline{\overline{f_{SAT}}}(x)$. Using the "restricted" approach of Equation 3, this bias is not incorporated, and only the difference at the overpass times is captured by $\overline{\overline{f_{SAT}}}(x)$. This is then assumed to be representative of these difference throughout the day.

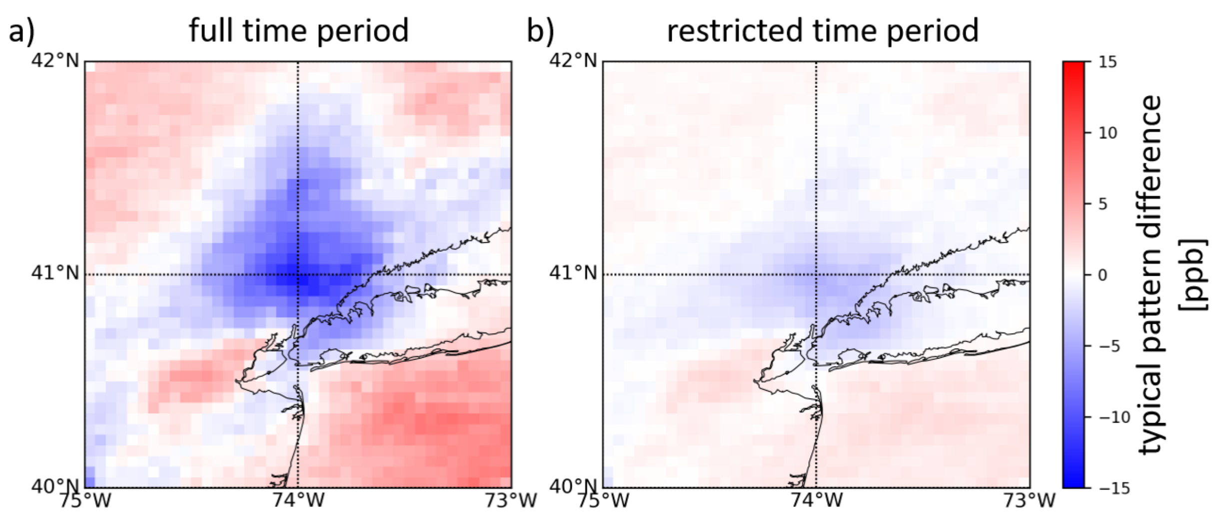

Figure 3. Patterns of systematic differences between rescaled TROPOMI satellite measurements and model-derived concentrations during an example calibration period. Rescaling was done considering either the entire calibration period, as in Equation 2 (**a**), or only the times of the satellite overpasses, as in Equation 3 (**b**).

Figure 3 depicts example $\overline{\overline{f_{TROPOMI}}}(x)$ patterns derived using the "full" and the "restricted" approaches. While the spatial patterns are similar, using the "full" approach leads to larger magnitudes in the pattern intensity, as in Figure 3a, compared to the "restricted" approach, as in Figure 3b. This is to be expected, since there is a better overall match between the spatial patterns when the averaging is restricted to satellite overpass times only.

Preliminary testing was conducted to determine which of these methods of typical pattern extraction should be used. The results were mixed; details are provided in the supplemental information, Section S2.1. For New Orleans, there was a clear improvement in correlation for patterns extracted at satellite overpass time as in Equation 3. Alternatively, for Las Vegas, there were reductions in average error and bias using the full calibration period to extract patterns as in Equation 2. For New York City, Salt Lake City, and San Francisco, results were substantially similar for either approach. Due to the rather large differences in correlation observed for New Orleans and the relatively small number of ground verification sites available in Las Vegas, as noted in Table 1, the method of Equation 3 is slightly preferred, and used for the results presented in Section 5.

3.3 Combination of model and typical patterns

Once typical patterns of the satellite data are extracted, these can be combined with the downscaled model estimates and forecasts. In the proposed method, this is done by direct addition:

$$f_{MODEL}(x,t) + \overline{\overline{f_{SAT}}}(x) \tag{6}$$

Note that this approach assumes that differences between model-predicted surface concentrations and true concentrations are constant. Alternatively, these patterns might be combined via linear regression, which would allow for the intensity of these differences to vary via tuning of the regression parameters. In comparing the use of patterns as regression inputs versus their direct addition as in Equation 6 during preliminary testing, there is a slight preference towards the combination of patterns via addition; see the supplemental information Section S2.1 for details. Furthermore, the combination of patterns via Equation 6 requires fewer free parameters compared to combination via regression. It may be that there were insufficient data during the calibration period to establish a robust regression to allow that combination approach to perform sufficiently well, and so simple addition achieved more stable performance. Regardless, Equation 6 is used in the results presented in Section 5.

3.4 Calibration to ground data

Next, a linear relationship is established during the calibration period between the indirect data sources, i.e., the model- and satellite-derived patterns, and the direct data source, i.e., the ground measurement data. The relationship established for this period is expressed as:

$$\theta_{INPUT\rightarrow GROUND} \leftarrow \text{regress } f_{INPUT}(x,t) \text{ to } f_{GROUND}(x,t) \;\; \forall\, x \in X_{ground}, t \in T_{calibration} \tag{7}$$

where the regression relationship $\theta_{INPUT\rightarrow GROUND}$ is developed by regressing the various input data sources $f_{INPUT}(x,t)$ as independent variables to the target $f_{GROUND}(x,t)$ dependent variables. The regression uses data collected during the calibration period $T_{calibration}$ and restricted to the sites where ground measurements are available during this time, X_{ground}. In this general formulation, $f_{INPUT}(x,t)$ is a stand-in for various data sources and/or combinations of sources. For example, using the sum of the model and satellite pattern data, as in Equation 6, as the input is denoted:

$$\theta_{MODEL+SAT\rightarrow GROUND} \leftarrow \text{regress } \left(f_{MODEL}(x,t) + \overline{\overline{f_{SAT}}}(x)\right) \text{ to } f_{GROUND}(x,t) \tag{8}$$

Note that additional regression inputs can also be considered within this framework. During preliminary testing, meteorological information such as temperature, relative humidity, planetary boundary layer height, and wind from the GEOS-CF system were considered as independent variables. Additionally, information on nighttime light intensity during the calibration period, as a representation of human activity, was considered as a possible proxy or predictor for surface NO_2 concentrations. Here, nighttime light intensity as measured by a day-night band sensor of the NASA Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership (SNPP) satellite was used for this purpose (NASA VIIRS Land Science Investigator-Led Processing System, 2019; Román et al., 2018). However, in both cases, there were no clear improvements over the use of the methodology without these additional data sources; see the supplemental information Section S2.2 for details. Results from these variations on the proposed methodology are therefore not presented in Section 5, and are only briefly discussed in Section 6.

3.5 Weighting schemes for calibration

In establishing linear regression relationships to ground data during the calibration period, different weighting schemes are used. These schemes can increase the relative emphasis placed on different subsets of data within the calibration period, in order to coerce the resulting regression to better represent these subsets. Various possible time-based weighting schemes are proposed here, in which weight varies as a function of prediction lead time t_{lead}, or the difference between the time at which the data are collected and the target prediction time t_{target}. Besides a null weighting scheme, where equal weight is given to all calibration data, four different time-varying weighting schemes are proposed, as illustrated in Figure 4.

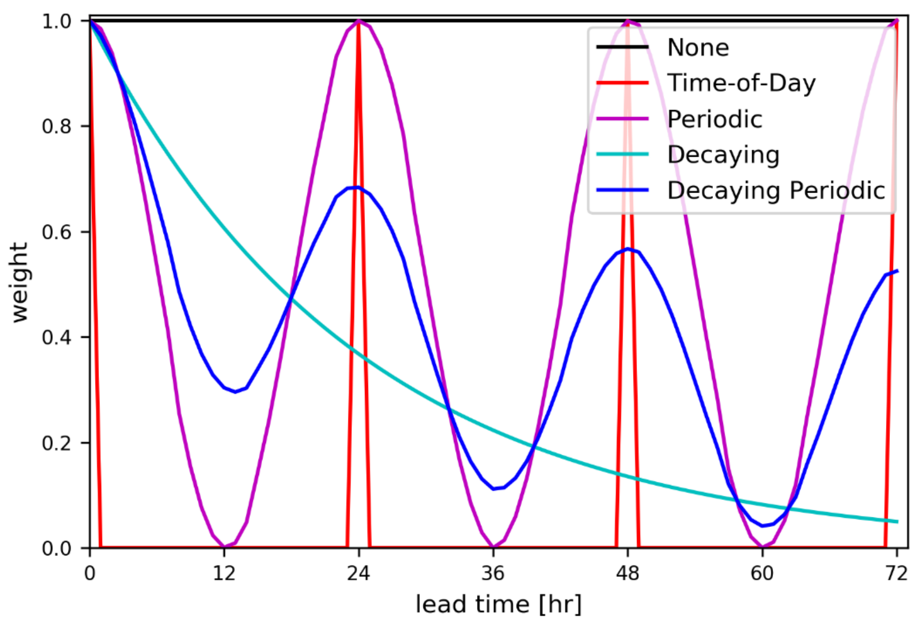

Figure 4. Different weighting schemes for linear regression to ground data.

First, the "time-of-day" weighting values only times separated from t_{target} by multiples of 24 hours, i.e., only data collected at the same time of day are to be used in establishing the linear regression, as depicted by the red line in Figure 4. The "periodic" weighting scheme follows a similar logic, but uses sinusoidally varying weight with a 24-hour period, shown as the purple line in Figure 4:

$$\mathrm{w}_{\mathrm{periodic}}(t_{lead}) = \cos^2\left(\pi \frac{t_{lead}}{24\ \mathrm{hrs}}\right) \tag{9}$$

Another approach uses a "decaying" weight function, where weight decreases monotonically with lead time. In this case, exponential decay with a scale parameter of 24 hours is used, shown by the cyan line in Figure 4:

$$\mathrm{w}_{\mathrm{decay}}(t_{lead}) = \exp\left(-\frac{|t_{lead}|}{24\ \mathrm{hrs}}\right) \tag{10}$$

Finally, a "decaying periodic" weight function combines periodicity with a decay rate, such that peaks occur at the same time of day as t_{target} but decrease in magnitude as lead time increases. This is plotted as the blue line in Figure 4, whose formula is:

$$\mathrm{w}_{\text{decaying periodic}}(t_{lead}) = 0.5 \exp\left(-\frac{|t_{lead}|}{24\text{ hrs}}\right) + 0.5 \cos^2\left(\pi \frac{t_{lead}}{24\text{ hrs}}\right) \tag{11}$$

Averaging rather than multiplication is used in order for there to be non-zero weightings at all lead times. Note that any regression $\theta_{INPUT \to GROUND}$ is specific both to the calibration period from which it is derived and, in the case of these weighted regression schemes, to a prediction lead time.

Among these weighting schemes, the decaying periodic scheme of Equation 11 gave the best and most robust performance across all regions during preliminary testing, and so is slightly preferred here; see the supplemental information Section S2.1 for details. Notably, however, the time-of-day approach to regression weighting led to noticeably better correlation in Salt Lake City and reduced bias in Las Vegas compared to the decaying periodic regression weighting scheme. Therefore, for those two areas, this weighting scheme is preferred. The exact reason for this difference is unclear but may be related to these areas being further inland than the other more coastal areas which are examined.

3.6 Surface concentration prediction

The regression relationship $\theta_{INPUT \to GROUND}$ can now be used to estimate the surface concentration at any time and location of interest, given appropriate input information.

$$\hat{f}(x_{target}, t_{target}) = \theta_{INPUT \to GROUND}\left(f_{INPUT}(x_{target}, t_{target})\right) \tag{12}$$

For example, following Equation 8:

$$\hat{f}(x_{target}, t_{target}) = \theta_{MODEL+SAT \to GROUND}\left(f_{MODEL}(x_{target}, t_{target}) + \overline{\overline{f_{SAT}}}(x_{target})\right) \tag{13}$$

Note that the input data sources should be matched to the calibration period, e.g., $\overline{\overline{f_{SAT}}}(x)$ should be extracted during $T_{calibration}$, while the regression relationship $\theta_{MODEL+SAT \to GROUND}$ should correspond to the same calibration period and to the target lead time if a time-varying weight scheme is used.

3.7 Updating predictions using correlation of ground data

In the proposed method so far, all information on the spatial distribution of surface pollutants is obtained indirectly, i.e., as GEOS-CF model outputs or satellite retrievals from TROPOMI. The direct ground measurements from EPA monitoring sites are only used to appropriately scale these data to better represent surface conditions. However, additional information can be extracted from the ground data directly. A method for this is outlined here, inspired by spatio-temporal kriging (Cressie & Wikle, 2011).

It is assumed that the residuals between estimates derived by the methods outlined above and true surface concentrations can be modeled as Gaussian random variables with simple correlation structures based on spatial distances and temporal differences. The most recent ground measurement data can then be used to perform a Bayesian updating of these residuals, providing a final correction for the estimate. This correction, denoted

$\delta\big(\mathbf{f}_{GROUND}(t_{current}), \hat{\mathbf{f}}(t_{current}), x, t\big)$, is used to update the estimation provided from Equation 12:

$$\hat{f}(x,t) = \theta_{INPUT\to GROUND}\big(f_{INPUT}(x,t)\big) + \delta\big(\mathbf{f}_{GROUND}(t_{current}), \hat{\mathbf{f}}(t_{current}), x, t\big) \quad (14)$$

The correction term is evaluated using the set of ground measurements at the current time, reflecting the latest available ground measurement data:

$$\mathbf{f}_{GROUND}(t_{current}) = \{f_{GROUND}(x, t_{current}) \;\; \forall\, x \in X_{ground}\} \quad (15)$$

In addition, a prior estimate of the current surface concentration is required as input for the Bayesian updating. This is:

$$\hat{\mathbf{f}}(t_{current}) = \{\theta_{INPUT\to GROUND}\big(f_{INPUT}(x, t_{current})\big) \;\; \forall\, x \in X_{ground}\} \quad (16)$$

Note that if a weighted regression scheme is used, $\theta_{INPUT\to GROUND}$ must be appropriately matched to the target time, in this case $t_{current}$. It may therefore be different than the regression used in Equations 12 or 14.

The update is derived using a standard Bayesian scheme for multivariate Gaussian distributions, assuming zero prior mean:

$$\delta\big(\mathbf{f}_{GROUND}(t_{current}), \hat{\mathbf{f}}(t_{current}), x, t\big) = \mathbf{\Sigma}_A(\mathbf{\Sigma}_B + \mathbf{\Sigma}_C)^{-1}\left(\mathbf{f}_{GROUND}(t_{current}) - \hat{\mathbf{f}}(t_{current})\right) \quad (17)$$

Matrix $\mathbf{\Sigma}_A$ denotes the spatio-temporal covariance between each of the ground measurement sites at the current time and the location and time of interest:

$$\mathbf{\Sigma}_A = \{\sigma^2_{variability}\, \rho_{space}(x, x_{ground})\, \rho_{time}(t, t_{current}) \;\forall\, x_{ground} \in X_{ground}\} \quad (18)$$

Matrix $\mathbf{\Sigma}_B$ denotes the spatial covariance between the ground measurement sites:

$$\mathbf{\Sigma}_B = \{\sigma^2_{variability}\, \rho_{space}(x_{ground,1}, x_{ground,2}) \;\forall\, x_{ground,1}, x_{ground,2} \in X_{ground}\} \quad (19)$$

Since only ground measurements at $t_{current}$ are used, temporal correlation is assumed to be 1. Matrix $\mathbf{\Sigma}_C$ denotes covariance of the errors in the ground measurements. These errors are assumed to be independent with the same variance:

$$\mathbf{\Sigma}_C = \sigma^2_{measure}\mathbf{I} \quad (20)$$

where **I** is an identity matrix of appropriate size.

Exponentially decaying spatial and temporal correlation structures are assumed. These are, for space:

$$\rho_{space}(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|}{\lambda_{space}}\right) \quad (21)$$

and for time:

$$\rho_{time}(t_1, t_2) = \exp\left(-\frac{|t_1 - t_2|}{\lambda_{time}}\right) \quad (22)$$

where $\|x_1 - x_2\|$ denotes the distance between two points x_1 and x_2 on the Earth's surface, computed via the Haversine formula approximation (Sinnott, 1984). Spatial and temporal correlations are assumed to be independent, such that overall spatio-temporal correlation is the product of the spatial and temporal correlations.

These correlation structures, together with parameters λ_{space}, λ_{time}, $\sigma_{variability}$, and $\sigma_{measure}$, are determined by examining the residual correlation structures of the underlying method for surface concentration estimation. The estimated parameters used in this paper are listed in the supplemental information, Table S1. An example of the spatial correlations, in **Figure 5**a, portrays a trend of higher correlation at shorter distances, declining to zero net correlation as distances between ground evaluation sites increase. Temporal correlations, in **Figure 5**b, also express a trend of decreasing correlations as time differences increase. Although there is a slight perturbation in this trend at about 24 hours, indicating some day-to-day correlations of the residuals, this was ignored for the purposes of this investigation. Future work may capture these effects with more sophisticated correlation structures.

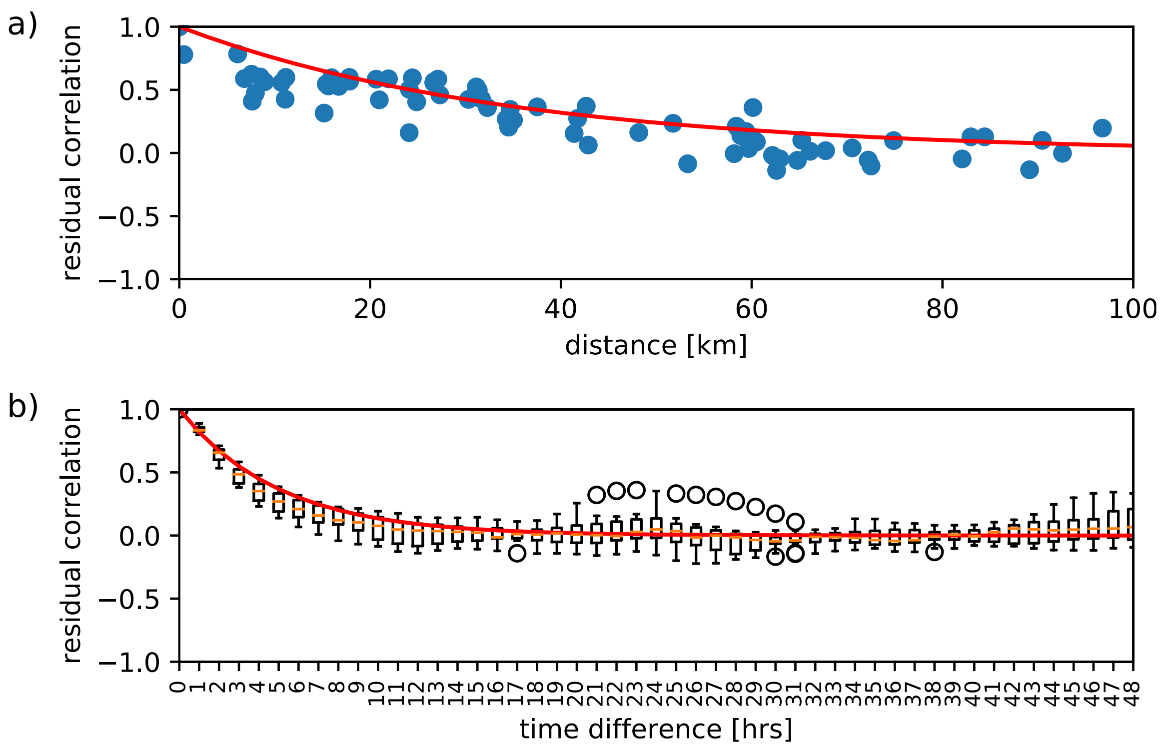

Figure 5. Spatial (**a**) and temporal (**b**) correlation of residuals for 1-hour-ahead prediction of surface concentrations using the proposed method of Sections 3.1-3.6 for New York City, September 2019. The fitted correlations are indicated by red lines.

4 Assessment Methodology

The methods described in Section 3 are tested to assess their ability to make accurate forecasts of surface-level NO_2 for the five city areas of Table 1 for the month of September 2019. To allow for a seven-day calibration period, performance evaluations begin on the eighth day and proceed hour-by-hour, with the calibration period covering the seven days prior to that hour, until the end of the month. For each hour, forecasts with various lead-times are made from that time forward, unless the forecast time falls after the end of the month, in which case no forecast is made. All data are aligned to an hourly timescale. For the EPA and GEOS-CF data, the native temporal resolutions are hourly-average. TROPOMI data are assigned to the hour during which the satellite pass occurs. Section 4.1 describes two baseline methods against which the proposed methods of Section 3 are to be compared, while Section 4.2 lists the metrics used to evaluate their performance.

4.1 Baseline methods

As a means of putting into context the performance of the surface concentration prediction approaches described in Section 3, two different baseline methods for making use of ground data only for estimation of surface pollutant concentrations are used. The “persistence” baseline assumes that the value at any point and time is the same as the most recent available measurement at the nearest ground measurement location:

$$\hat{f}_{persistence}(x,t) = f_{GROUND}(x_{nearest}, t_{latest}) \tag{23}$$

where $x_{nearest}$ is the closest ground monitor location to x in X_{ground} and t_{latest} is the most recent time in $T_{calibration}$, typically $t_{current}$.

The “climatology” baseline uses the measurement record of the nearest ground measurement location during the calibration period, and assumes that the value at any time is the same as the average value at that time of day, with the average being computed during the calibration period only:

$$\hat{f}_{climatology}(x,t) = \frac{\sum_{t' \in T_{calibration}} f_{GROUND}(x_{nearest}, t')\, \mathbb{I}\left(t' \in T_{time-of-day}(t)\right)}{\sum_{t' \in T_{calibration}} \mathbb{I}\left(t' \in T_{time-of-day}(t)\right)} \tag{24}$$

where $T_{time-of-day}(t)$ is a set of times at the same time of the day as t according to the temporal resolution being considered.

4.2 Performance metrics

To assess performance, for a given area, data from all but one ground site are allowed for use in calibration, while ground concentrations at the final site are estimated using the approach being tested. All ground sites in each area are cycled through in this manner, leading to one set of performance metrics being assessed for each ground site. The performance metrics assessed are the correlation coefficient (r), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Absolute Bias (AB). They are evaluated as follows:

$$\mathrm{r}(x) = \frac{\sum_{t \in T_{evaluation}} \left(\hat{f}(x,t) - \hat{f}_{mean}(x,t)\right)\left(f_{GROUND}(x,t) - \overline{f_{GROUND}}(x,t)\right)}{\sqrt{\sum_{t \in T_{evaluation}} \left(\hat{f}(x,t) - \hat{f}_{mean}(x,t)\right)^2} \sqrt{\sum_{t \in T_{evaluation}} \left(f_{GROUND}(x,t) - \overline{f_{GROUND}}(x,t)\right)^2}} \tag{25}$$

where

$$\hat{f}_{mean}(x,t) = \frac{1}{n_{evaluation}} \sum_{t \in T_{evaluation}} \hat{f}(x,t) \tag{26}$$

and

$$\overline{f_{GROUND}}(x,t) = \frac{1}{n_{evaluation}} \sum_{t \in T_{evaluation}} f_{GROUND}(x,t) \tag{27}$$

$$\mathrm{MAE}(x) = \frac{1}{n_{evaluation}} \sum_{t \in T_{evaluation}} \left|\hat{f}(x,t) - f_{GROUND}(x,t)\right| \tag{28}$$

$$\mathrm{RMSE}(x) = \sqrt{\frac{1}{n_{evaluation}} \sum_{t \in T_{evaluation}} \left(\hat{f}(x,t) - f_{GROUND}(x,t)\right)^2} \quad (29)$$

$$\mathrm{AB}(x) = \left|\frac{1}{n_{evaluation}} \sum_{t \in T_{evaluation}} \left(\hat{f}(x,t) - f_{GROUND}(x,t)\right)\right| \quad (30)$$

These metrics are also evaluated as a function of lead time, i.e., the difference between $t_{current}$ and t_{target}. Ten discrete lead times are investigated: 0, 1, 3, 6, 9, 12, 15, 18, 21, and 24 hours.

5 Results

The following sections present some key results regarding the performance of different combinations of data sources using the proposed methods described in Section 3, following the assessment methods of Section 4.

5.1 Comparison with baseline methods

Figure 6 compares the performance of the proposed method to that of the persistence baseline of Equation 23 for the different application areas. Here, the proposed method incorporates the GEOS-CF, TROPOMI, and EPA surface monitor data as outlined in Sections 3.1-3.6; the final updating of Section 3.7 has not yet been applied. For the persistence baseline, performance is fairly good at short lead times, but quickly drops off as the most recent measurements become increasingly outdated. There is a slight improvement again near the 24-hour lead time, due to similarities in diurnal profiles. In contrast, the proposed method has fairly consistent performance across lead times from about 3 to 24 hours. Performance is slightly better at short lead times, likely due to the calibration weighting schemes favoring such short-term performance. Although the performance of the proposed method at very short lead times is not as good as that of the persistence baseline by the correlation metric in some areas, overall, the proposed method dominates the persistence baseline for most areas and lead times by most metrics.

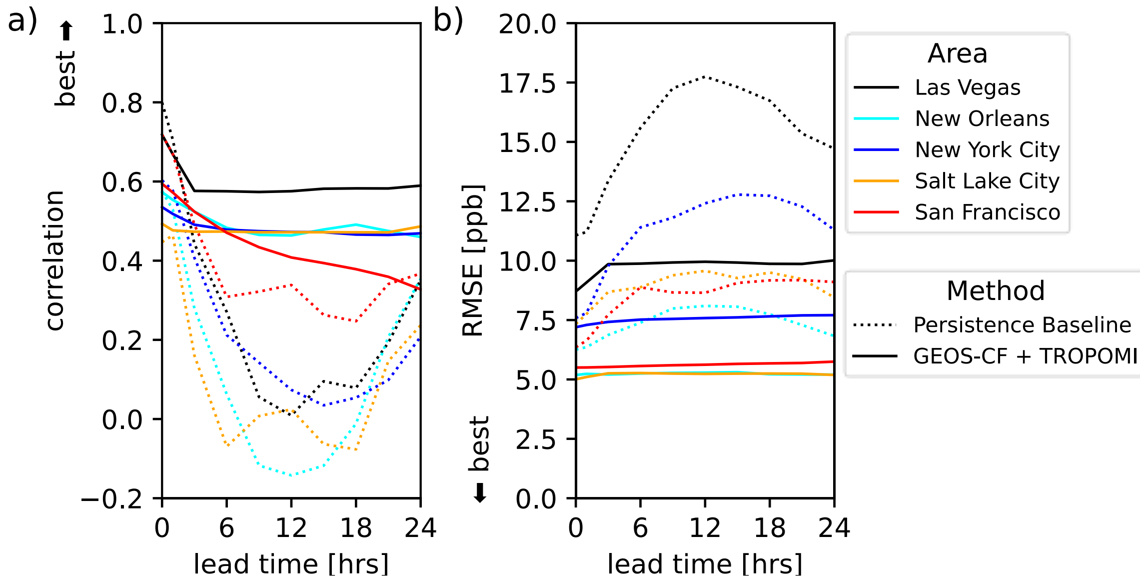

Figure 6. Comparison of the proposed method, in solid lines, to the persistence baseline, in dotted lines, for different color-coded application areas as a function of forecast lead time. Performance is presented in terms of the correlation (a) and RMSE (b) metrics; the direction of improved performance by each metric is indicated by the arrows adjacent to the vertical axes. The presented performance is the median performance across ground validation sites within each area.

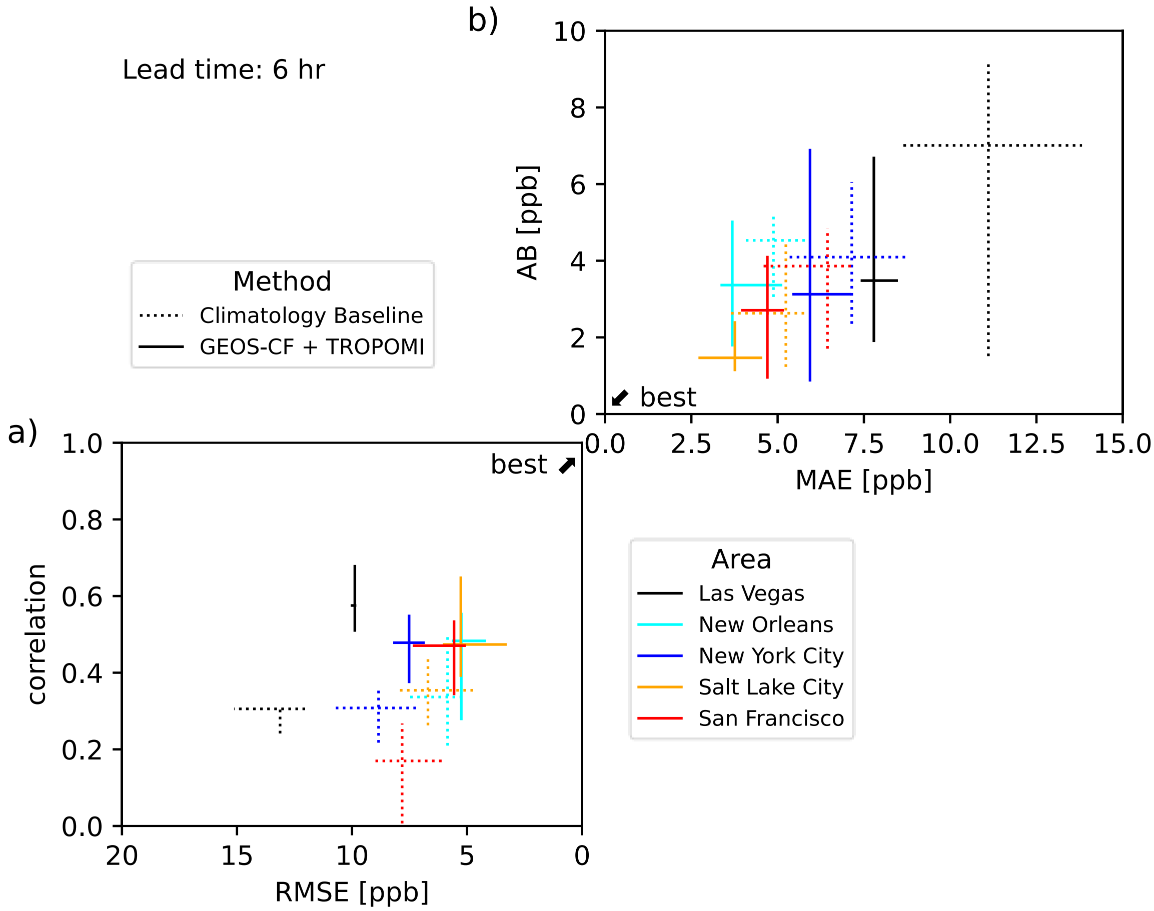

Figure 7. Comparison of the proposed method, in solid lines, to the climatology baseline, in dotted lines, in different color-coded application areas at a 6-hour forecast lead time. Performance is presented in terms of the correlation, RMSE (**a**), MAE, and AB (**b**) metrics. Performance is indicated with crosses, with the center of the cross indicating the median by each metric, and the arms of the cross denoting the 25th-to-75th percentile ranges of the metrics across the ground validation sites in each area. Axes are arranged such that the best performance by all metrics is towards the center of the figure overall.

For the climatology baseline of Equation 24, performance is roughly consistent across prediction lead times for all metrics. For this reason, Figure 7 compares the performance of this baseline with that of the proposed method for the 6-hour lead time only, as an illustrative example of the relative performance for all lead times. Note that the performance of the proposed method is typically consistent across lead times greater than about 3 houts, as indicated in Figure 6. In terms of median performance across all validation sites in each area, the proposed method universally improves over this baseline. The spread in performance in terms of the accuracy metrics of RMSE and MAE across ground validation sites, as indicated by the lengths of the horizontal bars of the crosses, is also typically smaller for the proposed method than for this baseline. This indicates that the proposed method has more consistent performance across different sites. Together, the results of these comparisons to the baselines illustrate the benefits of incorporating multiple data sources, as opposed to using the ground monitoring data only.

5.2 Impact of combining GEOS-CF with TROPOMI

The performance of the proposed method is also assessed both with and without the TROPOMI satellite information being included. Figure 8 breaks down this performance as a function of the local hour of the day, in order to examine the effects of these data in relation to the time at which the satellite passes occur. In terms of RMSE, there is typically little difference in performance due to the inclusion of these data. Satellite passes occur during a time of day when the performance is typically better by this metric anyway. A notable exception is Las Vegas, for which there are dramatic improvements in performance throughout the day due to the inclusion of the TROPOMI data, as shown by the solid versus dotted black lines in Figure 8. Satellite passes occur at a time of day when correlation is generally worse, although RMSE is better. In New York City there are notable improvements in the correlation for times of day around the satellite overpass time. This indicates that the TROPOMI information are having a temporally localized positive impact. Similar trends can be seen in other areas, but the effect is not universal. In San Francisco, there is a slight decrease in correlation for about 3 hours before and after the satellite overpass time, while overall the correlation was slightly improved by the addition of the TROPOMI data. Thus, while changes overall are fairly slight due to the inclusion of the TROPOMI satellite data in most areas, these can still have noticeable impacts to performance at specific times of the day, mostly around the satellite overpass times.

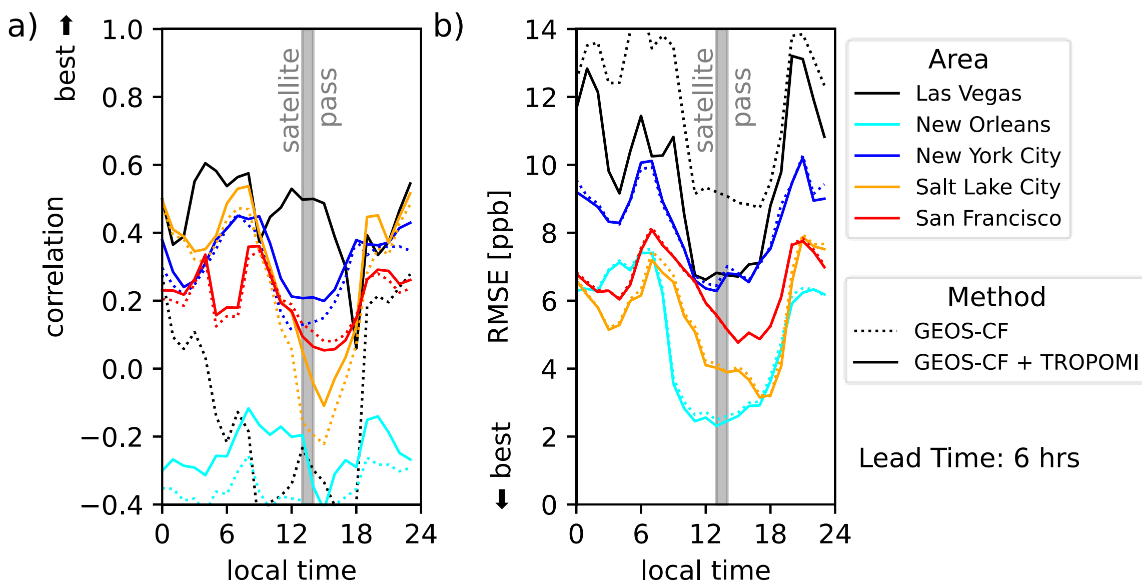

Figure 8. Performance of the proposed method applied with (solid lines) or without (dotted lines) the TROPOMI satellite information being included, evaluated in different color-coded areas as a function of the local time of day. Results are presented for the 6-hour lead time performance, as an illustrative example. Performance is presented in terms of the correlation (**a**) and RMSE (**b**) metrics. The times during which the TROPOMI instrument collects data, between 13:00 and 14:00 local time, are indicated with a gray band.

5.3 Forecast updating using residual correlations

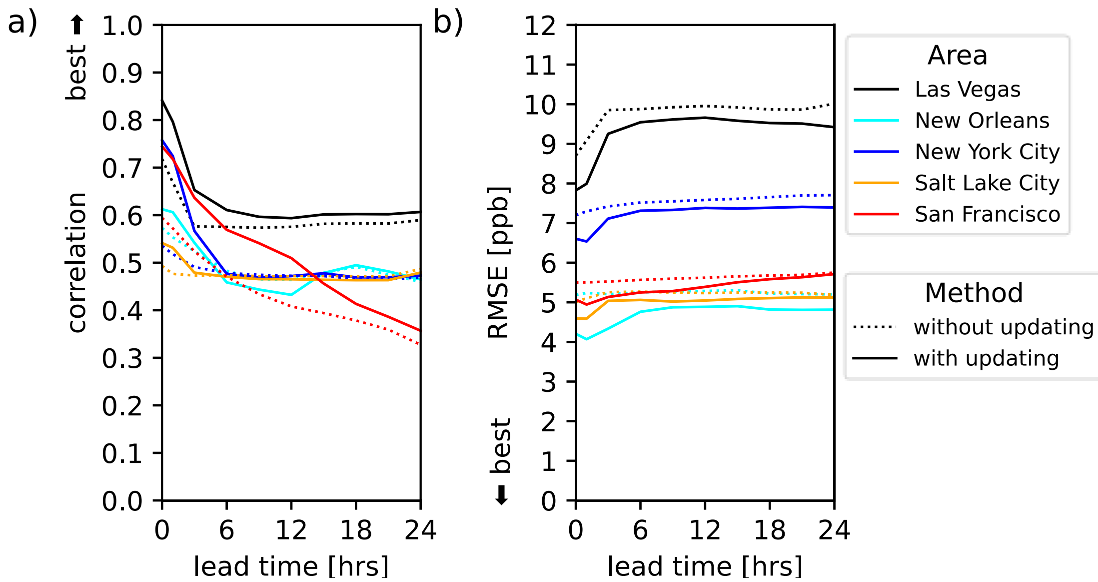

Figure 9. Performance of the proposed method applied with (solid lines) or without (dotted lines) the final updating step based on residual correlations, evaluated in different color-coded areas as a function of forecast lead time. Performance is presented in terms of the correlation (**a**) and RMSE (**b**) metrics.

Figure 9 presents the performance of applying the updating or kriging method based on residual correlations to the proposed approach, as outlined in Section 3.7. This updating leads to improved performance in all areas. As expected, these improvements are especially visible at shorter lead times, when temporal correlations are stronger. Some improvement is also noticeable even at longer lead times. Increases in correlation at short lead times are largest in New York City and San Francisco; this is likely due to the combination of the relatively large number of ground measurement sites in these areas and the relatively stronger observed spatial correlations in the forecast residuals. In Figure 10, performance is broken down by hour of the day. Results are presented for a 1-hour lead time, as an illustration of a case where temporal correlation with the latest ground measurements is generally high. While for the TROPOMI data, improvements in correlation are typically localized around the satellite overpass time, the continuous data provided by ground stations allow for positive impacts throughout the day. For New Orleans, where spatial heterogeneity of concentrations tends to be low, negative correlations throughout the day indicate that the model and satellite data do a poor job of capturing spatial patterns there. While overall correlations are positive in Figure 9a due to diurnal variability, the effect of this is removed when presenting the results as in Figure 10a. After updating with local data, these correlations, although still low, are positive, indicating a better representation of the spatial distribution of the pollutants is being made.

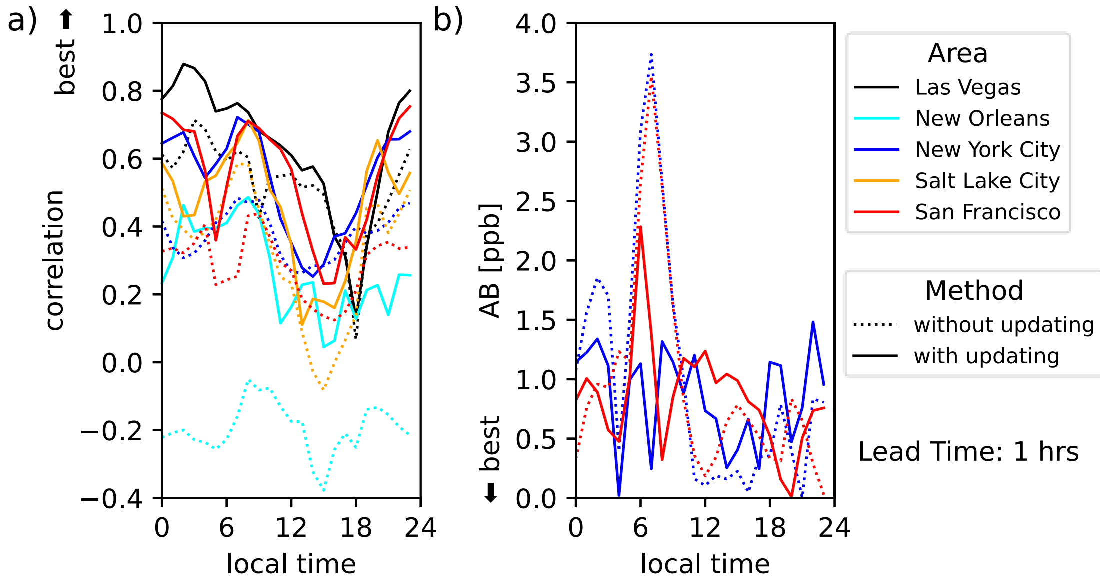

Figure 10. Performance of the proposed method applied with (solid lines) or without (dotted lines) the final updating step based on residual correlations, evaluated for different color-coded areas as a function of the local time of day. Results are presented for the 1-hour lead time performance. Performance is presented in terms of the correlation (**a**) and AB (**b**) metrics. Only results for New York City and San Francisco are presented in (**b**), for improved clarity.

In Figure 10b, there is a large peak in bias for the proposed method without updating in New York City and San Francisco centered around 7AM local time. This is likely due to a poor representation of the morning rush-hour patterns of pollutant distribution. Since the TROPOMI satellite passes occur after this rush-hour has passed, these data do not reflect the spatial patterns present during that period. Following updating with local data, this peak is substantially reduced, and in San Francisco shifted earlier in time by an amount equal to the forecast lead time. This reflects the impact of the near-real-time ground data. Forecasts for early rush-hour will still be biased as no measurements of that day's rush-hour concentrations are yet available. After these are collected and incorporated, subsequent forecasts for later during rush-hour will be more accurate. Overall, these results indicate the promise of examining residual spatial and temporal correlation patterns to make additional use of the latest ground measurement data in further improving the ability of these methods to capture local and transient pollution events which are only detectable in real-time by ground-based measurement.

6 Discussion

This paper has presented and demonstrated an approach for combining global atmospheric chemistry model outputs, specifically those from the GEOS-CF system, with satellite and ground-based measurements to generate high spatial and temporal resolution forecasts for surface-level air quality. Results for test cases of forecasting surface NO_2 across five US cities in September 2019 are presented and compared with baseline approaches which make use of surface monitoring data only. In all cases, except for very short-term forecasting, the proposed methods outperform both baselines by the metrics considered here.

Incorporating higher spatial resolution TROPOMI satellite information improves performance in most cases, with a substantial improvement observed in the Las Vegas area. This is not a universal result, however, and correlation in New York City actually declines slightly when TROPOMI information are included. However, examining performance by time of day, slight improvements are still observed even in New York City around the time of the satellite passes. When available, data from geostationary satellites for air quality monitoring missions, such as TEMPO for North America, GEMS for East Asia, and Sentinel 4 for Europe, should be considered. The use of such high temporal resolution information will overcome the limitation of using static typical satellite patterns, allowing for time-of-day-specific patterns instead. This could extend the benefits observed in certain areas around the TROPOMI satellite overpass time to the whole daytime when the geostationary instruments will make observations.

Attempts to include auxiliary information such as meteorological variables and VIIRS nighttime lights into the proposed approach as a proxy for human activity led to no notable improvements; see the supplemental information Section S2.2 for details. It is possible that different means of combining these data sources will provide different results. In particular, different temporal weightings of these sources might be used, since TROPOMI reflects day-time conditions while VIIRS may better represent night-time conditions. While in theory this might be achievable by combining patterns via regression as described in Section 3.3, in practice there may be insufficient calibration data to discern such relationships, or the relationships may be highly non-linear. Use of machine learning and/or Bayesian updating techniques to incorporate this information may prove more successful.

The final updating applied to surface concentration forecasts based on assumed residual correlations had a notable positive impact on performance by most metrics at short prediction lead times. These methods should be further investigated and expanded, using more sophisticated temporal correlation structures which take into account daily periodicity. This would allow, for example, in-situ information about yesterday’s rush-hour pollutant concentrations to play a larger role in updating today’s predictions. Incorporating the additional ground data sources available through low-cost sensor networks is also a promising area for future work. In that case, special consideration must be made for the relatively lower data quality of these sensors compared to the regulatory-grade instruments used here.

Finally, the techniques proposed in this work should be applicable to a variety of pollutants of interest for air quality applications. The same general techniques should still be applicable, although different relevant satellite retrievals and modeled pollutant species will have to be used. The relatively short atmospheric lifetime of NO_2 allowed column-integrated satellite retrievals to serve as a reasonable proxy for surface-level distribution patterns. For other pollutants, it may be beneficial to use ratios of surface-level to column-integrated pollutant concentrations, e.g., derived from GEOS-CF, to better relate satellite information to ground patterns.

Acknowledgements

Carl Malings is supported by an appointment to the NASA Postdoctoral Program at the Goddard Space Flight Center, administered by Universities Space Research Association (USRA) through a contract with NASA. K. Emma Knowland, Christoph Keller, and Stephen Cohn are supported through GMAO core funding administered by NASA’s Modeling, Analysis and Prediction (MAP) Program. Resources supporting GEOS-CF model simulations were provided by the NASA Center for Climate Simulation (NCCS) at the Goddard Space Flight Center (https://www.nccs.nasa.gov/services/discover).

Data Availability

GEOS-CF data are publicly available from the GMAO via the NCCS Data Portal (https://gmao.gsfc.nasa.gov/weather_prediction/GEOS-CF/data_access/). TROPOMI data are accessible via the Copernicus Open Access Data Hub (https://scihub.copernicus.eu/). EPA Network data are available from the EPA website (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw). All codes used for data analysis and prediction, as well as related outputs, are available at https://doi.org/10.5281/zenodo.4581090.

Author Contributions

Carl Malings: Conceptualization, Methodology, Software, Validation, Writing - Original Draft. K. Emma Knowland, Christoph Keller, and Stephen Cohn: Supervision, Writing - Review & Editing.

References

Ahangar, F., Freedman, F., & Venkatram, A. (2019). Using Low-Cost Air Quality Sensor Networks to Improve the Spatial and Temporal Resolution of Concentration Maps. *International Journal of Environmental Research and Public Health*, *16*(7), 1252. https://doi.org/10.3390/ijerph16071252

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., et al. (2001). Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, *106*(D19), 23073–23095. https://doi.org/10.1029/2001JD000807

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., et al. (2012). Exposure Assessment for Estimation of the Global Burden of Disease Attributable to Outdoor Air Pollution. *Environmental Science & Technology*, *46*(2), 652–660. https://doi.org/10.1021/es2025752

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V., Dentener, F., et al. (2016). Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environmental Science & Technology*, *50*(1), 79–88. https://doi.org/10.1021/acs.est.5b03709

Chang, K.-L., Cooper, O. R., West, J. J., Serre, M. L., Schultz, M. G., Lin, M., et al. (2019). A new method (M3Fusion v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution. *Geoscientific Model Development*, *12*(3), 955–978. https://doi.org/10.5194/gmd-12-955-2019

Chow, J. C. (1995). Measurement Methods to Determine Compliance with Ambient Air Quality Standards for Suspended Particles. *Journal of the Air & Waste Management Association*, *45*(5), 320–382. https://doi.org/10.1080/10473289.1995.10467369

Cleland, S. E., West, J. J., Jia, Y., Reid, S., Raffuse, S., O'Neill, S., & Serre, M. L. (2020). Estimating Wildfire Smoke Concentrations during the October 2017 California Fires through BME Space/Time Data Fusion of Observed, Modeled, and Satellite-Derived $PM_{2.5}$. *Environmental Science & Technology*, *54*(21), 13439–13447. https://doi.org/10.1021/acs.est.0c03761

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air

pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6
Cooper, M. J., Martin, R. V., McLinden, C. A., & Brook, J. R. (2020). Inferring ground-level nitrogen dioxide concentrations at fine spatial resolution applied to the TROPOMI satellite instrument. *Environmental Research Letters*, *15*(10), 104013. https://doi.org/10.1088/1748-9326/aba3a5
Cressie, N. A. C., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken, N.J: Wiley.
Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., et al. (2020). Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. *Remote Sensing*, *12*(6), 914. https://doi.org/10.3390/rs12060914
van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application. *Environmental Health Perspectives*, *118*(6), 847–855. https://doi.org/10.1289/ehp.0901623
Duncan, B. N., Prados, A. I., Lamsal, L. N., Liu, Y., Streets, D. G., Gupta, P., et al. (2014). Satellite data of atmospheric pollution for U.S. air quality applications: Examples of applications, summary of data end-user resources, answers to FAQs, and common mistakes to avoid. *Atmospheric Environment*, *94*, 647–662. https://doi.org/10.1016/j.atmosenv.2014.05.061
Eastham, S. D., Weisenstein, D. K., & Barrett, S. R. H. (2014). Development and evaluation of the unified tropospheric–stratospheric chemistry extension (UCX) for the global chemistry-transport model GEOS-Chem. *Atmospheric Environment*, *89*, 52–63. https://doi.org/10.1016/j.atmosenv.2014.02.001
Engel-Cox, J. A., Holloman, C. H., Coutant, B. W., & Hoff, R. M. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, *38*(16), 2495–2509. https://doi.org/10.1016/j.atmosenv.2004.01.039
Forouzanfar, M. H., Alexander, L., Anderson, H. R., Bachman, V. F., Biryukov, S., Brauer, M., et al. (2015). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *386*(10010), 2287–2323. https://doi.org/10.1016/S0140-6736(15)00128-2
Friberg, M. D., Zhai, X., Holmes, H. A., Chang, H. H., Strickland, M. J., Sarnat, S. E., et al. (2016). Method for Fusing Observational Data and Chemical Transport Model Simulations To Estimate Spatiotemporally Resolved Ambient Air Pollution. *Environmental Science & Technology*, *50*(7), 3695–3705. https://doi.org/10.1021/acs.est.5b05134
Goldberg, D. L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z., & Streets, D. G. (2019). Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM2.5 concentrations at 1 km resolution in the Eastern United States. *Atmospheric Environment*, *199*, 443–452. https://doi.org/10.1016/j.atmosenv.2018.11.049
Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., et al. (2018). Impact of California Fires on Local and Regional Air Quality: The Role of a Low-Cost

Sensor Network and Satellite Observations. *GeoHealth*, *2*(6), 172–181. https://doi.org/10.1029/2018GH000136

Han, W., Tong, L., Chen, Y., Li, R., Yan, B., & Liu, X. (2018). Estimation of High-Resolution Daily Ground-Level PM2.5 Concentration in Beijing 2013–2017 Using 1 km MAIAC AOT Data. *Applied Sciences*, *8*(12), 2624. https://doi.org/10.3390/app8122624

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057

Hu, L., Keller, C. A., Long, M. S., Sherwen, T., Auer, B., Da Silva, A., et al. (2018). Global simulation of tropospheric chemistry at 12.5 km resolution: performance and evaluation of the GEOS-Chem chemical module (v10-1) within the NASA GEOS Earth system model (GEOS-5 ESM). *Geoscientific Model Development*, *11*(11), 4603–4620. https://doi.org/10.5194/gmd-11-4603-2018

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., et al. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, *15*(2), 185–204. https://doi.org/10.1038/sj.jea.7500388

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A., Krewski, D., Newbold, K. B., et al. (2005). Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology*, *16*(6), 727–736. https://doi.org/10.1097/01.ede.0000181630.15826.7d

Jiang, T., Chen, B., Nie, Z., Ren, Z., Xu, B., & Tang, S. (2021). Estimation of hourly full-coverage PM2.5 concentrations at 1-km resolution in China using a two-stage random forest model. *Atmospheric Research*, *248*, 105146. https://doi.org/10.1016/j.atmosres.2020.105146

Just, A. C., Arfer, K. B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., & Kloog, I. (2020). Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. *Atmospheric Environment*, *239*, 117649. https://doi.org/10.1016/j.atmosenv.2020.117649

Karner, A. A., Eisinger, D. S., & Niemeier, D. A. (2010). Near-Roadway Air Quality: Synthesizing the Findings from Real-World Data. *Environmental Science & Technology*, *44*(14), 5334–5344. https://doi.org/10.1021/es100008x

Keller, C. A., Long, M. S., Yantosca, R. M., Da Silva, A. M., Pawson, S., & Jacob, D. J. (2014). HEMCO v1.0: a versatile, ESMF-compliant component for calculating emissions in atmospheric models. *Geoscientific Model Development*, *7*(4), 1409–1417. https://doi.org/10.5194/gmd-7-1409-2014

Keller, Christoph A., Evans, M. J., Knowland, K. E., Hasenkopf, C. A., Modekurty, S., Lucchesi, R. A., et al. (2020). *Global Impact of COVID-19 Restrictions on the Surface Concentrations of Nitrogen Dioxide and Ozone* (preprint). Gases/Atmospheric Modelling/Troposphere/Chemistry (chemical composition and reactions). https://doi.org/10.5194/acp-2020-685

Knowland, K. E., Keller, C. A., & Lucchesi, R. (2020). *File Specification for GEOS-CF Products* (No. Office Note No. 17 (Version 1.1)). Goddard Space Flight Center, Greenbelt, Maryland, USA: GMAO. Retrieved from https://gmao.gsfc.nasa.gov/pubs/docs/Knowland1204.pdf

Lamsal, L. N., Martin, R. V., van Donkelaar, A., Steinbacher, M., Celarier, E. A., Bucsela, E., et al. (2008). Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. *Journal of Geophysical Research*, *113*(D16), D16308. https://doi.org/10.1029/2007JD009235

Lamsal, L. N., Martin, R. V., van Donkelaar, A., Celarier, E. A., Bucsela, E. J., Boersma, K. F., et al. (2010). Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: Insight into the seasonal variation of nitrogen oxides at northern midlatitudes. *Journal of Geophysical Research*, *115*(D5), D05302. https://doi.org/10.1029/2009JD013351

Liu, L.-J. S., Tsai, M.-Y., Keidel, D., Gemperli, A., Ineichen, A., Hazenkamp-von Arx, M., et al. (2012). Long-term exposure models for traffic related NO2 across geographically diverse areas over separate years. *Atmospheric Environment*, *46*, 460–471. https://doi.org/10.1016/j.atmosenv.2011.09.021

Loh, M., Sarigiannis, D., Gotti, A., Karakitsios, S., Pronk, A., Kuijpers, E., et al. (2017). How Sensors Might Help Define the External Exposome. *International Journal of Environmental Research and Public Health*, *14*(4), 434. https://doi.org/10.3390/ijerph14040434

Long, M. S., Yantosca, R., Nielsen, J. E., Keller, C. A., da Silva, A., Sulprizio, M. P., et al. (2015). Development of a grid-independent GEOS-Chem chemical transport model (v9-02) as an atmospheric chemistry module for Earth system models. *Geoscientific Model Development*, *8*(3), 595–602. https://doi.org/10.5194/gmd-8-595-2015

Lyapustin, A., & Wang, Y. (2018). *MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006 [Data set]*. NASA EOSDIS Land Processes DAAC. Retrieved from https://doi.org/10.5067/MODIS/MCD19A2.006

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S. P. N., Zimmerman, N., Kara, L. B., et al. (2019). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, *12*(2), 903–920. https://doi.org/10.5194/amt-12-903-2019

Malings, C., Westervelt, D. M., Hauryliuk, A., Presto, A. A., Grieshop, A., Bittner, A., et al. (2020). Application of low-cost fine particulate mass monitors to convert satellite aerosol optical depth to surface concentrations in North America and Africa. *Atmospheric Measurement Techniques*, *13*(7), 3873–3892. https://doi.org/10.5194/amt-13-3873-2020

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, *42*(6), 1359–1369. https://doi.org/10.1016/j.atmosenv.2007.08.012

Mhawish, A., Banerjee, T., Sorek-Hamer, M., Bilal, M., Lyapustin, A. I., Chatfield, R., & Broday, D. M. (2020). Estimation of High-Resolution PM2.5 over the Indo-Gangetic Plain by Fusion of Satellite Data, Meteorology, and Land Use Variables. *Environmental Science & Technology*, *54*(13), 7891–7900. https://doi.org/10.1021/acs.est.0c01769

Michaelides, S., Paronis, D., Retalis, A., & Tymvios, F. (2017). Monitoring and Forecasting Air Pollution Levels by Exploiting Satellite, Ground-Based, and Synoptic Data, Elaborated with Regression Models. *Advances in Meteorology*, *2017*, 1–17. https://doi.org/10.1155/2017/2954010

Murray, N. L., Holmes, H. A., Liu, Y., & Chang, H. H. (2019). A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation. *Environmental Research*, *178*, 108601. https://doi.org/10.1016/j.envres.2019.108601

NASA VIIRS Land Science Investigator-Led Processing System. (2019). VIIRS/NPP Daily Gridded Day Night Band 500m Linear Lat Lon Grid Night [Data set]. NASA Level 1 and Atmosphere Archive and Distribution System. https://doi.org/10.5067/VIIRS/VNP46A1.001

Orbe, C., Oman, L. D., Strahan, S. E., Waugh, D. W., Pawson, S., Takacs, L. L., & Molod, A. M. (2017). Large-Scale Atmospheric Transport in GEOS Replay Simulations: TRANSPORT IN GEOS REPLAY SIMULATIONS. *Journal of Advances in Modeling Earth Systems*, *9*(7), 2545–2560. https://doi.org/10.1002/2017MS001053

Popoola, O. A. M., Stewart, G. B., Mead, M. I., & Jones, R. L. (2016). Development of a baseline-temperature correction methodology for electrochemical sensors and its implications for long-term stability. *Atmospheric Environment*, *147*, 330–343. https://doi.org/10.1016/j.atmosenv.2016.10.024

Román, M. O., Wang, Z., Sun, Q., Kalb, V., Miller, S. D., Molthan, A., et al. (2018). NASA's Black Marble nighttime lights product suite. *Remote Sensing of Environment*, *210*, 113–143. https://doi.org/10.1016/j.rse.2018.03.017

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., & Bartonova, A. (2017). Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International*, *106*, 234–247. https://doi.org/10.1016/j.envint.2017.05.005

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., et al. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(1), 231–253. https://doi.org/10.1111/rssc.12227

Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope*, *68*(2), 159.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., et al. (2013). The Changing Paradigm of Air Pollution Monitoring. *Environmental Science & Technology*, *47*(20), 11369–11377. https://doi.org/10.1021/es4022602

Tan, Y., Lipsky, E. M., Saleh, R., Robinson, A. L., & Presto, A. A. (2014). Characterizing the Spatial Variation of Air Pollutants and the Contributions of High Emitting Vehicles in Pittsburgh, PA. *Environmental Science & Technology*, *48*(24), 14186–14194. https://doi.org/10.1021/es5034074

TROPOMI Level 2 Nitrogen Dioxide. (n.d.). [Data set]. European Space Agency. https://doi.org/10.5270/S5P-s4ljg54

Turner, M. C., Nieuwenhuijsen, M., Anderson, K., Balshaw, D., Cui, Y., Dunton, G., et al. (2017). Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations. *Annual Review of Public Health*, *38*(1), 215–239. https://doi.org/10.1146/annurev-publhealth-082516-012802

US EPA. (2017). *Policy Assessment for the Review of the Primary National Ambient Air Quality Standards forOxides of Nitrogen* (No. EPA-452/R-17-003). Office of Air Quality Planning and Standards, Health and Environmental Impacts Division, Research Triangle Park, NC, USA: U.S. Environmental Protection Agency.

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., et al. (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, *120*, 70–83. https://doi.org/10.1016/j.rse.2011.09.027

Williams, R., Vasu Kilaru, Snyder, E., Kaufman, A., Dye, T., Rutter, A., et al. (2014). *Air Sensor Guidebook* (No. EPA/600/R-14/159 (NTIS PB2015-100610)). Washington, DC: United States Environmental Protection Agency. Retrieved from https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=519616

Zhang, H., Wang, J., García, L. C., Ge, C., Plessel, T., Szykman, J., et al. (2020). Improving Surface $PM_{2.5}$ Forecasts in the United States Using an Ensemble of Chemical Transport Model Outputs: 1. Bias Correction With Surface Observations in Nonrural Areas. *Journal of Geophysical Research: Atmospheres*, *125*(14). https://doi.org/10.1029/2019JD032293