# Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results

Carl Malings[1], K. Emma Knowland[1], Nathan Pavlovic[2], Justin G. Coughlin[2], Christoph A. Keller[3], Stephen E. Cohn[4], and Randall V Martin[5]

[1]Morgan State University
[2]Sonoma Technology, Inc.
[3]Universities Space Research Association
[4]Data Assimilation Office
[5]Washington University in St. Louis

March 15, 2024

**Abstract**

Integrating air quality information from models, satellites, and in-situ monitors allows for both better estimation of air quality and better quantification of uncertainties in this estimation. Uncertainty quantification is important to appropriately convey confidence in these estimates and forecasts to users who will base decisions on these. Uncertainty quantification also allows tracing the value of information provided by different data sources. This can identify gaps in the monitoring network where additional data could further reduce uncertainties. This paper presents a framework for data fusion with uncertainty quantification, applicable to multiple air-quality-relevant pollutants. Testing of this framework in the context of nitrogen dioxide forecasting at sub-city scales shows promising results, with confidence intervals typically encompassing the expected number of actual measurements during cross-validation. The framework is now being implemented into an online tool to support local air quality management decision-making. Future work will also include the incorporation of low-cost air sensor data and the quantification of uncertainty at hyper-local scales.

# Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results

**Carl Malings[1,2][0000-0002-2242-4328], K. Emma Knowland[1,2][0000-0003-0837-8502], Nathan Pavlovic[3][0000-0003-2127-3940], Justin G. Coughlin[3][0000-0003-3882-3064], Christoph Keller[1,2][0000-0002-0552-4298], Stephen Cohn[2][0000-0001-8506-9354], and Randall V. Martin[4][0000-0003-2632-8402]**

[1]Morgan State University, GESTAR II Cooperative Agreement, Baltimore, MD 21251, USA.

[2]NASA Goddard Space Flight Center, Global Modeling & Assimilation Office, Greenbelt, MD 20771, USA.

[3]Sonoma Technology, Inc., Petaluma, CA 94954, USA.

[4]Washington University in St. Louis., St. Louis, MO 63130, USA.

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

**Key Points:**

- The proposed data fusion method produces a-priori uncertainty assessments and confidence intervals for estimates and forecasts
- Confidence intervals were found to be mostly reasonable in a test case study for nitrogen dioxide across four months and two cities
- The method provided overconfident estimates for sites within 100 meters of highways

## Abstract

Integrating air quality information from models, satellites, and in-situ monitors allows for both better estimation of air quality and better quantification of uncertainties in this estimation. Uncertainty quantification is important to appropriately convey confidence in these estimates and forecasts to users who will base decisions on these. Uncertainty quantification also allows tracing the value of information provided by different data sources. This can identify gaps in the monitoring network where additional data could further reduce uncertainties. This paper presents a framework for data fusion with uncertainty quantification, applicable to multiple air-quality-relevant pollutants. Testing of this framework in the context of nitrogen dioxide forecasting at sub-city scales shows promising results, with confidence intervals typically encompassing the expected number of actual measurements during cross-validation. The framework is now being implemented into an online tool to support local air quality management decision-making. Future work will also include the incorporation of low-cost air sensor data and the quantification of uncertainty at hyper-local scales.



## Plain Language Summary

Poor air quality has adverse impacts on human and environmental health. Estimating and forecasting air quality accurately can improve early warnings and mitigation for poor air quality. Furthermore, understanding the uncertainties and degree of confidence in these forecasts and estimates can help air quality managers know when and where they can be relied upon, and where more data might still be needed. This paper outlines a method to combine air quality information from models, satellites, and ground-based monitors, and to assess the confidence in the combined output. Combining all these data sources can give us a better overall understanding of air quality, and making comparisons between them allows us to better understand uncertainties. Testing out the method proposed in this paper, we find that the method can produce reasonable assessments of the confidence it has in its estimates, with the expected numbers of actual measurements usually falling within the confidence intervals produced by the method. An exception is when this method is applied very close to a major pollution source (e.g., a highway, in our study). In such cases, since the method does not know that there is such a source nearby, it tends to be overconfident in its prediction.


## 1 Introduction

Poor air quality is a major global public health concern. The 2019 Global Burden of Disease study identified air pollution as the leading environmental risk factor for human premature mortality (Murray et al., 2020). To mitigate this public health problem on a global scale, air quality managers and practitioners first need access to accurate and comprehensive information on the state of air quality in their areas. Such information might come from a variety of disparate sources. In-situ measurements of air quality, typically obtained from instruments operated by regulatory bodies, e.g., the Environmental Protection Agency in the United States, are considered the trusted standard for assessing air quality. At a global scale, however, the relatively low density of such measurements means that regulatory instruments alone often cannot provide necessary air quality information to answer basic questions relevant to public health (Martin et al., 2019). Low-cost air quality sensors (LCS) are increasing in prominence to address this in-situ data gap (e.g., Tanzer et al., 2019; Rose Eilenberg et al., 2020). As the name implies, these provide a less expensive alternative to traditional regulatory-grade air quality monitors (RGM). As a tradeoff to achieve this

lower cost, LCS suffer from greater measurement uncertainties, and thus, require extensive calibration and validation efforts to generate useable data (Giordano et al., 2021). LCS can also be deployed to new areas which do not have the infrastructure to support RGM. LCS provide the only currently feasible means of routine air quality assessment in many low-and-middle-income countries (Hodoli et al., 2023; McFarlane, Isevulambire, et al., 2021; Raheja et al., 2022).

Even so, the availability of local air quality data from in-situ RGM or LCS may not provide sufficient situational awareness to air quality managers. Other, more globally available data sources may be required. One important source of such global data is satellite remote sensing retrievals of atmospheric composition. These data are provided by a fleet of instruments operated by national aerospace agencies and the private sector. By providing in many cases globe-spanning monitoring of the chemical and physical properties of the atmosphere at increasingly fine spatial resolution, satellite data can fill many gaps in our understanding of the composition of the atmosphere. However, satellite remote sensing has some key limitations with respect to air quality applications. Typically, remote sensing estimates take account of the entire atmospheric column, rather than the surface-level concentrations which are most relevant to air quality and the associated health exposure risk. The relationship between surface and column quantities is dependent on many factors. Thus, while promising, certain expertise and domain knowledge is required to correctly interpret satellite data for air quality purposes, which may be a barrier to its routine use in many areas (Anenberg et al., 2020; Duncan et al., 2021; Holloway et al., 2021).

Other sources of global air quality information are atmospheric chemistry and transport models (CTM). These models seek to estimate the state of the atmosphere, including parameters relevant for air quality, based on mathematical representations of chemical and physical processes combined with input data related to boundary conditions, e.g., the estimated emissions of various pollutants into the atmosphere. These models produce spatially comprehensive datasets and have the potential to forecast future air quality. However, their estimates may be biased due to incomplete and/or outdated input information or by inadequate representation of some chemical or physical processes. For example, inadequate temporal resolution for emissions data, differing vertical representations between the model and observations, as well as boundary layer mixing were found to impact the ability of the GEOS-Chem model to represent diel variations in fine particulate matter ($PM_{2.5}$) over the United States (Y. Li et al., 2023). Constraining CTM with observations from satellites, RGM, LCS, or a combination thereof via data assimilation is a widely used approach to addressing these model shortcomings. Assimilation of satellite data is more typical for global-scale CTM (Bocquet et al., 2015; Kelp et al., 2023), while in-situ data assimilation is more typical for sub-city to national scale CTM (Lopez-Restrepo et al., 2021; Schneider et al., 2023; Hassani et al., 2023).

Data fusion is an approach for bringing together various data sources. In contrast to data assimilation, where observations are used to update the state of a model, data fusion combines multiple data sources to produce a new data product, distinct from the inputs. A typical niche filled by data fusion is "downscaling" of coarser-resolution regional or global CTM output to produce more locally applicable outputs (Diao et al., 2019). A myriad of approaches using different inputs and methodologies has been proposed. On a local scale, data fusion of a dispersion model and LCS data has supported hourly $PM_{10}$ mapping in Nantes, France (Gressent et al., 2020). Regionally, satellite information is commonly used to support data fusion approaches; fusion of satellite aerosol optical depth (AOD), land use information, and meteorological data with surface observations from RGM and LCS allowed for daily 1-km resolution estimation of $PM_{2.5}$ over

California, USA (Bi et al., 2020). Satellite AOD, RGM, and LCS data were similarly combined for $PM_{2.5}$ mapping over Taiwan (J. Li et al., 2020). Globally, data fusion approaches are used to create yearly, monthly, or daily average surface $PM_{2.5}$ and constituent estimates (van Donkelaar et al., 2015, 2021; Wei et al., 2023). These estimates support analysis of the global impacts of air quality (Murray et al., 2020). For forecasting applications, i.e., prediction of surface concentrations in advance, bias correction for an ensemble of CTM was performed using surface RGM observations in both urban and rural areas to improve hourly $PM_{2.5}$ forecasting over the USA (Zhang et al., 2020, 2022). CTM, satellite and RGM data are combined to improve hourly $NO_2$ forecasts at sub-city scale (Malings et al., 2021). Machine learning methods have also been used for bias-correction of global CTM to produce daily $PM_{2.5}$ forecasts at 1-km resolution for applications at sub-city scale (Keller et al., 2020; Duncan et al., 2021; Bi et al., 2022). These studies demonstrate the wide applicability and flexibility of data fusion to incorporate models with various observational datasets.

In contrast to deterministic methods, probabilistic estimates and forecasts for air quality may be better suited to the needs of air quality managers and policy makers. For example, in a decision-focused analysis of ozone forecasting based on public health protection, it was found that single deterministic forecasts may produce less robust results compared to the use of multiple forecasts or an ensemble of forecasts for guiding air quality decision-making (Balashov et al., 2017; Garner & Thompson, 2012). This was because the ensemble forecasts more readily allowed for choosing actions which would be robust under a range of outcomes, i.e., robust under uncertainty. For global data fusion estimates of monthly $PM_{2.5}$, uncertainty quantification also supports analyzing the impact of this uncertainty on global health and epidemiological assessments (van Donkelaar et al., 2021). Several recent efforts have aimed at the quantification of uncertainty in air quality estimation and forecasting. Most of these approaches make use of ensembles of deterministic models (Garaud & Mallet, 2011; Gilliam et al., 2015; Riccio & Chianese, 2024) or machine learning methods, e.g., using generative models to produce a simulated ensemble (Fanfarillo et al., 2019). Data fusion approaches making use of geostatistical methods, especially Gaussian process or kriging approaches, have inherent capabilities to constrain estimates and quantify uncertainties for air quality estimation and forecasting (Wang et al., 2021). Kriging is referred to as "objective analysis" or "optimum interpolation" in the early numerical weather prediction literature (Diggle, 2010, p. 8). A major barrier to the wider use of probabilistic forecasts in air quality applications has been the difficulty associated with the interpretation of probabilistic forecasts by decision-makers and effectively communicating these to the public. Recent work has aimed at addressing these issues by explicitly analyzing different interpretation strategies corresponding with different desired outcomes (Balashov et al., 2023).

This paper presents a framework for combining CTM output, satellite remote sensing data, and in-situ measurements from a combination of RGM and LCS via a data fusion approach to support air quality estimation and/or forecasting. This framework includes explicit quantification of uncertainties associated with outputs from each stage, i.e., as each additional dataset is added. This paper aims at presenting a simple, generalizable method for data fusion with uncertainty quantification which can be implemented for near-real-time applications, with more limited computational requirements than a full data assimilation approach. We demonstrate this framework with a case study, focusing on estimation and forecasting of nitrogen dioxide in two US cities (San Francisco and New York City) in 2019. Nitrogen dioxide ($NO_2$), a regulated pollutant in the US (US EPA, 2017), represents a useful test case since it is known to vary on fine spatial scales in urban areas, which may not be captured even in high-resolution satellite datasets (e.g., Judd et al.,

2019). The ability to characterize this variability is an informative illustration of the capabilities of the proposed framework. The development of analysis tools and data products which combine multiple sources of air quality information, alongside methods to express confidence in or quantification of uncertainties in these products, has been suggested as a key need of air quality managers worldwide (Duncan et al., 2021). The methods presented in this paper are being implemented as part of a NASA-funded project to develop such tools for air quality data managers.

## 2 Methods

### 2.1 Input datasets

The proposed data fusion approach makes use of three categories of input information: CTM-based estimates and forecasts, satellite remote sensing data, and ground monitor data.

The NASA Global Earth Observing System Composition Forecast (GEOS-CF) system generates CTM outputs used in this paper. GEOS-CF couples the GEOS atmospheric general circulation model with the GEOS-Chem chemistry module (Keller et al., 2021). GEOS-CF produces 5-day forecasts initialized every day, following a 24-hour historical simulation for the previous day with the meteorology constrained by assimilated fields, to provide the best estimates for the past atmospheric composition. Both forecast and historical model output are used here. Hourly-average "surface-level" (average for the GEOS model's lowest level, nominally 130 m thick) nitrogen dioxide concentrations along with tropospheric column concentrations are used for the year 2019. GEOS-CF outputs are on a 0.25° or roughly 25 km latitude-longitude grid.

The TROPOMI instrument on the Sentinel 5P satellite provides retrievals related to tropospheric column concentrations of $NO_2$ (Veefkind et al., 2012). Through an agreement with the European Space Agency, TROPOMI data are also hosted at the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), searchable via the Common Metadata Repository system; these systems were used to identify and access relevant TROPOMI datasets. Tropospheric $NO_2$ concentration data products are used here, with recommended data quality filters for "good quality" retrievals. The latest high-resolution data product with a nominal pixel size of 5.5 by 3.5 km is used.

This paper presents a case study focused on San Francisco, California, USA (defined as between 37° N and 39° N and between 121° W and 123° W). Data for the month of September 2019 were used for the primary analysis; additional data from calendar year 2019 were also included as potential inputs for calibration purposes and for additional analysis presented in Section 3.3. An additional case study focused on New York City, New York, USA is also presented in the supplemental materials, described in supplemental text S1. These locations were selected due to their relatively high density of RGM for $NO_2$, as well as for comparability with previous related work (Malings et al., 2021). Ground monitoring data for hourly $NO_2$ were obtained from the US EPA's RGM network. Relevant data were queried using the Air Quality System API.

### 2.2 Data fusion approach and uncertainty quantification

The method for air quality data fusion outlined here is adapted from prior work (Malings et al., 2021). The major improvements presented here include (1) a generalization of the methodology and notation, where relevant changes to corresponding elements of the prior work will be noted, and (2) development of a framework for quantifying the uncertainty in fused estimates of surface air quality, which was not present in the prior work. The method is separated

into four phases: phase 1 involves model-based historical estimates and forecasts only; phase 2 fuses satellite with model data; phase 3 integrates in-situ measurements in an "offline" manner, useful mainly for bias correction; phase 4 integrates in-situ measurements in an "online" manner, useful for near-term estimate and forecast updating.

### *2.2.1 Phase 1: model-based estimation and uncertainty*

This data fusion approach starts with air quality estimate and forecast model outputs. Let $M(x,t,\tau)$ denote the estimated surface concentration of a given pollutant applicable at location $x$ and time $t$ produced by an air quality model (the GEOS-CF model in the current work). The forecasting lead-time is denoted by $\tau$. If target time $t$ is in the future, lead-time $\tau$ will be the difference between $t$ and when the model forecast was initialized. If $t$ is in the past, then $\tau = 0$, and the latest available model output covering time $t$ is used. Lead-time $\tau$ may not always be explicitly noted for notational convenience; when it is omitted, assume $\tau = 0$. The phase 1 estimate is simply the relevant model output:

$$E_1(x,t,\tau) = M(x,t,\tau). \quad (1)$$

Practically, it is important to note that while $x$ represents a location on the Earth's surface to arbitrary precision, the spatial resolution on which $E_1$ will be defined is limited to the spatial resolution of the model. In future work, it is considered that an ensemble of air quality models, either from different modeling systems or multiple initializations of the same model system, may be used to inform the data fusion. In that case, $E_1(x,t,\tau)$ could be the mean of multiple available models. Furthermore, the ensemble spread could be used for uncertainty quantification.

To better inform end-users on the uncertainty in data fusion estimates, we also aim to quantify the uncertainty of $E_1(x,t,\tau)$ in terms of the expected mean square error of the estimate with respect to the true concentration. We denote this uncertainty as $V_1(x,t,\tau)$. We estimate this uncertainty as the sum of four components, where independence between the components is assumed. These components are the uncertainty in the forecast due solely to its lead-time, $V_{F1}(x,t,\tau)$, the uncertainty due to local variability in the air quality model output, $V_M(x,t)$, the uncertainty due to potential bias in the air quality model, $V_{B1}(x,t)$, and the uncertainty due to the representational error of the model, $V_{R1}(x,t)$, due to its relatively coarse spatial resolution. Thus:

$$V_1(x,t,\tau) = V_{F1}(x,t,\tau) + V_M(x,t) + V_{B1}(x,t) + V_{R1}(x,t). \quad (2)$$

Model-based uncertainties $V_{F1}(x,t,\tau)$ and $V_M(x,t)$ are estimated empirically using model outputs. $V_{F1}(x,t,\tau)$ is estimated using the mean square difference of past model forecasts at lead-time $\tau$ and estimates at lead-time 0 for location $x$. This is evaluated over a set of times denoted $T_{c,t.o.d.}(t)$, representing times during a calibration period in the recent past, e.g., the prior week, at the same time-of-day (t.o.d.) as the time of interest $t$. This is meant to account for potential systematic differences in forecasting capabilities at different times of the day due to diel cycles or initialization times.

$$V_{F1}(x,t,\tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)}\left[\left(M(x,t',\tau) - M(x,t',0)\right)^2\right], \quad (3)$$

where $\mathbb{E}_i[\cdot]$ denotes the expected value, i.e., the mean, of the expression in brackets with respect to indexing parameter $i$. Note that $V_{F1}(x,t,0) = 0$ by design, and so this term can be ignored for $\tau = 0$.

$V_M(x,t)$ is estimated as the expected square difference of model outputs in the immediate vicinity of location $x$ and time $t$, i.e., the mean square difference of the model outputs in the grid cells immediately surrounding it in space and time:

$$V_M(x,t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[ \left( M(x',t') - M(x,t) \right)^2 \right], \quad (4)$$

where $X_n(x)$ represents the neighborhood of location $x$, i.e., its adjoining model grid cells depending on the model spatial resolution, and $T_n(t)$ represents the neighborhood of time $t$, i.e., the preceding and subsequent time steps according to the model temporal resolution. The logic behind this estimate is that, where model outputs are "smooth" in space and time, there is less uncertainty in the model outputs, while when the model outputs are more variable in space and time, there is greater uncertainty. This estimate depends on the model resolution, with lower uncertainties estimated for finer resolutions, all else being equal. We consider this to be reasonable, as finer resolution models will tend to explicitly represent processes at the relevant scale. However, simply interpolating model outputs to a finer resolution would artificially reduce the uncertainty estimate. This analysis should therefore be conducted at the native resolution of the model. A schematic for this phase is provided in Supplemental Figure S1.

The remaining terms $V_{B1}(x,t)$ and $V_{R1}(x,t)$ are impossible to assess using the model alone and must be estimated using external information, as will be discussed later (see Section 2.2.5). Note that, if an ensemble of models is used, it may be possible to estimate $V_{B1}(x,t)$ using the mean square differences between models in the ensemble (Riccio & Chianese, 2024). However, it may still be the case that all models within an ensemble are systematically biased due to some common underlying factor, e.g., all models using the same emissions dataset.

### *2.2.2 Phase 2: model downscaling with satellite data*

In phase 2, relationships between column concentrations from model and satellite data are used to inform the sub-model-grid variability of the pollutant of interest. The phase 2 estimate of the concentration of this pollutant at time $t$ and location $x$, $E_2(x,t,\tau)$, is the phase 1 estimate modified by the satellite-informed sub-grid difference pattern $D(x,t)$:

$$E_2(x,t,\tau) = E_1(x,t,\tau) + D(x,t), \quad (5)$$

where:

$$D(x,t) = \mathbb{E}_{t' \in T_{c,overpass}(t)} \left[ \left( S_{col}(x,t') - E_{1,col}(x,t') \right) \phi(x,t')\, \psi(x,t,t') \right]. \quad (6)$$

This difference pattern is the mean of the difference between the satellite-retrieved column concentration of the pollutant of interest, $S_{col}$, and the estimate of the same column quantity by the model used in phase 1, $E_{1,col}$, multiplied by two scaling factors $\phi$ and $\psi$. This mean is calculated during the calibration period associated with time of interest $t$ considering only times when the satellite was overhead, denoted $T_{c,overpass}(t)$. Practically, both $\phi$ and $\psi$ are informed by the model, which provides simulated data for all relevant surface and column quantities. Scaling factor $\phi(x,t)$ accounts for the change in surface concentration corresponding with a unit change in column concentration at location $x$ and time $t$. We approximate this sensitivity using a ratio of model values at this location and time:

$$\phi(x,t) \cong \frac{M(x,t,0)}{M_{col}(x,t,0)}. \quad (7)$$

Scaling factor $\psi(x,t,t')$ accounts for the ratio of changes in surface concentrations at location $x$ and time $t$ to changes at location $x$ and time $t'$. Again, we approximate this with a ratio of model values:

$$\psi(x,t,t') \cong \frac{M(x,t,0)}{M(x,t',0)}. \quad (8)$$

The definition of $D(x,t)$ presented in equation 6 is a generalization of "typical pattern" extraction described in equations 1 and 2 of Malings et al. (2021). This generalization now explicitly captures the relationship between surface concentrations and column quantities, which was only implicit before. Equation 5 here then replaces equation 3 of Malings et al. (2021). A schematic for this phase is provided in Supplemental Figure S2.

In general, it may be necessary to consider the observational operator and air mass factor used in the satellite retrieval algorithm, as these affect the comparability between satellite retrieved $S_{col}$ and modeled $E_{1,col}$ (e.g., Cooper et al., 2020). No explicit consideration of this is made here; instead, this will contribute to variability as discussed below. Future work may explicitly consider these impacts, likely leading to a reduced uncertainty. Note that in the case of $PM_{2.5}$, AOD would be the column quantity considered.

Similar to phase 1, the uncertainty of the phase 2 estimate, $V_2(x,t,\tau)$, is estimated as the sum of the uncertainty due to forecast lead-time, $V_{F2}(x,t,\tau)$, the local variability of the model, $V_M(x,t)$, the variance in the satellite-informed sub-grid difference pattern, $V_D(x,t)$, twice the co-variance of the model and sub-grid difference pattern, $V_{MD}(x,t)$, the uncertainty due to the potential bias in the model-and-satellite-derived surface concentration estimates, $\mathrm{V}_{B2}(x,t)$, and the uncertainty due to the representational error of the model-and-satellite-derived surface concentration estimates, $V_{R2}(x,t)$:

$$V_2(x,t,\tau) = V_{F2}(x,t,\tau) + V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t) + \mathrm{V}_{B2}(x,t) + V_{R2}(x,t). \quad (9)$$

Model local variability $V_M(x,t)$ is carried from phase 1, and as in phase 1, $V_{F2}(x,t,\tau)$ can be empirically estimated by examining the mean squared difference of forecasts with lead time $\tau$ over the calibration interval at the same time of day:

$$V_{F2}(x,t,\tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)} \left[ \left( E_2(x,t',\tau) - E_2(x,t',0) \right)^2 \right]. \quad (10)$$

$V_D(x,t)$ and $V_{MD}(x,t)$ can be estimated with the empirical variance and co-variance of relevant terms involved in computation of the satellite-informed sub-grid difference pattern:

$$V_D(x,t) \cong \mathbb{V}_{t' \in T_{c,overpass}(t)} \left[ \left( S_{col}(x,t') - E_{1,col}(x,t') \right) \phi(x,t') \, \psi(x,t,t') \right], \quad (11)$$

where $\mathbb{V}$ denotes a variance computation, and:

$$V_{MD}(x,t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[ \left( E_1(x',t') - E_1(x,t) \right) \left( D(x',t') - D(x,t) \right) \right]. \quad (12)$$

Note that in this formulation, $X_n(x)$ now denotes the neighboring locations of $x$ at the (finer) spatial resolution of the satellite data, i.e., the adjoining pixel centroids. The final terms related to bias $\mathrm{V}_{B2}(x,t)$ and representational errors $V_{R2}(x,t)$ again cannot be estimated using the model and satellite information alone and require surface-level information, as will be discussed later (see Section 2.2.5).

Comparing $V_1(x,t,\tau)$ with $V_2(x,t,\tau)$, and assuming a zero lead-time such that forecast-related uncertainty can be ignored, we can establish some constraints on the bias and representational error from phase 1 using phase 2 results. Due to the inclusion of satellite data in phase 2 compared to phase 1, we might assume that $V_2(x,t,\tau)$ will be less than or equal to $V_1(x,t,\tau)$ generally. Thus:

$$V_{B1}(x,t) + V_{R1}(x,t) \geq V_D(x,t) + 2V_{MD}(x,t) + \mathrm{V}_{B2}(x,t) + V_{R2}(x,t). \quad (13)$$

That is, uncertainty due to bias and representativity errors in phase 1 should be larger than the analogous terms from phase 2 plus the variance and co-variance related to the satellite-informed sub-model-grid difference patterns. Note that the inclusion of satellite information is informing both sub-model-grid variability, which would tend to reduce (though not eliminate) representational errors captured in $V_{R1}(x,t)$, as well as bringing in real-world measurement data, which would tend to reduce (though not eliminate) model bias as represented in $V_{B1}(x,t)$. Using this relationship, estimates of the phase 1 uncertainty terms can be made based on the relevant phase 2 uncertainty terms, e.g., using the average of these terms within each model grid cell.

### *2.2.3 Phase 3: linear correction with reliable surface measurements*

Phase 3 uses in-situ measurement data to correct for possible regional systematic errors in the model-and-satellite-derived estimates of surface air quality from phase 2. As a simple case, a linear correction is assumed with slope $\alpha$ and intercept $\beta$:

$$E_3(x,t,\tau) = \alpha\, E_2(x,t,\tau) + \beta. \quad (14)$$

This corresponds directly with equation 10 of Malings et al. (2021).

Coefficients $\alpha$ and $\beta$, as well as estimates of their variance $V_\alpha$ and $V_\beta$, co-variance $V_{\alpha\beta}$, and residual regression variance $V_{R3}$, are derived from a linear regression analysis between phase 2 estimates $E_2(x,t)$ as the independent variable and ground-based air quality measurements $G(x,t)$ as the dependent variable over the calibration time interval $T_c$ and the set of discrete surface monitoring sites in the region available during the calibration time interval $X_c$:

$$\alpha, \beta, V_\alpha, V_\beta, V_{\alpha\beta}, V_{R3} = \mathbb{LR}_{t' \in T_c(t), x' \in X_c(x)}[G(x',t') \sim E_2(x',t',0)], \quad (15)$$

where $\mathbb{LR}_{domain}[v_d \sim v_i]$ denotes a linear regression with independent variable $v_i$ and dependent variable $v_d$, conducted over a domain specified in the subscript of $\mathbb{LR}$. Since this regression is being applied for historical data collected during the calibration time interval, the phase 2 estimate with $\tau = 0$ is used, and so $\tau$ has been dropped here for notational convenience. Note that a weighted linear regression can be applied, e.g., using a weight factor related to the time-of-day as suggested in previous work (Malings et al., 2021, Section 3.5). In principle, other approaches to regression can also be applied, including for example machine learning techniques to account for non-linear relationships (e.g., as in Wei et al., 2023). In such a case, appropriate characterization of the variance of the regression estimates and their covariance with explanatory inputs would have to be performed. In this work, a linear regression approach is adopted as there are well known closed-form solutions for computing the variance and covariance of the parameters. A schematic for this phase is provided in Supplemental Figure S3.

In cases where both RGM and LCS provide in-situ data, a modified approach is recommended. First, available RGM are used in phase 3 as outlined above. Then, LCS are

regionally calibrated before incorporating their data in phase 4. Details are provided in supplemental text S2.

Uncertainty in the phase 3 estimate is based on the phase 2 estimated uncertainty, re-scaled with regression terms, and with the uncertainties in these regression terms and residual variance included:

$$V_3(x,t,\tau) = V_{F3}(x,t,\tau) + \alpha^2[V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)] + V_\alpha E_2(x,t)^2 + 2V_{\alpha\beta}E_2(x,t) + V_\beta + V_{R3}. \quad (16)$$

Now that in-situ data have been included, systematic bias due to the misrepresentation of the surface air quality due to model and satellite information only, as well as representational issues due to the limited spatial resolutions of the model and satellite data with respect to specific points represented in the surface data, are considered to be captured in terms related to regression coefficient variance and residual variance. However, practical limitations on the availability of surface air quality measurement sites, as well as the tendencies of such sites to be clustered in high-population-density areas, might mean that there are some residual biases which are not fully captured in this formulation. In other words, by necessity, the data fusion process will be tailored towards better representing locations where surface monitors already exist, and the above formulation for phase 3 uncertainty will tend to be more appropriate in those types of areas, rather than, e.g., more rural areas which are not covered by surface-based monitors. Furthermore, biases in the in-situ data will not be accounted for, e.g., the known sensitivity of $NO_2$ monitors to other species (e.g., Steinbacher et al., 2007).

Comparing the phase 2 and 3 variance estimates, assuming zero lead-time, and assuming that inclusion of surface information will tend to decrease phase 3 uncertainty with respect to phase 2, we can establish that:

$$\mathrm{V}_{B2}(x,t) + V_{R2}(x,t) \geq (\alpha^2 - 1)[V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)] + V_\alpha E_2(x,t)^2 + 2V_{\alpha\beta}E_2(x,t) + V_\beta + V_{R3}. \quad (17)$$

Note that we have now established a "chain" of relationships connecting various bias and representational error terms, which could not be directly quantified, to terms which can be empirically estimated based on the data fusion process. This gives us a basis for quantifying these uncertainties in earlier phases as well; this will be discussed further in Section 2.2.5.

### *2.2.4 Phase 4: updating with recent, nearby in-situ data*

Phase 4 enables the use of recent and nearby surface measurement data to provide updates to estimates and forecasts from phase 3 via a spatio-temporal kriging approach. This process is expressed as:

$$E_4(x,t,\tau) = E_3(x,t,\tau) + \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x,x',t,t')\,[G(x',t') - E_3(x',t')], \quad (18)$$

where $X_{near}(x)$ denotes surface measurement locations arbitrarily "nearby" to $x$, $T_{recent}(t)$ denotes times arbitrarily "recent" with respect to $t$, and $K(x,x',t,t')$ is the kriging update factor, encompassing the relationship between concentrations at spatio-temporal coordinates $x,t$ and $x',t'$. This relationship is a combination of variance and co-variance relationships between the locations as well as the measurement noise. $K(x,x',t,t')$ is evaluated with the assistance of a kernel function, used in Gaussian process regression to parameterize these co-variances based on,

e.g., the difference in space and time between the two sets of coordinates (Rasmussen & Williams, 2006). Recent work has proposed the use of Gaussian process regression for interpolating air quality data in space and/or time based on sparse measurements, and have proposed using square exponential, Matérn, and periodic kernel functions for this purpose for different pollutants of interest (Jang et al., 2020; Malings et al., 2021; Wang et al., 2021). The approach used here to determine appropriate kernel functions and parameters is described in (Malings et al., 2021, section 3.7). Equation 18 combines equations 11 and 14 of Malings et al. (2021), using a more generic notation of the kernel. A schematic for this phase is provided in Supplemental Figure S4.

Spatio-temporal kriging also quantifies the resulting uncertainty reduction:

$$V_4(x,t,\tau) = V_3(x,t,\tau) - \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x,x',t,t') \operatorname{cov}[E_3(x',t'), E_3(x,t)], \quad (19)$$

where $\operatorname{cov}[E_3(x',t'), E_3(x,t)]$ denotes the covariance between surface concentrations of the pollutant of interest between spatio-temporal coordinates $x, t$ and $x', t'$, which is again evaluated using the kernel function.

For practical purposes, appropriate definitions for $X_{near}(x)$ and $T_{recent}(t)$ will have to be chosen to balance accuracy with the computational intensiveness of considering many measurements in this updating, which is a typical limitation of Gaussian process regression. In this paper, we use all surface measurement locations in our application region but use only the most recent measurement from each location.

### *2.2.5 Quantifying uncertainties in phases 1 and 2*

Following phases 1 and 2 of the data fusion approach outlined above, there remain several terms related to potential bias and representativity errors which are not quantifiable given the inputs available at these phases. However, following phase 3, the inclusion of ground-based monitor data allowed the full quantification of uncertainty as expressed in equation (16). Using this fact, alongside the inequality relationships presented in equations (13) and (17), we conducted an empirical analysis comparing the quantified uncertainties at different phases. Based on this analysis, we propose the following parametric estimates for the unquantified portions of the uncertainties in phases 1 and 2:

$$V_{B1}(x,t) + V_{R1}(x,t) \cong \eta_1^2(t \bmod 24\text{h})\, \mathbb{E}_{t' \in T_c(t)}\, V_M(x,t'), \quad (20)$$

$$\mathrm{V}_{B2}(x,t) + V_{R2}(x,t) \cong \eta_2^2(t \bmod 24\text{h})\, \mathbb{E}_{t' \in T_c(t)}[V_M(x,t') + V_D(x,t') + 2V_{MD}(x,t')]. \quad (21)$$

In these estimates, the unquantified portions of the uncertainty are related to the quantified performance via empirically determined factors $\eta_1$ for phase 1 and $\eta_2$ for phase 2. These factors are assumed to vary as a function of time-of-day, based on observations for how relationships between different portions of the quantified uncertainty varied over the calibration period investigated here. Empirically determined values of $\eta_1$ and $\eta_2$ for San Francisco are presented Supplemental Figure S5; values for New York City are presented in Supplemental Figure S6.

This proposed approach has important limitations. Most notably, it relies on proceeding to phase 3 of the data fusion approach. In regions without ground-based monitoring, or where only a small number of ground-based monitors are available, the results from phase 3 of the data fusion approach will be unavailable or highly unreliable. Empirically determined values of $\eta_1$ and $\eta_2$

from another region might be used, but there is no reason to expect these to generalize well. Thus, in the absence of surface data, full uncertainty quantification in phase 1 or 2 of the data fusion approach becomes unreliable.

### 2.3 Confidence interval determination

Following the approaches for data fusion with uncertainty quantification presented in the previous section, for a location of interest $x$ and time of interest $t$, with forecast lead time $\tau$, and for data fusion phase $p$, a data fusion "best estimate" for the quantity of interest $E_p(x,t,\tau)$ will be available, along with an uncertainty estimate for this quantity, $V_p(x,t,\tau)$. To make practical use of these outputs, in this work, we use them to define confidence intervals (CI) for our estimates or forecasts. To do this, a probabilistic distribution must be assumed for the quantity of interest. In this work, we assume a lognormal distribution, which is a typical assumption for many non-negative quantities relevant to air quality. This distribution is parameterized by the mean $\mu$ and standard deviation $\sigma$ of the associated normal distribution. These are calculated from the outputs of the data fusion process as follows:

$$\mu_p(x,t,\tau) = \log\left[\frac{E_p(x,t,\tau)}{\sqrt{1+\frac{V_p(x,t,\tau)}{E_p(x,t,\tau)^2}}}\right], \tag{22}$$

$$\sigma_p(x,t,\tau) = \sqrt{\log\left[1+\frac{V_p(x,t,\tau)}{E_p(x,t,\tau)^2}\right]}. \tag{23}$$

The quantity of interest $F_p(x,t,\tau)$ is then a lognormally distributed random variable:

$$F_p(x,t,\tau) \sim \mathrm{LN}\left(\mu_p(x,t,\tau), \sigma_p(x,t,\tau)\right). \tag{24}$$

where $\mathrm{LN}(\mu,\sigma)$ denotes a lognormal distribution with mean $\mu$ and standard deviation $\sigma$ for the associated normal distribution. This distribution can be used to determine a CI for the quantity of interest. For example, the 75 % confidence range is defined with a lower bound, representing the 12.5$^{\text{th}}$ percentile of the lognormal distribution, and an upper bound, representing the 87.5$^{\text{th}}$ percentile of the lognormal distribution.

The lognormal distribution assumption is of course an approximation of the true distribution of the quantity of interest. Therefore, the CI determined as described above would not necessarily correspond to the actual CI for the quantity of interest, even if the mean and variance were known exactly. However, some assumption about the distribution of the quantity of interest is necessary, as its true distribution will not be known a priori.

## 3 Results

In this section, we investigate the performance of the proposed data fusion framework described above through testing with actual data. In all cases, a leave-one-site-out cross-validation approach is used. For the given domain of interest, data from all but one of the active ground monitoring sites are considered as inputs to the data fusion algorithm. Concentrations are estimated or forecast via the data fusion approach for the location of the single held-out site. All sites are cycled through in this manner, resulting in estimates and forecasts of concentrations at each monitoring site using data from all other sites. This allows for comparisons to be made between

actual concentration measurements at each site and the estimates or forecasts from the data fusion using all information except for any measurements at the site in question. This allows for evaluating how the method would perform at an arbitrary location without in-situ data. A 14-day moving calibration time window is used across all phases, i.e., for a given time of interest $t$ and forecast lead time $\tau$, the calibration interval $T_c$ ranges from $t - \tau - 14$ days to $t - \tau$. This ensures that only input data available at or before a given time are used, with lead time measured from the time of the most recently available data. However, data latency effects are not considered, e.g., satellite data are assumed to be available as soon as the satellite passes overhead. Data latency effects can be estimated by inflating the lead time, e.g., performance of a 1-day forecast using inputs with a 1-day data latency is assumed to be similar to a 2-day forecast.

For illustrative purposes, an example of time series output from the data fusion approach is presented in Figure 1. Outputs from phase 4 of the data fusion process, the colored line, including a 50 % CI, the colored area, are compared to actual measurements from the RGM at this location, the black line. In the figure, local midnight of September 17$^{th}$ is considered to be "the present" (marked by grey dotted vertical line). Before this time, estimates are shown considering zero lead time, i.e., GEOS-CF historical outputs are used together with satellite and RGM data available up to and including the indicated time. After midnight of September 17$^{th}$, forecasts are shown with increasing lead time, i.e., the latest GEOS-CF forecast initialized 12 UTC the previous day is used, together with satellite and RGM data collected prior to September 17$^{th}$. For the historical estimates, availability of in-situ measurements at other RGM sites has allowed short-term spikes to be better represented, with the CI likewise being wider to capture the variability. For the forecasts, such spikes are not specifically captured, but the CI tends to be wider throughout the timeseries, accounting for the potential for such spikes to occur. In this example, the estimated CI tend to be underconfident: 75 % of actual measurements fell within the 50 % CI depicted. An analysis of the accuracy and precision of the forecasts (not considering their confidence estimates) is presented in Supplemental Figure S7.

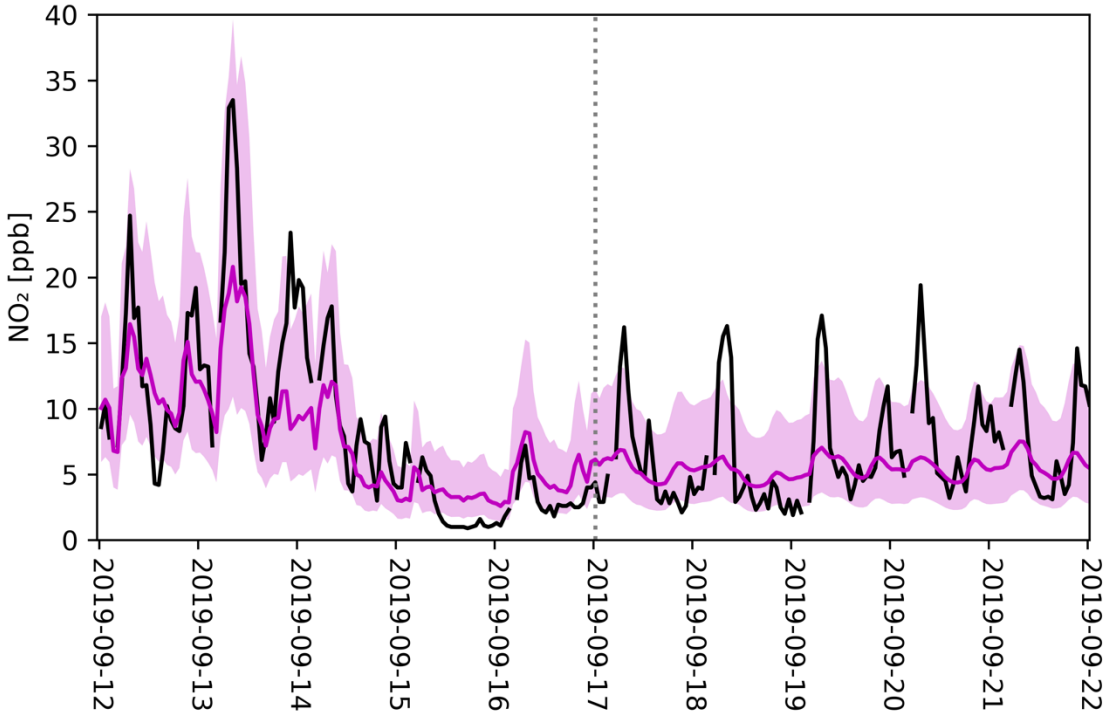



**Figure 1. Representative example of probabilistic estimates and forecasts for hourly surface-level $NO_2$ concentrations at the Redwood City monitor site (AQS ID 06-081-1001) in San Francisco, between September 12 and 22, 2019 local time. The black line indicates the reported concentrations from the regulatory monitor, i.e., the true concentration. The colored line indicates the mean estimated concentration from phase 4 of the data fusion process, $\boldsymbol{E_4(x,t)}$. The colored shaded areas denote the 50 % CI for the estimates. Estimates are presented with zero lead time up to midnight on September 17th, denoted with a vertical dotted line. Beyond this, forecasts with an increasing lead time are presented.**

### 3.1 Assessment of confidence interval coverage for different phases of data fusion

To investigate the accuracy of the assessed uncertainties in the data fusion, the fraction of actual measurements falling within the estimated 75 % CI across different phases of the data fusion approach is presented in Figure 2. This analysis considers all $NO_2$ monitor sites operating during September 2019 in the San Francisco study region, a total of 25 sites. The fraction of measurements falling within the 75 % CI is calculated for each site and considering the estimates for each phase of the data fusion process. Total uncertainties for phases 1 and 2 are estimated as outlined in section 2.2.5. Horizontal colored solid and dotted lines indicate the median, 25th percentile, and 75th percentile values of these fractions across all sites for each phase. Furthermore, sites are divided into types based on their assumed scale of spatial representativity, which is assessed for each monitoring site by US EPA. The five site types are microscale (0-0.1 km; 5 sites), middle (0.1-0.5 km; 3 sites), neighborhood (0.5-4 km; 13 sites), urban (4-50 km; 3 sites) and regional (50+ km; 1 site), as defined in 40 CFR Part 58. By investigating the capacity of the data fusion system to capture uncertainties at different spatial scales in this way, its benefits and limitations can be better understood.

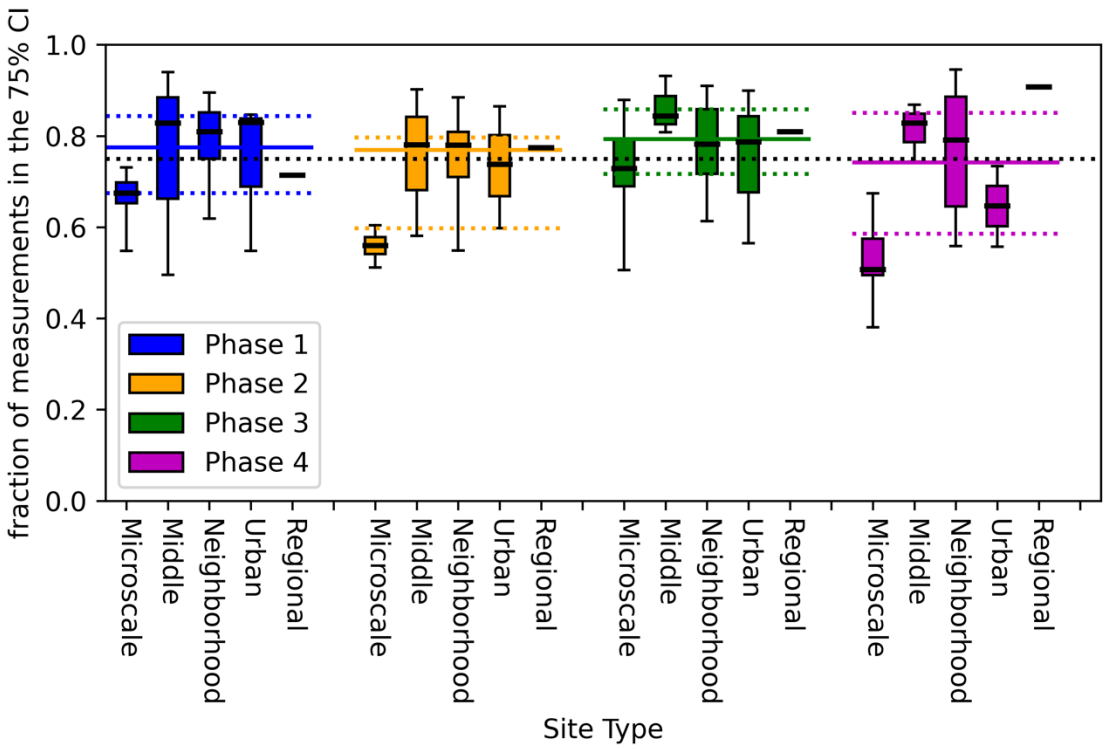



**Figure 2. Assessment of the fraction of actual measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors. The analysis represents data from 25 active $NO_2$ ground monitoring sites in the San Francisco study region for September 2019. A horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI. For each ground monitor site, the fraction of measurements at that site falling within the 75 % CI is calculated. For each phase, a solid horizonal line in the corresponding color indicates the median of these fractions across sites, and two horizontal dotted colored lines indicate the 25$^{th}$ percentile and 75$^{th}$ percentile of these fractions across sites. Furthermore, monitoring sites are divided into different site types. The spread in fraction of measurements falling within the 75 % CI for each site type is indicated with a box-and-whisker plot. In each box-and-whisker plot, the horizontal line inside the box denotes the median, the box denotes the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denote the full range.**

Overall, for all phases of the data fusion process, the estimated 75 % CI captures roughly 75 % of measured data. Performance is most consistent for phases 1 and 3, which have the smallest inter-quartile spreads in fraction of measurements falling within the 75 % CI. Focusing on phase 1, where only model outputs are considered, performance is consistent across most site types. There is a slight bias towards underconfidence, i.e., more measurements falling within the 75 % CI than expected. For microscale sites, however, estimates are systematically overconfident, with fewer measurements falling within the 75 % CI than expected. Considering the native spatial resolution of the model, better representation of uncertainties at urban and regional scales is to be expected. There is a lack of information at this stage to make informed assessments of confidence at finer spatial scales. This manifests in the results with a slightly larger spread in performance for middle scale sites and the overconfidence noted for microscale sites.

In phase 2, this is exacerbated, with increased overconfidence for estimates of microscale sites. Again, this can be explained by considering that, at phase 2, satellite data from TROPOMI with a nominal spatial resolution on the order of 5 km has been incorporated. This would be expected to improve assessments at neighborhood sites. This is reflected in the results with a slight decrease in the underconfidence of estimates for sites at this scale. However, there continues to be a lack of relevant information at finer spatial scales, and so while uncertainty estimates seem to have been improved for most scales, they have substantially degraded for microscale sites.

In phase 3, with the incorporation of ground-based data, uncertainties at microscale sites are now better represented overall, although one microscale site (denoted with the lower whisker) continues to be quite overconfidently estimated. However, middle scale sites are now being represented with systematic underconfidence. This might be a consequence of the relative numbers of sites in each type. There are 5 microscale and 3 middle scale sites in the study domain. Furthermore, because of the cross-validation approach, data from the site being evaluated are not included, underrepresenting that type. Thus, the approach of phase 3 would tend to better represent the more numerous site type. This could be accounted for by assigning lesser weights to certain types of sites when conducting the linear regression in phase 3. However, because one would not know a-priori the characteristics of the site at which concentrations are to be estimated, weighting different types of sites differently might not be an appropriate approach. Uncertainty estimates for neighborhood, urban, and regional sites appear reasonable, if slightly underconfident overall.

In phase 4, while uncertainty estimates seem to be most accurate in the median, the spread in performance has increased. Microscale sites are again exhibiting systematic overconfidence, along with urban scale sites, while middle scale and regional sites are underconfident. With only a single regional site, however, that latter result is not necessarily robust. This varied performance might be understood by considering that, due to the heterogeneity of urban areas, monitoring sites of different types will tend to be interspersed with one another. For a given site, the closest site which will have the greatest influence in the kriging approach of phase 4 is likely to be of a different type than the site being estimated for in the cross-validation. Neighborhood sites are least susceptible to this effect since, as the most numerous site type in the study area, the closest RGM to a neighborhood site is often another neighborhood scale site. The microscale sites, on the other hand, are closest to either neighborhood or urban scale sites, and the neighborhood or urban scale sites likewise are often closest to microscale sites. A kernel function for the kriging approach not based solely on distance might alleviate this difficulty, e.g., by defining similarities based on similar land use and land cover factors (e.g., Gilpin et al., 2023). Such an approach would require additional input information and is left as a subject for future improvements.

Across all phases, the best and most consistent results were observed for neighborhood scale sites. This is probably due in part to their relative abundance, but also to the fact that their representative scale (0.5-4 km) is of the same order as the satellite input data, which provides the most relevant information about spatial heterogeneity of pollutant concentrations. Overall, this is consistent with what might be expected, given the way in which the data fusion and associated uncertainty quantification are being conducted. Results were also similar for different CI (see Supplemental Figure S8).

### 3.2 Assessment of confidence interval coverage for different forecast lead times

Figure 3 presents an analysis of the fraction of measurements falling within the 75 % CI of the uncertainty estimate as a function of the forecasting lead time. Several discrete lead times are

considered, and results for zero lead time are also presented for comparison; these were previously presented in Figure 2.

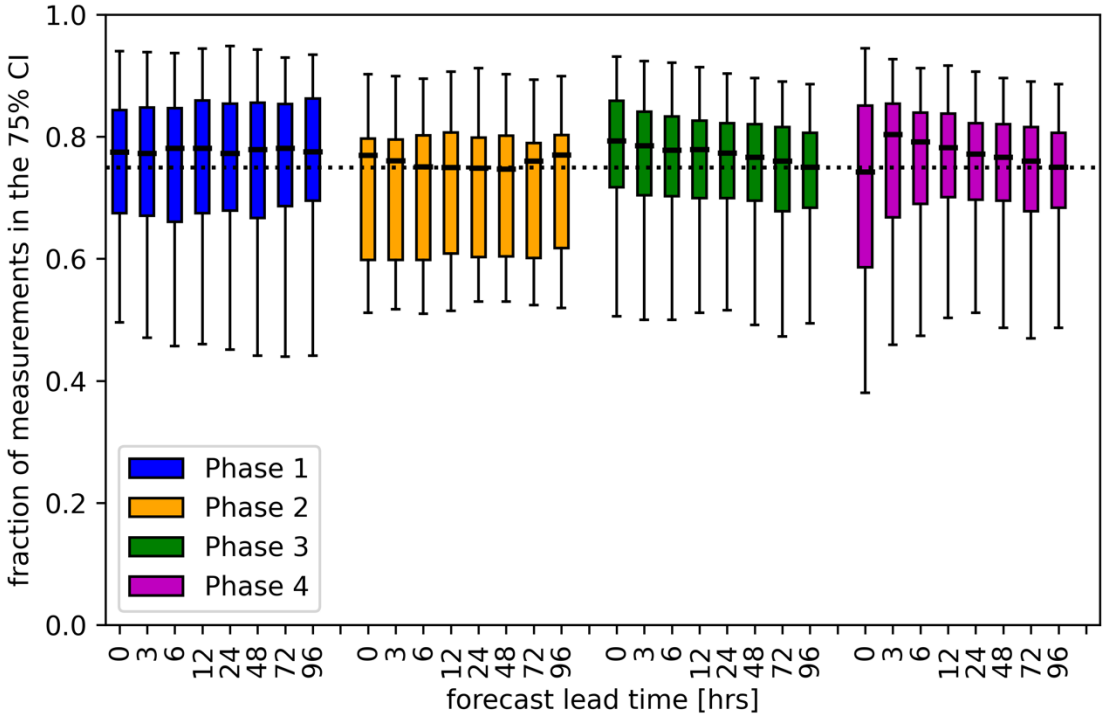



**Figure 3. Assessment of the fraction of actual measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, as a function of forecasting lead time, in hours. The analysis represents data from 25 active $NO_2$ ground monitoring sites in the San Francisco study region for September 2019. A horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI. For each ground monitor site, the fraction of measurements at that site falling within the 75 % CI is calculated. The box-and-whisker plots denote the ranges of these fractions across sites, with the horizontal line in the box denoting the median, the box denoting the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denoting the full range.**

Overall, there is little variation in the CI coverage as lead time increases, indicating that the uncertainty quantification approach is applicable for forecasts as well as historical estimates. For phase 3, there appears to be a tendency towards underconfidence at shorter lead times. For phase 4, the spread in coverage decreases as the forecasting lead time increases. As noted previously, the kriging approach of phase 4 with a distance-based kernel tends to induce under- or overconfidence at nearby sites. As the forecasting lead time increases, the influence of the most recent measurement data decreases, and the uncertainty quantification resembles that of phase 3. While the incorporation of near-real-time data in phase 4 has notable benefits in terms of near-term forecast accuracy, as noted in previous work (Malings et al., 2021), these results indicate that there is also a trade-off in terms of slightly less realistic uncertainty estimates in the phase 4 near-term forecasts compared to the other phases and to longer lead times.

### 3.3 Assessment of confidence interval coverage across different times of year

As an additional assessment, the methodology was applied across different months. Results for CI coverage at zero forecast lead time in March 2019, June 2019, September 2019 (as presented previously), and December 2019 are shown in Figure 4. There is some variability in performance for different phases in different months. For example, in December 2019, phases 1, 3, and 4 show a tendency for underconfidence in their estimates, although this is not apparent in phase 2. Conversely, phase 2 exhibits overconfidence in June 2019, while this is not apparent for other phases. This might indicate monthly or seasonally varying biases in the input data sources which are not accounted for in the current method.

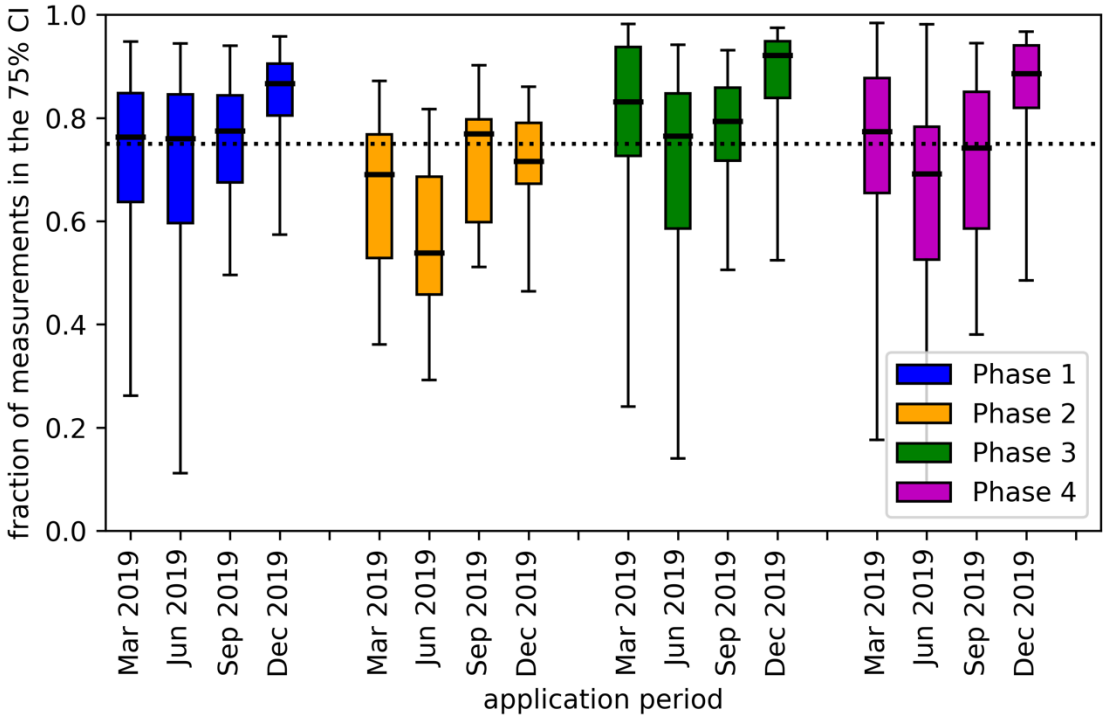



**Figure 4. Fractions of measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, presented for different application months. Box-and-whisker plots denote ranges of these fractions across active $NO_2$ monitor sites in San Francisco during that month, with the horizontal line in the box denoting the median, the box denoting the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denoting the full range. The horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI.**

A similar assessment was conducted for the region of New York City, as discussed in the supplemental materials. Results for CI coverage at zero forecast lead time in March 2019, June 2019, September 2019, and December 2019 are shown in Supplemental Figure S9. Similar variability in performance for different phases in different months is observed as was noted above. Underconfidence in December 2019 seems to be more extreme, especially in phase 1, than in the case of San Francisco. Overconfidence in phase 2 also appears to be more severe. Again, monthly or seasonal differences in relevant parameters, especially the factors $\eta_1$ and $\eta_2$ calculated for the domain and kriging spatial and temporal scales associated with phase 4, might be influencing this.

The fact that month-to-month differences appear to be greater in New York City, where seasonal differences in prevailing meteorological conditions are relatively greater than in San Francisco, where such changes are relatively smaller, seems to corroborate this hypothesis. Thus, future development should focus on better capturing such seasonal changes through dynamically recalculating relevant parameters as part of the calibration process.

## 4 Conclusions

Overall, the proposed framework to estimate uncertainties and CI for concentration estimates from data fusion produced reasonable results in most cases, with most CI coverage being within about 10 percentage points of the theoretical value. There were also few instances of extreme overconfidence (few measurements falling within the prescribed CI) or extreme underconfidence (almost all measurements falling within the prescribed CI) observed in the results presented here. These findings are encouraging given the various assumptions made in defining the uncertainty quantification framework, including the assumption of lognormally distributed concentrations.

The uncertainty quantification was found to be least accurate overall for microscale sites, which are most impacted by hyperlocal sources. In the San Francisco case study, these sites were adjacent to highways, which are most heavily impacted by $NO_2$ pollution. This finding is useful to convey to any user of this system, i.e., that results may not be reliable within about 100 meters of a major source like a highway or other intense combustion activity. Similar limitations are likely, should the method be applied to other constituents measured near their respective sources.

It is also important to note that CI assessments are not being provided for independent data, but rather there is significant autocorrelation in the data. For example, while a measurement might have a 50 % chance of falling within a 50 % CI a-priori, if it is known that a recent measurement fell outside this CI, it becomes much less likely that a new measurement will fall within the CI. This effect can be noted on September 15$^{th}$ in Figure 1, when multiple measurements in sequence were observed outside the 50 % CI.

Several areas of theoretical and practical improvement are noted for future work. As suggested in Section 2.2.1, use of an ensemble of models rather than a single model in phase 1 would allow for estimating uncertainties at that phase based on variability across the ensemble. For incorporating satellite data in phase 2, multiple sources of satellite data might be considered, offering coverage at different times of day. Geostationary instruments like the recently launched TEMPO might be particularly useful in establishing different values of $D(x,t)$ corresponding to different times of day. Better definitions for the calibration dataset might also be explored, in contrast to a simple moving time window as presented in Section 2.2.2. For example, forecasted conditions might be matched to similar past conditions for which satellite data were available, in an attempt to identify past situations which approximately match forecasted future conditions in order to define a more suitable calibration dataset. There is also the possibility to include ancillary datasets, such as land use information, as additional co-variates to explain local variability. These might be incorporated using more sophisticated regression techniques, such as machine learning approaches, in contrast to the linear techniques presented for phase 3 in Section 2.2.3. While it would be necessary to develop customized uncertainty quantification schemes for these techniques, they might be better suited to capturing non-linear relationships in the data. Finally, the limitation of ground data availability and the resulting tendency of the approach to be biased towards such areas, as mentioned in Section 2.2.3, might be addressed in a more systematic way,

e.g., via resampling or application of different weightings to data from different types of monitoring sites in order to create a more unbiased calibration dataset.

Nevertheless, the framework established here presents a reasonable prior CI for the estimates and forecasts of the proposed data fusion system, and this fact supports effective and appropriate interpretation of its output by users. For example, these uncertainty estimates might be applied with respect to a given regulatory pollutant threshold to estimate the probability of exceeding that threshold. Such information could support air quality management decision-making. In an ongoing project supported by the NASA Health and Air Quality Applied Sciences Program, the authors are implementing the data fusion and uncertainty quantification scheme presented here in an online application via the Google Earth Engine platform. It is hoped that this application will present a useful tool for local air quality managers to visualize sub-city-scale atmospheric composition and variability using a combination of model, satellite, and in-situ data. This project is being conducted in collaboration with local environmental managers in the USA, Brazil, and Senegal. An example prototype for this tool is presented in Figure 5. As part of this project, the framework will also be extended to other relevant pollutants, primarily $PM_{2.5}$ and $O_3$.

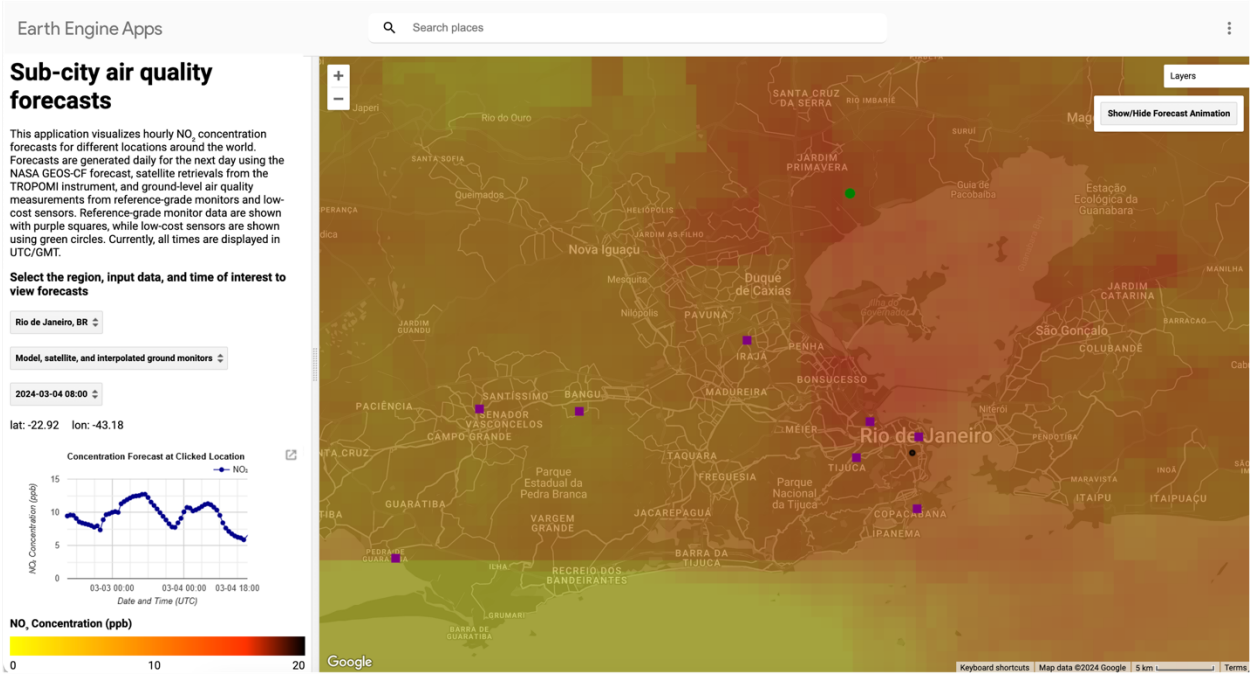



**Figure 5. Screenshot of an application currently under development which will implement the data fusion framework presented here, including uncertainty quantification, via the Google Earth Engine platform. This application will enable air quality managers to access and visualize estimates and forecasts of relevant air quality parameters such as $NO_2$, $O_3$, $PM_{2.5}$, along with associated expressions of confidence. Example outputs are presented for the city of Rio de Janeiro, Brazil, one of the partners for this project.**

## Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) under Grants 80NSSC22K1473 and WBS 389018.02.09.02.72 issued through the NASA Health and Air Quality Applied Sciences Program. The authors would also like to acknowledge the participation of Alan Chan, Sean Khan, John White, Daniel Westervelt, and Sean Wihera in

that grant project. The authors would like to thank Callum Wayman for consultations related to the implementation of the data fusion and uncertainty quantification scheme on the Google Earth Engine platform, Daniel King for software implementation to support the Google Earth Engine application, and Karin Tuxen-Bettman for guidance and assistance with ingesting necessary input datasets into Google Earth Engine. Finally, the authors would like to thank Felipe Mandarino, Bruno Boscaro, and Oswaldo Cruz for their comments and feedback during the development of the prototype depicted in Figure 5.

## Open Research

GEOS-CF outputs are available via the GMAO website; "AQC" and "XQC" collection files have been used here. Other input data are available via NASA GES DISC and the US EPA Air Quality System. Data and code used to generate the figures presented in this paper are available in an online Zenodo archive (Malings, 2024), governed under a CC BY-NC License.

## Author Contributions

Carl Malings: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft; K. Emma Knowland: Conceptualization, Supervision, Writing – Review & Editing; Nathan Pavlovic: Conceptualization, Writing – Review & Editing; Justin Coughlin: Software, Visualization, Writing – Review & Editing; Christoph Keller: Conceptualization; Stephen Cohn: Conceptualization, Methodology, Writing – Review & Editing; Randall Martin: Writing – Review & Editing.

## References

Anenberg, S. C., Bindl, M., Brauer, M., Castillo, J. J., Cavalieri, S., Duncan, B. N., Fiore, A. M., Fuller, R., Goldberg, D. L., Henze, D. K., Hess, J., Holloway, T., James, P., Jin, X., Kheirbek, I., Kinney, P. L., Liu, Y., Mohegh, A., Patz, J., … West, J. J. (2020). Using Satellites to Track Indicators of Global Air Pollution and Climate Change Impacts: Lessons Learned From a NASA-Supported Science-Stakeholder Collaborative. *GeoHealth*, *4*(7). https://doi.org/10.1029/2020GH000270

Balashov, N. V., Huff, A. K., & Thompson, A. M. (2023). Interpretation of Probabilistic Surface Ozone Forecasts: A Case Study for Philadelphia. *Weather and Forecasting*, *38*(10), 1895–1906. https://doi.org/10.1175/WAF-D-22-0185.1

Balashov, N. V., Thompson, A. M., & Young, G. S. (2017). Probabilistic Forecasting of Surface Ozone with a Novel Statistical Approach. *Journal of Applied Meteorology and Climatology*, *56*(2), 297–316. https://doi.org/10.1175/JAMC-D-16-0110.1

Bi, J., Knowland, K. E., Keller, C. A., & Liu, Y. (2022). Combining Machine Learning and Numerical Simulation for High-Resolution PM $_{2.5}$ Concentration Forecast. *Environmental Science & Technology*, *56*(3), 1544–1556. https://doi.org/10.1021/acs.est.1c05578

Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating Low-Cost Sensor Measurements into High-Resolution PM $_{2.5}$ Modeling at a Large Spatial Scale. *Environmental Science & Technology*, *54*(4), 2152–2162. https://doi.org/10.1021/acs.est.9b06046

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov,

A., Carnevale, C., Grell, G., & Seigneur, C. (2015). Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*, *15*(10), 5325–5358. https://doi.org/10.5194/acp-15-5325-2015

Cooper, M. J., Martin, R. V., Henze, D. K., & Jones, D. B. A. (2020). Effects of a priori profile shape assumptions on comparisons between satellite NO<sub>2</sub> columns and model simulations. *Atmospheric Chemistry and Physics*, *20*(12), 7231–7241. https://doi.org/10.5194/acp-20-7231-2020

Diao, M., Holloway, T., Choi, S., O'Neill, S. M., Al-Hamdan, M. Z., Van Donkelaar, A., Martin, R. V., Jin, X., Fiore, A. M., Henze, D. K., Lacey, F., Kinney, P. L., Freedman, F., Larkin, N. K., Zou, Y., Kelly, J. T., & Vaidyanathan, A. (2019). Methods, availability, and applications of PM $_{2.5}$ exposure estimates derived from ground measurements, satellite, and atmospheric models. *Journal of the Air & Waste Management Association*, *69*(12), 1391–1414. https://doi.org/10.1080/10962247.2019.1668498

Diggle, P. J. (2010). Historical Introduction. In A. E. Gelfand, M. Fuentes, P. Guttorp, & P. J. Diggle (Eds.), *Handbook of spatial statistics* (pp. 3–14). CRC Press.

Duncan, B. N., Malings, C. A., Knowland, K. E., Anderson, D. C., Prados, A. I., Keller, C. A., Cromar, K. R., Pawson, S., & Ensz, H. (2021). Augmenting the Standard Operating Procedures of Health and Air Quality Stakeholders With NASA Resources. *GeoHealth*, *5*(9). https://doi.org/10.1029/2021GH000451

Fanfarillo, A., Roozitalab, B., Hu, W., & Cervone, G. (2019). *Probabilistic Forecasting using Deep Generative Models*. https://doi.org/10.48550/ARXIV.1909.11865

Garaud, D., & Mallet, V. (2011). Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality. *Journal of Geophysical Research*, *116*(D19), D19304. https://doi.org/10.1029/2011JD015780

Garner, G. G., & Thompson, A. M. (2012). The Value of Air Quality Forecasting in the Mid-Atlantic Region. *Weather, Climate, and Society*, *4*(1), 69–79. https://doi.org/10.1175/WCAS-D-10-05010.1

Gilliam, R. C., Hogrefe, C., Godowitch, J. M., Napelenok, S., Mathur, R., & Rao, S. T. (2015). Impact of inherent meteorology uncertainty on air quality model predictions. *Journal of Geophysical Research: Atmospheres*, *120*(23). https://doi.org/10.1002/2015JD023674

Gilpin, S., Matsuo, T., & Cohn, S. E. (2023). A generalized, compactly supported correlation function for data assimilation applications. *Quarterly Journal of the Royal Meteorological Society*, *149*(754), 1953–1989. https://doi.org/10.1002/qj.4490

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., & Subramanian, R. (2021). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, *158*, 105833. https://doi.org/10.1016/j.jaerosci.2021.105833

Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environment International*, *143*, 105965. https://doi.org/10.1016/j.envint.2020.105965

788 Hassani, A., Schneider, P., Vogt, M., & Castell, N. (2023). Low-Cost Particulate Matter Sensors
789 for Monitoring Residential Wood Burning. *Environmental Science & Technology*, *57*(40), 15162–
790 15172. https://doi.org/10.1021/acs.est.3c03661

791 Hodoli, C. G., Coulon, F., & Mead, M. I. (2023). Source identification with high-temporal
792 resolution data from low-cost sensors using bivariate polar plots in urban areas of Ghana.
793 *Environmental Pollution*, *317*, 120448. https://doi.org/10.1016/j.envpol.2022.120448

794 Holloway, T., Miller, D., Anenberg, S., Diao, M., Duncan, B., Fiore, A. M., Henze, D. K., Hess, J.,
795 Kinney, P. L., Liu, Y., Neu, J. L., O'Neill, S. M., Odman, M. T., Pierce, R. B., Russell, A. G., Tong,
796 D., West, J. J., & Zondlo, M. A. (2021). Satellite Monitoring for Air Quality and Health. *Annual*
797 *Review of Biomedical Data Science*, *4*(1), 417–447. https://doi.org/10.1146/annurev-biodatasci-
798 110920-093120

799 Jang, J., Shin, S., Lee, H., & Moon, I.-C. (2020). Forecasting the Concentration of Particulate
800 Matter in the Seoul Metropolitan Area Using a Gaussian Process Model. *Sensors*, *20*(14), 3845.
801 https://doi.org/10.3390/s20143845

802 Judd, L. M., Al-Saadi, J. A., Janz, S. J., Kowalewski, M. G., Pierce, R. B., Szykman, J. J., Valin,
803 L. C., Swap, R., Cede, A., Mueller, M., Tiefengraber, M., Abuhassan, N., & Williams, D. (2019).
804 Evaluating the impact of spatial resolution on tropospheric NO2 column comparisons within urban
805 areas using high-resolution airborne data. *Atmospheric Measurement Techniques*, *12*(11), 6091–
806 6111. https://doi.org/10.5194/amt-12-6091-2019

807 Keller, C. A., Evans, M. J., Knowland, K. E., Hasenkopf, C. A., Modekurty, S., Lucchesi, R. A.,
808 Oda, T., Franca, B. B., Mandarino, F. C., Díaz Suárez, M. V., Ryan, R. G., Fakes, L. H., & Pawson,
809 S. (2020). *Global Impact of COVID-19 Restrictions on the Surface Concentrations of Nitrogen*
810 *Dioxide and Ozone* [Preprint]. Gases/Atmospheric Modelling/Troposphere/Chemistry (chemical
811 composition and reactions). https://doi.org/10.5194/acp-2020-685

812 Keller, C. A., Knowland, K. E., Duncan, B. N., Liu, J., Anderson, D. C., Das, S., Lucchesi, R. A.,
813 Lundgren, E. W., Nicely, J. M., Nielsen, E., Ott, L. E., Saunders, E., Strode, S. A., Wales, P. A.,
814 Jacob, D. J., & Pawson, S. (2021). Description of the NASA GEOS Composition Forecast
815 Modeling System GEOS-CF v1.0. *Journal of Advances in Modeling Earth Systems*, *13*(4).
816 https://doi.org/10.1029/2020MS002413

817 Kelp, M. M., Keller, C. A., Wargan, K., Karpowicz, B. M., & Jacob, D. J. (2023). Tropospheric
818 ozone data assimilation in the NASA GEOS Composition Forecast modeling system (GEOS-CF
819 v2.0) using satellite data for ozone vertical profiles (MLS), total ozone columns (OMI), and
820 thermal infrared radiances (AIRS, IASI). *Environmental Research Letters*, *18*(9), 094036.
821 https://doi.org/10.1088/1748-9326/acf0b7

822 Li, J., Zhang, H., Chao, C.-Y., Chien, C.-H., Wu, C.-Y., Luo, C. H., Chen, L.-J., & Biswas, P.
823 (2020). Integrating low-cost air quality sensor networks with fixed and satellite monitoring
824 systems to study ground-level PM2.5. *Atmospheric Environment*, *223*, 117293.
825 https://doi.org/10.1016/j.atmosenv.2020.117293

826 Li, Y., Martin, R. V., Li, C., Boys, B. L., van Donkelaar, A., Meng, J., & Pierce, J. R. (2023).
827 *Development and evaluation of processes affecting simulation of diel fine particulate matter*
828 *variation in the GEOS-Chem model* [Preprint]. Aerosols/Atmospheric Modelling and Data
829 Analysis/Troposphere/Chemistry (chemical composition and reactions).
830 https://doi.org/10.5194/egusphere-2023-704

Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O. L., Segers, A., & Heemink, A. W. (2021). Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia. *Atmosphere*, *12*(1), 91. https://doi.org/10.3390/atmos12010091

Malings, C. (2024). *Supporting Data for "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results"* (1.0) [Python]. Zenodo. https://doi.org/10.5281/zenodo.10650853

Malings, C., Knowland, K. E., Keller, C. A., & Cohn, S. E. (2021). Sub-City Scale Hourly Air Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements. *Earth and Space Science*, *8*(7). https://doi.org/10.1029/2021EA001743

Martin, R. V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., & Dey, S. (2019). No one knows which city has the highest concentration of fine particulate matter. *Atmospheric Environment: X*, *3*, 100040. https://doi.org/10.1016/j.aeaoa.2019.100040

McFarlane, C., Isevulambire, P. K., Lumbuenamo, R. S., Ndinga, A. M. E., Dhammapala, R., Jin, X., McNeill, V. F., Malings, C., Subramanian, R., & Westervelt, D. M. (2021). First Measurements of Ambient PM2.5 in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of Congo Using Field-calibrated Low-cost Sensors. *Aerosol and Air Quality Research*, *21*. https://doi.org/10.4209/aaqr.200619

McFarlane, C., Raheja, G., Malings, C., Appoh, E. K. E., Hughes, A. F., & Westervelt, D. M. (2021). Application of Gaussian Mixture Regression for the Correction of Low Cost PM $_{2.5}$ Monitoring Data in Accra, Ghana. *ACS Earth and Space Chemistry*, acsearthspacechem.1c00217. https://doi.org/10.1021/acsearthspacechem.1c00217

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abegaz, K. H., Abolhassani, H., Aboyans, V., Abreu, L. G., Abrigo, M. R. M., Abualhasan, A., Abu-Raddad, L. J., Abushouk, A. I., Adabi, M., … Lim, S. S. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Raheja, G., Sabi, K., Sonla, H., Gbedjangni, E. K., McFarlane, C. M., Hodoli, C. G., & Westervelt, D. M. (2022). A Network of Field-Calibrated Low-Cost Sensor Measurements of $PM_{2.5}$ in Lomé, Togo, Over One to Two Years. *ACS Earth and Space Chemistry*, *6*(4), 1011–1021. https://doi.org/10.1021/acsearthspacechem.1c00391

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Riccio, A., & Chianese, E. (2024). Technical note: Accurate, reliable, and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method. *Atmospheric Chemistry and Physics*, *24*(3), 1673–1689. https://doi.org/10.5194/acp-24-1673-2024

Rose Eilenberg, S., Subramanian, R., Malings, C., Hauryliuk, A., Presto, A. A., & Robinson, A. L. (2020). Using a network of lower-cost monitors to identify the influence of modifiable factors driving spatial patterns in fine particulate matter concentrations in an urban environment. *Journal of Exposure Science & Environmental Epidemiology*. https://doi.org/10.1038/s41370-020-0255-x

Schneider, P., Vogt, M., Haugen, R., Hassani, A., Castell, N., Dauge, F. R., & Bartonova, A. (2023). Deployment and Evaluation of a Network of Open Low-Cost Air Quality Sensor Systems. *Atmosphere*, *14*(3), 540. https://doi.org/10.3390/atmos14030540

Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., & Hueglin, C. (2007). Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques. *Journal of Geophysical Research: Atmospheres*, *112*(D11), 2006JD007971. https://doi.org/10.1029/2006JD007971

Tanzer, R., Malings, C., Hauryliuk, A., Subramanian, R., & Presto, A. A. (2019). Demonstration of a Low-Cost Multi-Pollutant Network to Quantify Intra-Urban Spatial Variations in Air Pollutant Source Impacts and to Evaluate Environmental Justice. *International Journal of Environmental Research and Public Health*, *16*(14), 2523. https://doi.org/10.3390/ijerph16142523

US EPA. (2017). *Policy Assessment for the Review of the Primary National Ambient Air Quality Standards for Oxides of Nitrogen* (EPA-452/R-17-003). U.S. Environmental Protection Agency.

van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., & Martin, R. V. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology*, *55*(22), 15287–15300. https://doi.org/10.1021/acs.est.1c05309

van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2015). Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. *Environmental Health Perspectives*, *123*(2), 135–143. https://doi.org/10.1289/ehp.1408646

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., … Levelt, P. F. (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

Wang, P., Mihaylova, L., Chakraborty, R., Munir, S., Mayfield, M., Alam, K., Khokhar, M. F., Zheng, Z., Jiang, C., & Fang, H. (2021). A Gaussian Process Method with Uncertainty Quantification for Air Quality Monitoring. *Atmosphere*, *12*(10), 1344. https://doi.org/10.3390/atmos12101344

Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., & Zhu, T. (2023). First close insight into global daily gapless 1 km PM2.5 pollution, variability, and health impact. *Nature Communications*, *14*(1), 8349. https://doi.org/10.1038/s41467-023-43862-3

Zhang, H., Wang, J., García, L. C., Ge, C., Plessel, T., Szykman, J., Murphy, B., & Spero, T. L. (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

Zhang, H., Wang, J., García, L. C., Zhou, M., Ge, C., Plessel, T., Szykman, J., Levy, R. C., Murphy, B., & Spero, T. L. (2022). Improving Surface PM $_{2.5}$ Forecasts in the United States Using an Ensemble of Chemical Transport Model Outputs: 2. Bias Correction With Satellite Data for Rural

913 Areas. *Journal of Geophysical Research: Atmospheres*, *127*(1).
914 https://doi.org/10.1029/2021JD035563

915

# Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results

**Carl Malings[1,2][0000-0002-2242-4328], K. Emma Knowland[1,2][0000-0003-0837-8502], Nathan Pavlovic[3][0000-0003-2127-3940], Justin G. Coughlin[3][0000-0003-3882-3064], Christoph Keller[1,2][0000-0002-0552-4298], Stephen Cohn[2][0000-0001-8506-9354], and Randall V. Martin[4][0000-0003-2632-8402]**

[1]Morgan State University, GESTAR II Cooperative Agreement, Baltimore, MD 21251, USA.

[2]NASA Goddard Space Flight Center, Global Modeling & Assimilation Office, Greenbelt, MD 20771, USA.

[3]Sonoma Technology, Inc., Petaluma, CA 94954, USA.

[4]Washington University in St. Louis., St. Louis, MO 63130, USA.

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

**Key Points:**

- The proposed data fusion method produces a-priori uncertainty assessments and confidence intervals for estimates and forecasts
- Confidence intervals were found to be mostly reasonable in a test case study for nitrogen dioxide across four months and two cities
- The method provided overconfident estimates for sites within 100 meters of highways

## Abstract

Integrating air quality information from models, satellites, and in-situ monitors allows for both better estimation of air quality and better quantification of uncertainties in this estimation. Uncertainty quantification is important to appropriately convey confidence in these estimates and forecasts to users who will base decisions on these. Uncertainty quantification also allows tracing the value of information provided by different data sources. This can identify gaps in the monitoring network where additional data could further reduce uncertainties. This paper presents a framework for data fusion with uncertainty quantification, applicable to multiple air-quality-relevant pollutants. Testing of this framework in the context of nitrogen dioxide forecasting at sub-city scales shows promising results, with confidence intervals typically encompassing the expected number of actual measurements during cross-validation. The framework is now being implemented into an online tool to support local air quality management decision-making. Future work will also include the incorporation of low-cost air sensor data and the quantification of uncertainty at hyper-local scales.



## Plain Language Summary

Poor air quality has adverse impacts on human and environmental health. Estimating and forecasting air quality accurately can improve early warnings and mitigation for poor air quality. Furthermore, understanding the uncertainties and degree of confidence in these forecasts and estimates can help air quality managers know when and where they can be relied upon, and where more data might still be needed. This paper outlines a method to combine air quality information from models, satellites, and ground-based monitors, and to assess the confidence in the combined output. Combining all these data sources can give us a better overall understanding of air quality, and making comparisons between them allows us to better understand uncertainties. Testing out the method proposed in this paper, we find that the method can produce reasonable assessments of the confidence it has in its estimates, with the expected numbers of actual measurements usually falling within the confidence intervals produced by the method. An exception is when this method is applied very close to a major pollution source (e.g., a highway, in our study). In such cases, since the method does not know that there is such a source nearby, it tends to be overconfident in its prediction.


## 1 Introduction

Poor air quality is a major global public health concern. The 2019 Global Burden of Disease study identified air pollution as the leading environmental risk factor for human premature mortality (Murray et al., 2020). To mitigate this public health problem on a global scale, air quality managers and practitioners first need access to accurate and comprehensive information on the state of air quality in their areas. Such information might come from a variety of disparate sources. In-situ measurements of air quality, typically obtained from instruments operated by regulatory bodies, e.g., the Environmental Protection Agency in the United States, are considered the trusted standard for assessing air quality. At a global scale, however, the relatively low density of such measurements means that regulatory instruments alone often cannot provide necessary air quality information to answer basic questions relevant to public health (Martin et al., 2019). Low-cost air quality sensors (LCS) are increasing in prominence to address this in-situ data gap (e.g., Tanzer et al., 2019; Rose Eilenberg et al., 2020). As the name implies, these provide a less expensive alternative to traditional regulatory-grade air quality monitors (RGM). As a tradeoff to achieve this

lower cost, LCS suffer from greater measurement uncertainties, and thus, require extensive calibration and validation efforts to generate useable data (Giordano et al., 2021). LCS can also be deployed to new areas which do not have the infrastructure to support RGM. LCS provide the only currently feasible means of routine air quality assessment in many low-and-middle-income countries (Hodoli et al., 2023; McFarlane, Isevulambire, et al., 2021; Raheja et al., 2022).

Even so, the availability of local air quality data from in-situ RGM or LCS may not provide sufficient situational awareness to air quality managers. Other, more globally available data sources may be required. One important source of such global data is satellite remote sensing retrievals of atmospheric composition. These data are provided by a fleet of instruments operated by national aerospace agencies and the private sector. By providing in many cases globe-spanning monitoring of the chemical and physical properties of the atmosphere at increasingly fine spatial resolution, satellite data can fill many gaps in our understanding of the composition of the atmosphere. However, satellite remote sensing has some key limitations with respect to air quality applications. Typically, remote sensing estimates take account of the entire atmospheric column, rather than the surface-level concentrations which are most relevant to air quality and the associated health exposure risk. The relationship between surface and column quantities is dependent on many factors. Thus, while promising, certain expertise and domain knowledge is required to correctly interpret satellite data for air quality purposes, which may be a barrier to its routine use in many areas (Anenberg et al., 2020; Duncan et al., 2021; Holloway et al., 2021).

Other sources of global air quality information are atmospheric chemistry and transport models (CTM). These models seek to estimate the state of the atmosphere, including parameters relevant for air quality, based on mathematical representations of chemical and physical processes combined with input data related to boundary conditions, e.g., the estimated emissions of various pollutants into the atmosphere. These models produce spatially comprehensive datasets and have the potential to forecast future air quality. However, their estimates may be biased due to incomplete and/or outdated input information or by inadequate representation of some chemical or physical processes. For example, inadequate temporal resolution for emissions data, differing vertical representations between the model and observations, as well as boundary layer mixing were found to impact the ability of the GEOS-Chem model to represent diel variations in fine particulate matter ($PM_{2.5}$) over the United States (Y. Li et al., 2023). Constraining CTM with observations from satellites, RGM, LCS, or a combination thereof via data assimilation is a widely used approach to addressing these model shortcomings. Assimilation of satellite data is more typical for global-scale CTM (Bocquet et al., 2015; Kelp et al., 2023), while in-situ data assimilation is more typical for sub-city to national scale CTM (Lopez-Restrepo et al., 2021; Schneider et al., 2023; Hassani et al., 2023).

Data fusion is an approach for bringing together various data sources. In contrast to data assimilation, where observations are used to update the state of a model, data fusion combines multiple data sources to produce a new data product, distinct from the inputs. A typical niche filled by data fusion is "downscaling" of coarser-resolution regional or global CTM output to produce more locally applicable outputs (Diao et al., 2019). A myriad of approaches using different inputs and methodologies has been proposed. On a local scale, data fusion of a dispersion model and LCS data has supported hourly $PM_{10}$ mapping in Nantes, France (Gressent et al., 2020). Regionally, satellite information is commonly used to support data fusion approaches; fusion of satellite aerosol optical depth (AOD), land use information, and meteorological data with surface observations from RGM and LCS allowed for daily 1-km resolution estimation of $PM_{2.5}$ over

California, USA (Bi et al., 2020). Satellite AOD, RGM, and LCS data were similarly combined for $PM_{2.5}$ mapping over Taiwan (J. Li et al., 2020). Globally, data fusion approaches are used to create yearly, monthly, or daily average surface $PM_{2.5}$ and constituent estimates (van Donkelaar et al., 2015, 2021; Wei et al., 2023). These estimates support analysis of the global impacts of air quality (Murray et al., 2020). For forecasting applications, i.e., prediction of surface concentrations in advance, bias correction for an ensemble of CTM was performed using surface RGM observations in both urban and rural areas to improve hourly $PM_{2.5}$ forecasting over the USA (Zhang et al., 2020, 2022). CTM, satellite and RGM data are combined to improve hourly $NO_2$ forecasts at sub-city scale (Malings et al., 2021). Machine learning methods have also been used for bias-correction of global CTM to produce daily $PM_{2.5}$ forecasts at 1-km resolution for applications at sub-city scale (Keller et al., 2020; Duncan et al., 2021; Bi et al., 2022). These studies demonstrate the wide applicability and flexibility of data fusion to incorporate models with various observational datasets.

In contrast to deterministic methods, probabilistic estimates and forecasts for air quality may be better suited to the needs of air quality managers and policy makers. For example, in a decision-focused analysis of ozone forecasting based on public health protection, it was found that single deterministic forecasts may produce less robust results compared to the use of multiple forecasts or an ensemble of forecasts for guiding air quality decision-making (Balashov et al., 2017; Garner & Thompson, 2012). This was because the ensemble forecasts more readily allowed for choosing actions which would be robust under a range of outcomes, i.e., robust under uncertainty. For global data fusion estimates of monthly $PM_{2.5}$, uncertainty quantification also supports analyzing the impact of this uncertainty on global health and epidemiological assessments (van Donkelaar et al., 2021). Several recent efforts have aimed at the quantification of uncertainty in air quality estimation and forecasting. Most of these approaches make use of ensembles of deterministic models (Garaud & Mallet, 2011; Gilliam et al., 2015; Riccio & Chianese, 2024) or machine learning methods, e.g., using generative models to produce a simulated ensemble (Fanfarillo et al., 2019). Data fusion approaches making use of geostatistical methods, especially Gaussian process or kriging approaches, have inherent capabilities to constrain estimates and quantify uncertainties for air quality estimation and forecasting (Wang et al., 2021). Kriging is referred to as "objective analysis" or "optimum interpolation" in the early numerical weather prediction literature (Diggle, 2010, p. 8). A major barrier to the wider use of probabilistic forecasts in air quality applications has been the difficulty associated with the interpretation of probabilistic forecasts by decision-makers and effectively communicating these to the public. Recent work has aimed at addressing these issues by explicitly analyzing different interpretation strategies corresponding with different desired outcomes (Balashov et al., 2023).

This paper presents a framework for combining CTM output, satellite remote sensing data, and in-situ measurements from a combination of RGM and LCS via a data fusion approach to support air quality estimation and/or forecasting. This framework includes explicit quantification of uncertainties associated with outputs from each stage, i.e., as each additional dataset is added. This paper aims at presenting a simple, generalizable method for data fusion with uncertainty quantification which can be implemented for near-real-time applications, with more limited computational requirements than a full data assimilation approach. We demonstrate this framework with a case study, focusing on estimation and forecasting of nitrogen dioxide in two US cities (San Francisco and New York City) in 2019. Nitrogen dioxide ($NO_2$), a regulated pollutant in the US (US EPA, 2017), represents a useful test case since it is known to vary on fine spatial scales in urban areas, which may not be captured even in high-resolution satellite datasets (e.g., Judd et al.,

2019). The ability to characterize this variability is an informative illustration of the capabilities of the proposed framework. The development of analysis tools and data products which combine multiple sources of air quality information, alongside methods to express confidence in or quantification of uncertainties in these products, has been suggested as a key need of air quality managers worldwide (Duncan et al., 2021). The methods presented in this paper are being implemented as part of a NASA-funded project to develop such tools for air quality data managers.

## 2 Methods

### 2.1 Input datasets

The proposed data fusion approach makes use of three categories of input information: CTM-based estimates and forecasts, satellite remote sensing data, and ground monitor data.

The NASA Global Earth Observing System Composition Forecast (GEOS-CF) system generates CTM outputs used in this paper. GEOS-CF couples the GEOS atmospheric general circulation model with the GEOS-Chem chemistry module (Keller et al., 2021). GEOS-CF produces 5-day forecasts initialized every day, following a 24-hour historical simulation for the previous day with the meteorology constrained by assimilated fields, to provide the best estimates for the past atmospheric composition. Both forecast and historical model output are used here. Hourly-average "surface-level" (average for the GEOS model's lowest level, nominally 130 m thick) nitrogen dioxide concentrations along with tropospheric column concentrations are used for the year 2019. GEOS-CF outputs are on a 0.25° or roughly 25 km latitude-longitude grid.

The TROPOMI instrument on the Sentinel 5P satellite provides retrievals related to tropospheric column concentrations of $NO_2$ (Veefkind et al., 2012). Through an agreement with the European Space Agency, TROPOMI data are also hosted at the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), searchable via the Common Metadata Repository system; these systems were used to identify and access relevant TROPOMI datasets. Tropospheric $NO_2$ concentration data products are used here, with recommended data quality filters for "good quality" retrievals. The latest high-resolution data product with a nominal pixel size of 5.5 by 3.5 km is used.

This paper presents a case study focused on San Francisco, California, USA (defined as between 37° N and 39° N and between 121° W and 123° W). Data for the month of September 2019 were used for the primary analysis; additional data from calendar year 2019 were also included as potential inputs for calibration purposes and for additional analysis presented in Section 3.3. An additional case study focused on New York City, New York, USA is also presented in the supplemental materials, described in supplemental text S1. These locations were selected due to their relatively high density of RGM for $NO_2$, as well as for comparability with previous related work (Malings et al., 2021). Ground monitoring data for hourly $NO_2$ were obtained from the US EPA's RGM network. Relevant data were queried using the Air Quality System API.

### 2.2 Data fusion approach and uncertainty quantification

The method for air quality data fusion outlined here is adapted from prior work (Malings et al., 2021). The major improvements presented here include (1) a generalization of the methodology and notation, where relevant changes to corresponding elements of the prior work will be noted, and (2) development of a framework for quantifying the uncertainty in fused estimates of surface air quality, which was not present in the prior work. The method is separated

into four phases: phase 1 involves model-based historical estimates and forecasts only; phase 2 fuses satellite with model data; phase 3 integrates in-situ measurements in an "offline" manner, useful mainly for bias correction; phase 4 integrates in-situ measurements in an "online" manner, useful for near-term estimate and forecast updating.

### *2.2.1 Phase 1: model-based estimation and uncertainty*

This data fusion approach starts with air quality estimate and forecast model outputs. Let $M(x,t,\tau)$ denote the estimated surface concentration of a given pollutant applicable at location $x$ and time $t$ produced by an air quality model (the GEOS-CF model in the current work). The forecasting lead-time is denoted by $\tau$. If target time $t$ is in the future, lead-time $\tau$ will be the difference between $t$ and when the model forecast was initialized. If $t$ is in the past, then $\tau = 0$, and the latest available model output covering time $t$ is used. Lead-time $\tau$ may not always be explicitly noted for notational convenience; when it is omitted, assume $\tau = 0$. The phase 1 estimate is simply the relevant model output:

$$E_1(x,t,\tau) = M(x,t,\tau). \quad (1)$$

Practically, it is important to note that while $x$ represents a location on the Earth's surface to arbitrary precision, the spatial resolution on which $E_1$ will be defined is limited to the spatial resolution of the model. In future work, it is considered that an ensemble of air quality models, either from different modeling systems or multiple initializations of the same model system, may be used to inform the data fusion. In that case, $E_1(x,t,\tau)$ could be the mean of multiple available models. Furthermore, the ensemble spread could be used for uncertainty quantification.

To better inform end-users on the uncertainty in data fusion estimates, we also aim to quantify the uncertainty of $E_1(x,t,\tau)$ in terms of the expected mean square error of the estimate with respect to the true concentration. We denote this uncertainty as $V_1(x,t,\tau)$. We estimate this uncertainty as the sum of four components, where independence between the components is assumed. These components are the uncertainty in the forecast due solely to its lead-time, $V_{F1}(x,t,\tau)$, the uncertainty due to local variability in the air quality model output, $V_M(x,t)$, the uncertainty due to potential bias in the air quality model, $V_{B1}(x,t)$, and the uncertainty due to the representational error of the model, $V_{R1}(x,t)$, due to its relatively coarse spatial resolution. Thus:

$$V_1(x,t,\tau) = V_{F1}(x,t,\tau) + V_M(x,t) + V_{B1}(x,t) + V_{R1}(x,t). \quad (2)$$

Model-based uncertainties $V_{F1}(x,t,\tau)$ and $V_M(x,t)$ are estimated empirically using model outputs. $V_{F1}(x,t,\tau)$ is estimated using the mean square difference of past model forecasts at lead-time $\tau$ and estimates at lead-time 0 for location $x$. This is evaluated over a set of times denoted $T_{c,t.o.d.}(t)$, representing times during a calibration period in the recent past, e.g., the prior week, at the same time-of-day (t.o.d.) as the time of interest $t$. This is meant to account for potential systematic differences in forecasting capabilities at different times of the day due to diel cycles or initialization times.

$$V_{F1}(x,t,\tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)}\left[\left(M(x,t',\tau) - M(x,t',0)\right)^2\right], \quad (3)$$

where $\mathbb{E}_i[\cdot]$ denotes the expected value, i.e., the mean, of the expression in brackets with respect to indexing parameter $i$. Note that $V_{F1}(x,t,0) = 0$ by design, and so this term can be ignored for $\tau = 0$.

$V_M(x,t)$ is estimated as the expected square difference of model outputs in the immediate vicinity of location $x$ and time $t$, i.e., the mean square difference of the model outputs in the grid cells immediately surrounding it in space and time:

$$V_M(x,t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[ \left( M(x',t') - M(x,t) \right)^2 \right], \quad (4)$$

where $X_n(x)$ represents the neighborhood of location $x$, i.e., its adjoining model grid cells depending on the model spatial resolution, and $T_n(t)$ represents the neighborhood of time $t$, i.e., the preceding and subsequent time steps according to the model temporal resolution. The logic behind this estimate is that, where model outputs are "smooth" in space and time, there is less uncertainty in the model outputs, while when the model outputs are more variable in space and time, there is greater uncertainty. This estimate depends on the model resolution, with lower uncertainties estimated for finer resolutions, all else being equal. We consider this to be reasonable, as finer resolution models will tend to explicitly represent processes at the relevant scale. However, simply interpolating model outputs to a finer resolution would artificially reduce the uncertainty estimate. This analysis should therefore be conducted at the native resolution of the model. A schematic for this phase is provided in Supplemental Figure S1.

The remaining terms $V_{B1}(x,t)$ and $V_{R1}(x,t)$ are impossible to assess using the model alone and must be estimated using external information, as will be discussed later (see Section 2.2.5). Note that, if an ensemble of models is used, it may be possible to estimate $V_{B1}(x,t)$ using the mean square differences between models in the ensemble (Riccio & Chianese, 2024). However, it may still be the case that all models within an ensemble are systematically biased due to some common underlying factor, e.g., all models using the same emissions dataset.

### *2.2.2 Phase 2: model downscaling with satellite data*

In phase 2, relationships between column concentrations from model and satellite data are used to inform the sub-model-grid variability of the pollutant of interest. The phase 2 estimate of the concentration of this pollutant at time $t$ and location $x$, $E_2(x,t,\tau)$, is the phase 1 estimate modified by the satellite-informed sub-grid difference pattern $D(x,t)$:

$$E_2(x,t,\tau) = E_1(x,t,\tau) + D(x,t), \quad (5)$$

where:

$$D(x,t) = \mathbb{E}_{t' \in T_{c,overpass}(t)} \left[ \left( S_{col}(x,t') - E_{1,col}(x,t') \right) \phi(x,t')\, \psi(x,t,t') \right]. \quad (6)$$

This difference pattern is the mean of the difference between the satellite-retrieved column concentration of the pollutant of interest, $S_{col}$, and the estimate of the same column quantity by the model used in phase 1, $E_{1,col}$, multiplied by two scaling factors $\phi$ and $\psi$. This mean is calculated during the calibration period associated with time of interest $t$ considering only times when the satellite was overhead, denoted $T_{c,overpass}(t)$. Practically, both $\phi$ and $\psi$ are informed by the model, which provides simulated data for all relevant surface and column quantities. Scaling factor $\phi(x,t)$ accounts for the change in surface concentration corresponding with a unit change in column concentration at location $x$ and time $t$. We approximate this sensitivity using a ratio of model values at this location and time:

$$\phi(x,t) \cong \frac{M(x,t,0)}{M_{col}(x,t,0)}. \quad (7)$$

Scaling factor $\psi(x,t,t')$ accounts for the ratio of changes in surface concentrations at location $x$ and time $t$ to changes at location $x$ and time $t'$. Again, we approximate this with a ratio of model values:

$$\psi(x,t,t') \cong \frac{M(x,t,0)}{M(x,t',0)}. \quad (8)$$

The definition of $D(x,t)$ presented in equation 6 is a generalization of "typical pattern" extraction described in equations 1 and 2 of Malings et al. (2021). This generalization now explicitly captures the relationship between surface concentrations and column quantities, which was only implicit before. Equation 5 here then replaces equation 3 of Malings et al. (2021). A schematic for this phase is provided in Supplemental Figure S2.

In general, it may be necessary to consider the observational operator and air mass factor used in the satellite retrieval algorithm, as these affect the comparability between satellite retrieved $S_{col}$ and modeled $E_{1,col}$ (e.g., Cooper et al., 2020). No explicit consideration of this is made here; instead, this will contribute to variability as discussed below. Future work may explicitly consider these impacts, likely leading to a reduced uncertainty. Note that in the case of $PM_{2.5}$, AOD would be the column quantity considered.

Similar to phase 1, the uncertainty of the phase 2 estimate, $V_2(x,t,\tau)$, is estimated as the sum of the uncertainty due to forecast lead-time, $V_{F2}(x,t,\tau)$, the local variability of the model, $V_M(x,t)$, the variance in the satellite-informed sub-grid difference pattern, $V_D(x,t)$, twice the co-variance of the model and sub-grid difference pattern, $V_{MD}(x,t)$, the uncertainty due to the potential bias in the model-and-satellite-derived surface concentration estimates, $\mathrm{V}_{B2}(x,t)$, and the uncertainty due to the representational error of the model-and-satellite-derived surface concentration estimates, $V_{R2}(x,t)$:

$$V_2(x,t,\tau) = V_{F2}(x,t,\tau) + V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t) + \mathrm{V}_{B2}(x,t) + V_{R2}(x,t). \quad (9)$$

Model local variability $V_M(x,t)$ is carried from phase 1, and as in phase 1, $V_{F2}(x,t,\tau)$ can be empirically estimated by examining the mean squared difference of forecasts with lead time $\tau$ over the calibration interval at the same time of day:

$$V_{F2}(x,t,\tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)} \left[ \left( E_2(x,t',\tau) - E_2(x,t',0) \right)^2 \right]. \quad (10)$$

$V_D(x,t)$ and $V_{MD}(x,t)$ can be estimated with the empirical variance and co-variance of relevant terms involved in computation of the satellite-informed sub-grid difference pattern:

$$V_D(x,t) \cong \mathbb{V}_{t' \in T_{c,overpass}(t)} \left[ \left( S_{col}(x,t') - E_{1,col}(x,t') \right) \phi(x,t') \, \psi(x,t,t') \right], \quad (11)$$

where $\mathbb{V}$ denotes a variance computation, and:

$$V_{MD}(x,t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[ \left( E_1(x',t') - E_1(x,t) \right) \left( D(x',t') - D(x,t) \right) \right]. \quad (12)$$

Note that in this formulation, $X_n(x)$ now denotes the neighboring locations of $x$ at the (finer) spatial resolution of the satellite data, i.e., the adjoining pixel centroids. The final terms related to bias $\mathrm{V}_{B2}(x,t)$ and representational errors $V_{R2}(x,t)$ again cannot be estimated using the model and satellite information alone and require surface-level information, as will be discussed later (see Section 2.2.5).

Comparing $V_1(x,t,\tau)$ with $V_2(x,t,\tau)$, and assuming a zero lead-time such that forecast-related uncertainty can be ignored, we can establish some constraints on the bias and representational error from phase 1 using phase 2 results. Due to the inclusion of satellite data in phase 2 compared to phase 1, we might assume that $V_2(x,t,\tau)$ will be less than or equal to $V_1(x,t,\tau)$ generally. Thus:

$$V_{B1}(x,t) + V_{R1}(x,t) \geq V_D(x,t) + 2V_{MD}(x,t) + \mathrm{V}_{B2}(x,t) + V_{R2}(x,t). \quad (13)$$

That is, uncertainty due to bias and representativity errors in phase 1 should be larger than the analogous terms from phase 2 plus the variance and co-variance related to the satellite-informed sub-model-grid difference patterns. Note that the inclusion of satellite information is informing both sub-model-grid variability, which would tend to reduce (though not eliminate) representational errors captured in $V_{R1}(x,t)$, as well as bringing in real-world measurement data, which would tend to reduce (though not eliminate) model bias as represented in $V_{B1}(x,t)$. Using this relationship, estimates of the phase 1 uncertainty terms can be made based on the relevant phase 2 uncertainty terms, e.g., using the average of these terms within each model grid cell.

### *2.2.3 Phase 3: linear correction with reliable surface measurements*

Phase 3 uses in-situ measurement data to correct for possible regional systematic errors in the model-and-satellite-derived estimates of surface air quality from phase 2. As a simple case, a linear correction is assumed with slope $\alpha$ and intercept $\beta$:

$$E_3(x,t,\tau) = \alpha\, E_2(x,t,\tau) + \beta. \quad (14)$$

This corresponds directly with equation 10 of Malings et al. (2021).

Coefficients $\alpha$ and $\beta$, as well as estimates of their variance $V_\alpha$ and $V_\beta$, co-variance $V_{\alpha\beta}$, and residual regression variance $V_{R3}$, are derived from a linear regression analysis between phase 2 estimates $E_2(x,t)$ as the independent variable and ground-based air quality measurements $G(x,t)$ as the dependent variable over the calibration time interval $T_c$ and the set of discrete surface monitoring sites in the region available during the calibration time interval $X_c$:

$$\alpha, \beta, V_\alpha, V_\beta, V_{\alpha\beta}, V_{R3} = \mathbb{LR}_{t' \in T_c(t), x' \in X_c(x)}[G(x',t') \sim E_2(x',t',0)], \quad (15)$$

where $\mathbb{LR}_{domain}[v_d \sim v_i]$ denotes a linear regression with independent variable $v_i$ and dependent variable $v_d$, conducted over a domain specified in the subscript of $\mathbb{LR}$. Since this regression is being applied for historical data collected during the calibration time interval, the phase 2 estimate with $\tau = 0$ is used, and so $\tau$ has been dropped here for notational convenience. Note that a weighted linear regression can be applied, e.g., using a weight factor related to the time-of-day as suggested in previous work (Malings et al., 2021, Section 3.5). In principle, other approaches to regression can also be applied, including for example machine learning techniques to account for non-linear relationships (e.g., as in Wei et al., 2023). In such a case, appropriate characterization of the variance of the regression estimates and their covariance with explanatory inputs would have to be performed. In this work, a linear regression approach is adopted as there are well known closed-form solutions for computing the variance and covariance of the parameters. A schematic for this phase is provided in Supplemental Figure S3.

In cases where both RGM and LCS provide in-situ data, a modified approach is recommended. First, available RGM are used in phase 3 as outlined above. Then, LCS are

regionally calibrated before incorporating their data in phase 4. Details are provided in supplemental text S2.

Uncertainty in the phase 3 estimate is based on the phase 2 estimated uncertainty, re-scaled with regression terms, and with the uncertainties in these regression terms and residual variance included:

$$V_3(x,t,\tau) = V_{F3}(x,t,\tau) + \alpha^2[V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)] + V_\alpha E_2(x,t)^2 + 2V_{\alpha\beta}E_2(x,t) + V_\beta + V_{R3}. \quad (16)$$

Now that in-situ data have been included, systematic bias due to the misrepresentation of the surface air quality due to model and satellite information only, as well as representational issues due to the limited spatial resolutions of the model and satellite data with respect to specific points represented in the surface data, are considered to be captured in terms related to regression coefficient variance and residual variance. However, practical limitations on the availability of surface air quality measurement sites, as well as the tendencies of such sites to be clustered in high-population-density areas, might mean that there are some residual biases which are not fully captured in this formulation. In other words, by necessity, the data fusion process will be tailored towards better representing locations where surface monitors already exist, and the above formulation for phase 3 uncertainty will tend to be more appropriate in those types of areas, rather than, e.g., more rural areas which are not covered by surface-based monitors. Furthermore, biases in the in-situ data will not be accounted for, e.g., the known sensitivity of $NO_2$ monitors to other species (e.g., Steinbacher et al., 2007).

Comparing the phase 2 and 3 variance estimates, assuming zero lead-time, and assuming that inclusion of surface information will tend to decrease phase 3 uncertainty with respect to phase 2, we can establish that:

$$\mathrm{V}_{B2}(x,t) + V_{R2}(x,t) \geq (\alpha^2 - 1)[V_M(x,t) + V_D(x,t) + 2V_{MD}(x,t)] + V_\alpha E_2(x,t)^2 + 2V_{\alpha\beta}E_2(x,t) + V_\beta + V_{R3}. \quad (17)$$

Note that we have now established a "chain" of relationships connecting various bias and representational error terms, which could not be directly quantified, to terms which can be empirically estimated based on the data fusion process. This gives us a basis for quantifying these uncertainties in earlier phases as well; this will be discussed further in Section 2.2.5.

### *2.2.4 Phase 4: updating with recent, nearby in-situ data*

Phase 4 enables the use of recent and nearby surface measurement data to provide updates to estimates and forecasts from phase 3 via a spatio-temporal kriging approach. This process is expressed as:

$$E_4(x,t,\tau) = E_3(x,t,\tau) + \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x,x',t,t')\,[G(x',t') - E_3(x',t')], \quad (18)$$

where $X_{near}(x)$ denotes surface measurement locations arbitrarily "nearby" to $x$, $T_{recent}(t)$ denotes times arbitrarily "recent" with respect to $t$, and $K(x,x',t,t')$ is the kriging update factor, encompassing the relationship between concentrations at spatio-temporal coordinates $x,t$ and $x',t'$. This relationship is a combination of variance and co-variance relationships between the locations as well as the measurement noise. $K(x,x',t,t')$ is evaluated with the assistance of a kernel function, used in Gaussian process regression to parameterize these co-variances based on,

e.g., the difference in space and time between the two sets of coordinates (Rasmussen & Williams, 2006). Recent work has proposed the use of Gaussian process regression for interpolating air quality data in space and/or time based on sparse measurements, and have proposed using square exponential, Matérn, and periodic kernel functions for this purpose for different pollutants of interest (Jang et al., 2020; Malings et al., 2021; Wang et al., 2021). The approach used here to determine appropriate kernel functions and parameters is described in (Malings et al., 2021, section 3.7). Equation 18 combines equations 11 and 14 of Malings et al. (2021), using a more generic notation of the kernel. A schematic for this phase is provided in Supplemental Figure S4.

Spatio-temporal kriging also quantifies the resulting uncertainty reduction:

$$V_4(x,t,\tau) = V_3(x,t,\tau) - \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x,x',t,t') \operatorname{cov}[E_3(x',t'), E_3(x,t)], \quad (19)$$

where $\operatorname{cov}[E_3(x',t'), E_3(x,t)]$ denotes the covariance between surface concentrations of the pollutant of interest between spatio-temporal coordinates $x, t$ and $x', t'$, which is again evaluated using the kernel function.

For practical purposes, appropriate definitions for $X_{near}(x)$ and $T_{recent}(t)$ will have to be chosen to balance accuracy with the computational intensiveness of considering many measurements in this updating, which is a typical limitation of Gaussian process regression. In this paper, we use all surface measurement locations in our application region but use only the most recent measurement from each location.

### *2.2.5 Quantifying uncertainties in phases 1 and 2*

Following phases 1 and 2 of the data fusion approach outlined above, there remain several terms related to potential bias and representativity errors which are not quantifiable given the inputs available at these phases. However, following phase 3, the inclusion of ground-based monitor data allowed the full quantification of uncertainty as expressed in equation (16). Using this fact, alongside the inequality relationships presented in equations (13) and (17), we conducted an empirical analysis comparing the quantified uncertainties at different phases. Based on this analysis, we propose the following parametric estimates for the unquantified portions of the uncertainties in phases 1 and 2:

$$V_{B1}(x,t) + V_{R1}(x,t) \cong \eta_1^2(t \bmod 24\text{h})\, \mathbb{E}_{t' \in T_c(t)}\, V_M(x,t'), \quad (20)$$

$$\mathrm{V}_{B2}(x,t) + V_{R2}(x,t) \cong \eta_2^2(t \bmod 24\text{h})\, \mathbb{E}_{t' \in T_c(t)}[V_M(x,t') + V_D(x,t') + 2V_{MD}(x,t')]. \quad (21)$$

In these estimates, the unquantified portions of the uncertainty are related to the quantified performance via empirically determined factors $\eta_1$ for phase 1 and $\eta_2$ for phase 2. These factors are assumed to vary as a function of time-of-day, based on observations for how relationships between different portions of the quantified uncertainty varied over the calibration period investigated here. Empirically determined values of $\eta_1$ and $\eta_2$ for San Francisco are presented Supplemental Figure S5; values for New York City are presented in Supplemental Figure S6.

This proposed approach has important limitations. Most notably, it relies on proceeding to phase 3 of the data fusion approach. In regions without ground-based monitoring, or where only a small number of ground-based monitors are available, the results from phase 3 of the data fusion approach will be unavailable or highly unreliable. Empirically determined values of $\eta_1$ and $\eta_2$

from another region might be used, but there is no reason to expect these to generalize well. Thus, in the absence of surface data, full uncertainty quantification in phase 1 or 2 of the data fusion approach becomes unreliable.

### 2.3 Confidence interval determination

Following the approaches for data fusion with uncertainty quantification presented in the previous section, for a location of interest $x$ and time of interest $t$, with forecast lead time $\tau$, and for data fusion phase $p$, a data fusion "best estimate" for the quantity of interest $E_p(x,t,\tau)$ will be available, along with an uncertainty estimate for this quantity, $V_p(x,t,\tau)$. To make practical use of these outputs, in this work, we use them to define confidence intervals (CI) for our estimates or forecasts. To do this, a probabilistic distribution must be assumed for the quantity of interest. In this work, we assume a lognormal distribution, which is a typical assumption for many non-negative quantities relevant to air quality. This distribution is parameterized by the mean $\mu$ and standard deviation $\sigma$ of the associated normal distribution. These are calculated from the outputs of the data fusion process as follows:

$$\mu_p(x,t,\tau) = \log\left[\frac{E_p(x,t,\tau)}{\sqrt{1+\frac{V_p(x,t,\tau)}{E_p(x,t,\tau)^2}}}\right], \tag{22}$$

$$\sigma_p(x,t,\tau) = \sqrt{\log\left[1+\frac{V_p(x,t,\tau)}{E_p(x,t,\tau)^2}\right]}. \tag{23}$$

The quantity of interest $F_p(x,t,\tau)$ is then a lognormally distributed random variable:

$$F_p(x,t,\tau) \sim \mathrm{LN}\left(\mu_p(x,t,\tau), \sigma_p(x,t,\tau)\right). \tag{24}$$

where $\mathrm{LN}(\mu,\sigma)$ denotes a lognormal distribution with mean $\mu$ and standard deviation $\sigma$ for the associated normal distribution. This distribution can be used to determine a CI for the quantity of interest. For example, the 75 % confidence range is defined with a lower bound, representing the 12.5$^{\text{th}}$ percentile of the lognormal distribution, and an upper bound, representing the 87.5$^{\text{th}}$ percentile of the lognormal distribution.

The lognormal distribution assumption is of course an approximation of the true distribution of the quantity of interest. Therefore, the CI determined as described above would not necessarily correspond to the actual CI for the quantity of interest, even if the mean and variance were known exactly. However, some assumption about the distribution of the quantity of interest is necessary, as its true distribution will not be known a priori.

## 3 Results

In this section, we investigate the performance of the proposed data fusion framework described above through testing with actual data. In all cases, a leave-one-site-out cross-validation approach is used. For the given domain of interest, data from all but one of the active ground monitoring sites are considered as inputs to the data fusion algorithm. Concentrations are estimated or forecast via the data fusion approach for the location of the single held-out site. All sites are cycled through in this manner, resulting in estimates and forecasts of concentrations at each monitoring site using data from all other sites. This allows for comparisons to be made between

actual concentration measurements at each site and the estimates or forecasts from the data fusion using all information except for any measurements at the site in question. This allows for evaluating how the method would perform at an arbitrary location without in-situ data. A 14-day moving calibration time window is used across all phases, i.e., for a given time of interest $t$ and forecast lead time $\tau$, the calibration interval $T_c$ ranges from $t - \tau - 14$ days to $t - \tau$. This ensures that only input data available at or before a given time are used, with lead time measured from the time of the most recently available data. However, data latency effects are not considered, e.g., satellite data are assumed to be available as soon as the satellite passes overhead. Data latency effects can be estimated by inflating the lead time, e.g., performance of a 1-day forecast using inputs with a 1-day data latency is assumed to be similar to a 2-day forecast.

For illustrative purposes, an example of time series output from the data fusion approach is presented in Figure 1. Outputs from phase 4 of the data fusion process, the colored line, including a 50 % CI, the colored area, are compared to actual measurements from the RGM at this location, the black line. In the figure, local midnight of September 17$^{th}$ is considered to be "the present" (marked by grey dotted vertical line). Before this time, estimates are shown considering zero lead time, i.e., GEOS-CF historical outputs are used together with satellite and RGM data available up to and including the indicated time. After midnight of September 17$^{th}$, forecasts are shown with increasing lead time, i.e., the latest GEOS-CF forecast initialized 12 UTC the previous day is used, together with satellite and RGM data collected prior to September 17$^{th}$. For the historical estimates, availability of in-situ measurements at other RGM sites has allowed short-term spikes to be better represented, with the CI likewise being wider to capture the variability. For the forecasts, such spikes are not specifically captured, but the CI tends to be wider throughout the timeseries, accounting for the potential for such spikes to occur. In this example, the estimated CI tend to be underconfident: 75 % of actual measurements fell within the 50 % CI depicted. An analysis of the accuracy and precision of the forecasts (not considering their confidence estimates) is presented in Supplemental Figure S7.

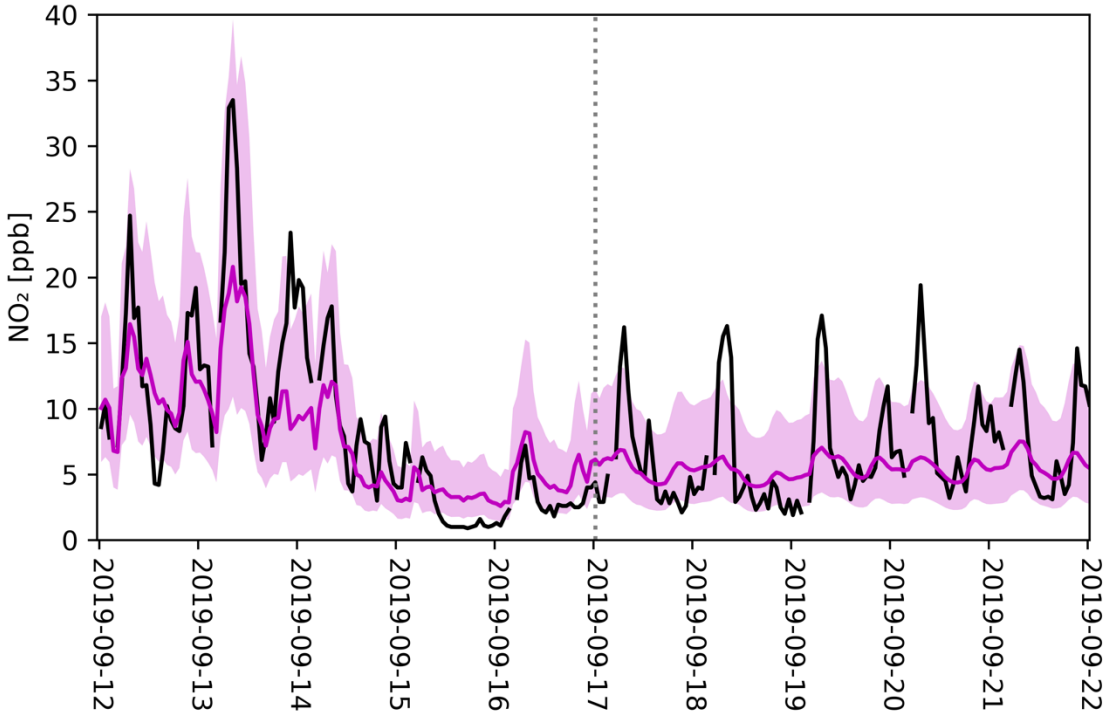



**Figure 1. Representative example of probabilistic estimates and forecasts for hourly surface-level $NO_2$ concentrations at the Redwood City monitor site (AQS ID 06-081-1001) in San Francisco, between September 12 and 22, 2019 local time. The black line indicates the reported concentrations from the regulatory monitor, i.e., the true concentration. The colored line indicates the mean estimated concentration from phase 4 of the data fusion process, $\boldsymbol{E_4(x,t)}$. The colored shaded areas denote the 50 % CI for the estimates. Estimates are presented with zero lead time up to midnight on September 17th, denoted with a vertical dotted line. Beyond this, forecasts with an increasing lead time are presented.**

### 3.1 Assessment of confidence interval coverage for different phases of data fusion

To investigate the accuracy of the assessed uncertainties in the data fusion, the fraction of actual measurements falling within the estimated 75 % CI across different phases of the data fusion approach is presented in Figure 2. This analysis considers all $NO_2$ monitor sites operating during September 2019 in the San Francisco study region, a total of 25 sites. The fraction of measurements falling within the 75 % CI is calculated for each site and considering the estimates for each phase of the data fusion process. Total uncertainties for phases 1 and 2 are estimated as outlined in section 2.2.5. Horizontal colored solid and dotted lines indicate the median, 25th percentile, and 75th percentile values of these fractions across all sites for each phase. Furthermore, sites are divided into types based on their assumed scale of spatial representativity, which is assessed for each monitoring site by US EPA. The five site types are microscale (0-0.1 km; 5 sites), middle (0.1-0.5 km; 3 sites), neighborhood (0.5-4 km; 13 sites), urban (4-50 km; 3 sites) and regional (50+ km; 1 site), as defined in 40 CFR Part 58. By investigating the capacity of the data fusion system to capture uncertainties at different spatial scales in this way, its benefits and limitations can be better understood.

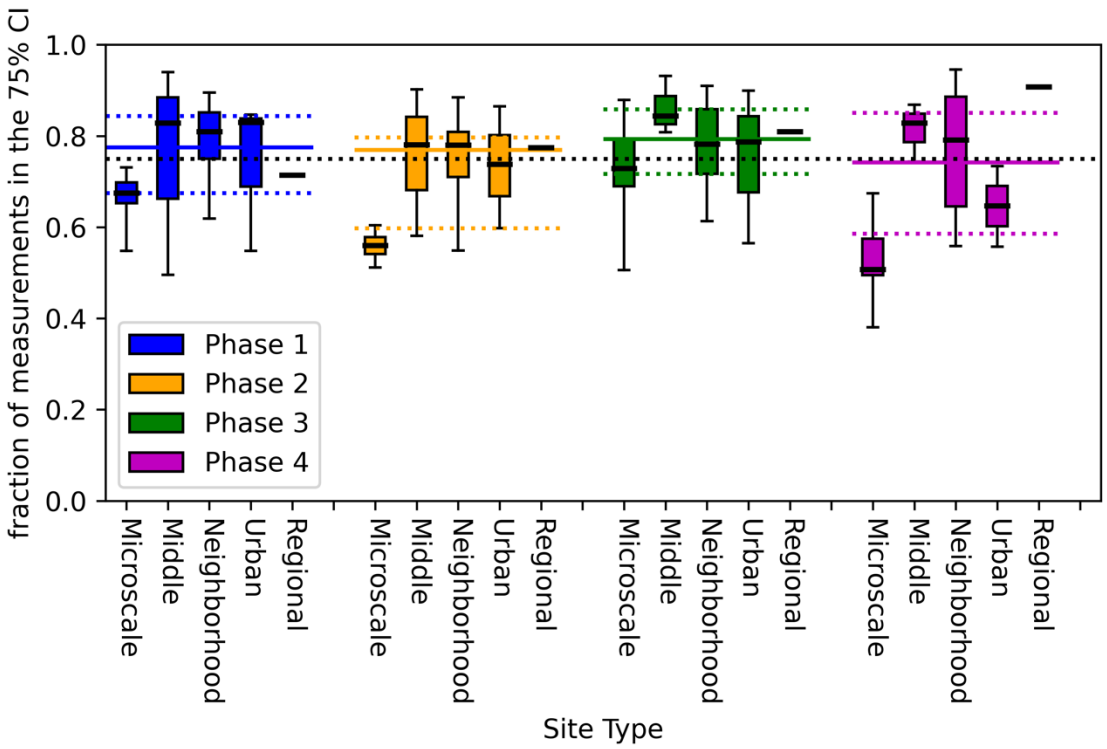



**Figure 2. Assessment of the fraction of actual measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors. The analysis represents data from 25 active $NO_2$ ground monitoring sites in the San Francisco study region for September 2019. A horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI. For each ground monitor site, the fraction of measurements at that site falling within the 75 % CI is calculated. For each phase, a solid horizonal line in the corresponding color indicates the median of these fractions across sites, and two horizontal dotted colored lines indicate the 25$^{th}$ percentile and 75$^{th}$ percentile of these fractions across sites. Furthermore, monitoring sites are divided into different site types. The spread in fraction of measurements falling within the 75 % CI for each site type is indicated with a box-and-whisker plot. In each box-and-whisker plot, the horizontal line inside the box denotes the median, the box denotes the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denote the full range.**

Overall, for all phases of the data fusion process, the estimated 75 % CI captures roughly 75 % of measured data. Performance is most consistent for phases 1 and 3, which have the smallest inter-quartile spreads in fraction of measurements falling within the 75 % CI. Focusing on phase 1, where only model outputs are considered, performance is consistent across most site types. There is a slight bias towards underconfidence, i.e., more measurements falling within the 75 % CI than expected. For microscale sites, however, estimates are systematically overconfident, with fewer measurements falling within the 75 % CI than expected. Considering the native spatial resolution of the model, better representation of uncertainties at urban and regional scales is to be expected. There is a lack of information at this stage to make informed assessments of confidence at finer spatial scales. This manifests in the results with a slightly larger spread in performance for middle scale sites and the overconfidence noted for microscale sites.

In phase 2, this is exacerbated, with increased overconfidence for estimates of microscale sites. Again, this can be explained by considering that, at phase 2, satellite data from TROPOMI with a nominal spatial resolution on the order of 5 km has been incorporated. This would be expected to improve assessments at neighborhood sites. This is reflected in the results with a slight decrease in the underconfidence of estimates for sites at this scale. However, there continues to be a lack of relevant information at finer spatial scales, and so while uncertainty estimates seem to have been improved for most scales, they have substantially degraded for microscale sites.

In phase 3, with the incorporation of ground-based data, uncertainties at microscale sites are now better represented overall, although one microscale site (denoted with the lower whisker) continues to be quite overconfidently estimated. However, middle scale sites are now being represented with systematic underconfidence. This might be a consequence of the relative numbers of sites in each type. There are 5 microscale and 3 middle scale sites in the study domain. Furthermore, because of the cross-validation approach, data from the site being evaluated are not included, underrepresenting that type. Thus, the approach of phase 3 would tend to better represent the more numerous site type. This could be accounted for by assigning lesser weights to certain types of sites when conducting the linear regression in phase 3. However, because one would not know a-priori the characteristics of the site at which concentrations are to be estimated, weighting different types of sites differently might not be an appropriate approach. Uncertainty estimates for neighborhood, urban, and regional sites appear reasonable, if slightly underconfident overall.

In phase 4, while uncertainty estimates seem to be most accurate in the median, the spread in performance has increased. Microscale sites are again exhibiting systematic overconfidence, along with urban scale sites, while middle scale and regional sites are underconfident. With only a single regional site, however, that latter result is not necessarily robust. This varied performance might be understood by considering that, due to the heterogeneity of urban areas, monitoring sites of different types will tend to be interspersed with one another. For a given site, the closest site which will have the greatest influence in the kriging approach of phase 4 is likely to be of a different type than the site being estimated for in the cross-validation. Neighborhood sites are least susceptible to this effect since, as the most numerous site type in the study area, the closest RGM to a neighborhood site is often another neighborhood scale site. The microscale sites, on the other hand, are closest to either neighborhood or urban scale sites, and the neighborhood or urban scale sites likewise are often closest to microscale sites. A kernel function for the kriging approach not based solely on distance might alleviate this difficulty, e.g., by defining similarities based on similar land use and land cover factors (e.g., Gilpin et al., 2023). Such an approach would require additional input information and is left as a subject for future improvements.

Across all phases, the best and most consistent results were observed for neighborhood scale sites. This is probably due in part to their relative abundance, but also to the fact that their representative scale (0.5-4 km) is of the same order as the satellite input data, which provides the most relevant information about spatial heterogeneity of pollutant concentrations. Overall, this is consistent with what might be expected, given the way in which the data fusion and associated uncertainty quantification are being conducted. Results were also similar for different CI (see Supplemental Figure S8).

### 3.2 Assessment of confidence interval coverage for different forecast lead times

Figure 3 presents an analysis of the fraction of measurements falling within the 75 % CI of the uncertainty estimate as a function of the forecasting lead time. Several discrete lead times are

considered, and results for zero lead time are also presented for comparison; these were previously presented in Figure 2.

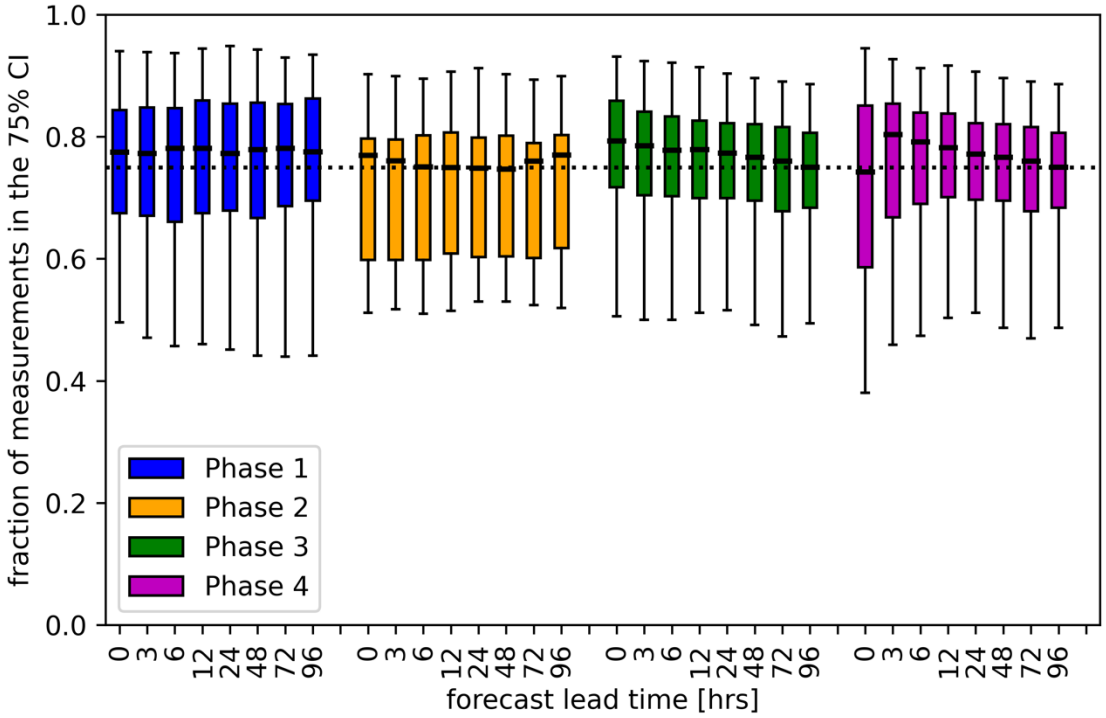



**Figure 3. Assessment of the fraction of actual measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, as a function of forecasting lead time, in hours. The analysis represents data from 25 active $NO_2$ ground monitoring sites in the San Francisco study region for September 2019. A horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI. For each ground monitor site, the fraction of measurements at that site falling within the 75 % CI is calculated. The box-and-whisker plots denote the ranges of these fractions across sites, with the horizontal line in the box denoting the median, the box denoting the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denoting the full range.**

Overall, there is little variation in the CI coverage as lead time increases, indicating that the uncertainty quantification approach is applicable for forecasts as well as historical estimates. For phase 3, there appears to be a tendency towards underconfidence at shorter lead times. For phase 4, the spread in coverage decreases as the forecasting lead time increases. As noted previously, the kriging approach of phase 4 with a distance-based kernel tends to induce under- or overconfidence at nearby sites. As the forecasting lead time increases, the influence of the most recent measurement data decreases, and the uncertainty quantification resembles that of phase 3. While the incorporation of near-real-time data in phase 4 has notable benefits in terms of near-term forecast accuracy, as noted in previous work (Malings et al., 2021), these results indicate that there is also a trade-off in terms of slightly less realistic uncertainty estimates in the phase 4 near-term forecasts compared to the other phases and to longer lead times.

### 3.3 Assessment of confidence interval coverage across different times of year

As an additional assessment, the methodology was applied across different months. Results for CI coverage at zero forecast lead time in March 2019, June 2019, September 2019 (as presented previously), and December 2019 are shown in Figure 4. There is some variability in performance for different phases in different months. For example, in December 2019, phases 1, 3, and 4 show a tendency for underconfidence in their estimates, although this is not apparent in phase 2. Conversely, phase 2 exhibits overconfidence in June 2019, while this is not apparent for other phases. This might indicate monthly or seasonally varying biases in the input data sources which are not accounted for in the current method.

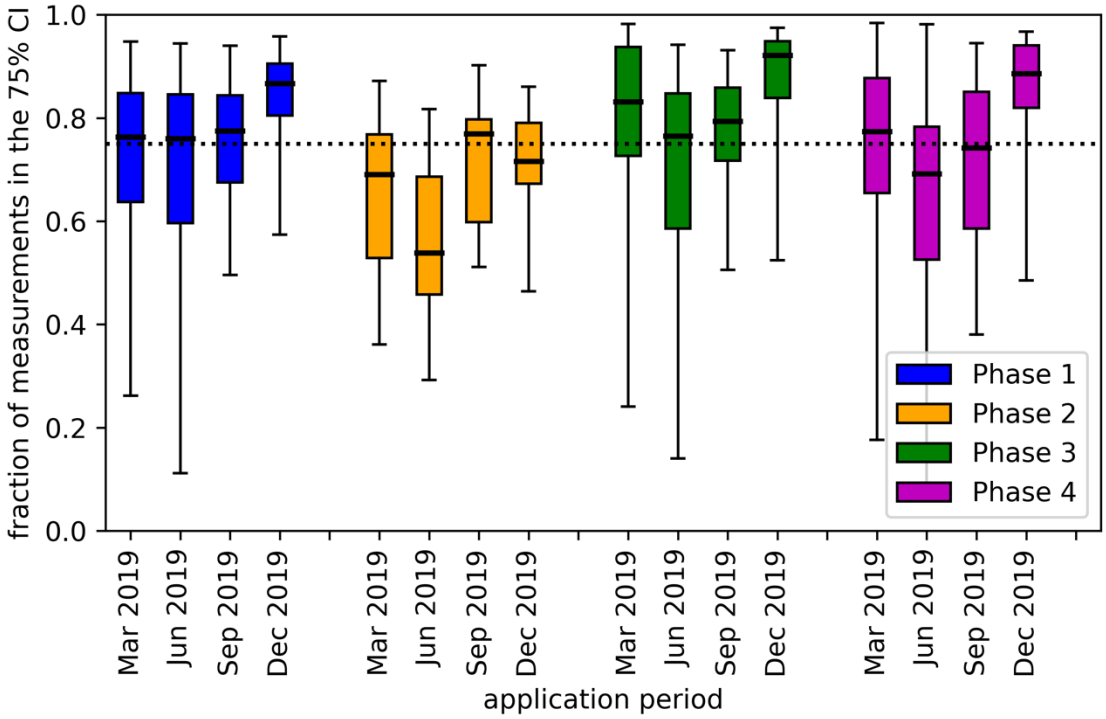



**Figure 4. Fractions of measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, presented for different application months. Box-and-whisker plots denote ranges of these fractions across active $NO_2$ monitor sites in San Francisco during that month, with the horizontal line in the box denoting the median, the box denoting the 25$^{th}$-to-75$^{th}$-percentile range, and the whiskers denoting the full range. The horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI.**

A similar assessment was conducted for the region of New York City, as discussed in the supplemental materials. Results for CI coverage at zero forecast lead time in March 2019, June 2019, September 2019, and December 2019 are shown in Supplemental Figure S9. Similar variability in performance for different phases in different months is observed as was noted above. Underconfidence in December 2019 seems to be more extreme, especially in phase 1, than in the case of San Francisco. Overconfidence in phase 2 also appears to be more severe. Again, monthly or seasonal differences in relevant parameters, especially the factors $\eta_1$ and $\eta_2$ calculated for the domain and kriging spatial and temporal scales associated with phase 4, might be influencing this.

The fact that month-to-month differences appear to be greater in New York City, where seasonal differences in prevailing meteorological conditions are relatively greater than in San Francisco, where such changes are relatively smaller, seems to corroborate this hypothesis. Thus, future development should focus on better capturing such seasonal changes through dynamically recalculating relevant parameters as part of the calibration process.

## 4 Conclusions

Overall, the proposed framework to estimate uncertainties and CI for concentration estimates from data fusion produced reasonable results in most cases, with most CI coverage being within about 10 percentage points of the theoretical value. There were also few instances of extreme overconfidence (few measurements falling within the prescribed CI) or extreme underconfidence (almost all measurements falling within the prescribed CI) observed in the results presented here. These findings are encouraging given the various assumptions made in defining the uncertainty quantification framework, including the assumption of lognormally distributed concentrations.

The uncertainty quantification was found to be least accurate overall for microscale sites, which are most impacted by hyperlocal sources. In the San Francisco case study, these sites were adjacent to highways, which are most heavily impacted by $NO_2$ pollution. This finding is useful to convey to any user of this system, i.e., that results may not be reliable within about 100 meters of a major source like a highway or other intense combustion activity. Similar limitations are likely, should the method be applied to other constituents measured near their respective sources.

It is also important to note that CI assessments are not being provided for independent data, but rather there is significant autocorrelation in the data. For example, while a measurement might have a 50 % chance of falling within a 50 % CI a-priori, if it is known that a recent measurement fell outside this CI, it becomes much less likely that a new measurement will fall within the CI. This effect can be noted on September 15$^{th}$ in Figure 1, when multiple measurements in sequence were observed outside the 50 % CI.

Several areas of theoretical and practical improvement are noted for future work. As suggested in Section 2.2.1, use of an ensemble of models rather than a single model in phase 1 would allow for estimating uncertainties at that phase based on variability across the ensemble. For incorporating satellite data in phase 2, multiple sources of satellite data might be considered, offering coverage at different times of day. Geostationary instruments like the recently launched TEMPO might be particularly useful in establishing different values of $D(x,t)$ corresponding to different times of day. Better definitions for the calibration dataset might also be explored, in contrast to a simple moving time window as presented in Section 2.2.2. For example, forecasted conditions might be matched to similar past conditions for which satellite data were available, in an attempt to identify past situations which approximately match forecasted future conditions in order to define a more suitable calibration dataset. There is also the possibility to include ancillary datasets, such as land use information, as additional co-variates to explain local variability. These might be incorporated using more sophisticated regression techniques, such as machine learning approaches, in contrast to the linear techniques presented for phase 3 in Section 2.2.3. While it would be necessary to develop customized uncertainty quantification schemes for these techniques, they might be better suited to capturing non-linear relationships in the data. Finally, the limitation of ground data availability and the resulting tendency of the approach to be biased towards such areas, as mentioned in Section 2.2.3, might be addressed in a more systematic way,

e.g., via resampling or application of different weightings to data from different types of monitoring sites in order to create a more unbiased calibration dataset.

Nevertheless, the framework established here presents a reasonable prior CI for the estimates and forecasts of the proposed data fusion system, and this fact supports effective and appropriate interpretation of its output by users. For example, these uncertainty estimates might be applied with respect to a given regulatory pollutant threshold to estimate the probability of exceeding that threshold. Such information could support air quality management decision-making. In an ongoing project supported by the NASA Health and Air Quality Applied Sciences Program, the authors are implementing the data fusion and uncertainty quantification scheme presented here in an online application via the Google Earth Engine platform. It is hoped that this application will present a useful tool for local air quality managers to visualize sub-city-scale atmospheric composition and variability using a combination of model, satellite, and in-situ data. This project is being conducted in collaboration with local environmental managers in the USA, Brazil, and Senegal. An example prototype for this tool is presented in Figure 5. As part of this project, the framework will also be extended to other relevant pollutants, primarily $PM_{2.5}$ and $O_3$.

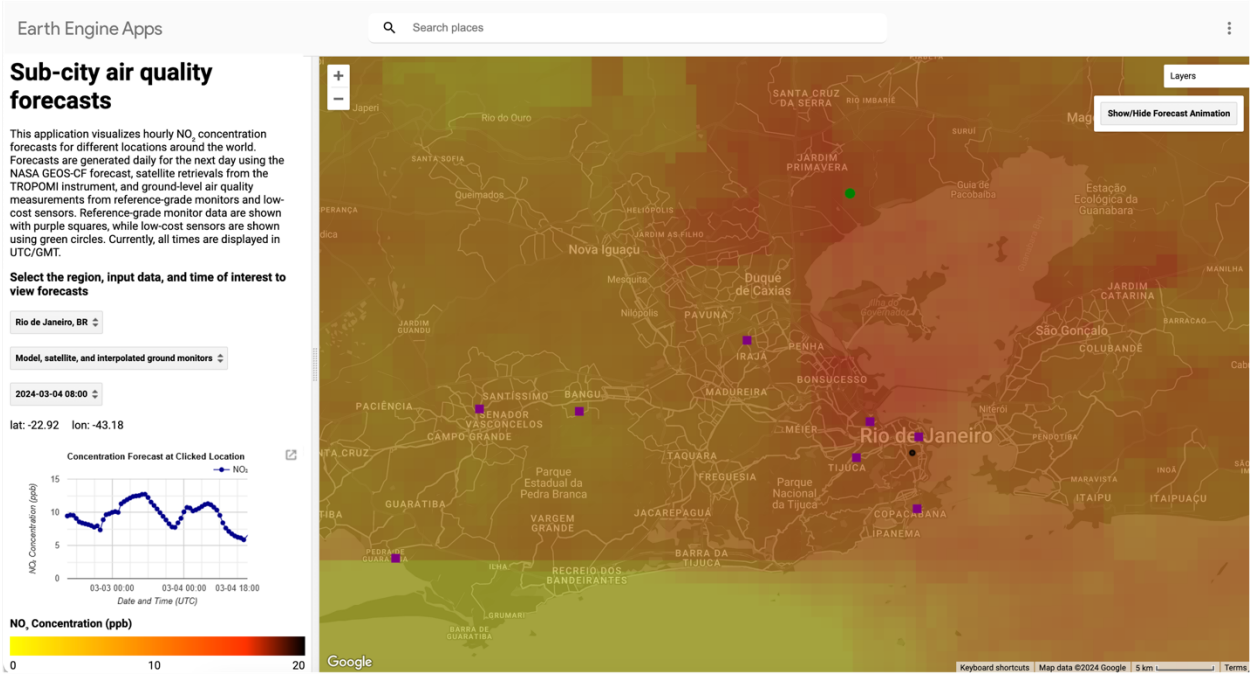



**Figure 5. Screenshot of an application currently under development which will implement the data fusion framework presented here, including uncertainty quantification, via the Google Earth Engine platform. This application will enable air quality managers to access and visualize estimates and forecasts of relevant air quality parameters such as $NO_2$, $O_3$, $PM_{2.5}$, along with associated expressions of confidence. Example outputs are presented for the city of Rio de Janeiro, Brazil, one of the partners for this project.**

## Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) under Grants 80NSSC22K1473 and WBS 389018.02.09.02.72 issued through the NASA Health and Air Quality Applied Sciences Program. The authors would also like to acknowledge the participation of Alan Chan, Sean Khan, John White, Daniel Westervelt, and Sean Wihera in

that grant project. The authors would like to thank Callum Wayman for consultations related to the implementation of the data fusion and uncertainty quantification scheme on the Google Earth Engine platform, Daniel King for software implementation to support the Google Earth Engine application, and Karin Tuxen-Bettman for guidance and assistance with ingesting necessary input datasets into Google Earth Engine. Finally, the authors would like to thank Felipe Mandarino, Bruno Boscaro, and Oswaldo Cruz for their comments and feedback during the development of the prototype depicted in Figure 5.

## Open Research

GEOS-CF outputs are available via the GMAO website; "AQC" and "XQC" collection files have been used here. Other input data are available via NASA GES DISC and the US EPA Air Quality System. Data and code used to generate the figures presented in this paper are available in an online Zenodo archive (Malings, 2024), governed under a CC BY-NC License.

## Author Contributions

Carl Malings: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft; K. Emma Knowland: Conceptualization, Supervision, Writing – Review & Editing; Nathan Pavlovic: Conceptualization, Writing – Review & Editing; Justin Coughlin: Software, Visualization, Writing – Review & Editing; Christoph Keller: Conceptualization; Stephen Cohn: Conceptualization, Methodology, Writing – Review & Editing; Randall Martin: Writing – Review & Editing.

## References

Anenberg, S. C., Bindl, M., Brauer, M., Castillo, J. J., Cavalieri, S., Duncan, B. N., Fiore, A. M., Fuller, R., Goldberg, D. L., Henze, D. K., Hess, J., Holloway, T., James, P., Jin, X., Kheirbek, I., Kinney, P. L., Liu, Y., Mohegh, A., Patz, J., … West, J. J. (2020). Using Satellites to Track Indicators of Global Air Pollution and Climate Change Impacts: Lessons Learned From a NASA-Supported Science-Stakeholder Collaborative. *GeoHealth*, *4*(7). https://doi.org/10.1029/2020GH000270

Balashov, N. V., Huff, A. K., & Thompson, A. M. (2023). Interpretation of Probabilistic Surface Ozone Forecasts: A Case Study for Philadelphia. *Weather and Forecasting*, *38*(10), 1895–1906. https://doi.org/10.1175/WAF-D-22-0185.1

Balashov, N. V., Thompson, A. M., & Young, G. S. (2017). Probabilistic Forecasting of Surface Ozone with a Novel Statistical Approach. *Journal of Applied Meteorology and Climatology*, *56*(2), 297–316. https://doi.org/10.1175/JAMC-D-16-0110.1

Bi, J., Knowland, K. E., Keller, C. A., & Liu, Y. (2022). Combining Machine Learning and Numerical Simulation for High-Resolution PM $_{2.5}$ Concentration Forecast. *Environmental Science & Technology*, *56*(3), 1544–1556. https://doi.org/10.1021/acs.est.1c05578

Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating Low-Cost Sensor Measurements into High-Resolution PM $_{2.5}$ Modeling at a Large Spatial Scale. *Environmental Science & Technology*, *54*(4), 2152–2162. https://doi.org/10.1021/acs.est.9b06046

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov,

A., Carnevale, C., Grell, G., & Seigneur, C. (2015). Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*, *15*(10), 5325–5358. https://doi.org/10.5194/acp-15-5325-2015

Cooper, M. J., Martin, R. V., Henze, D. K., & Jones, D. B. A. (2020). Effects of a priori profile shape assumptions on comparisons between satellite NO<sub>2</sub> columns and model simulations. *Atmospheric Chemistry and Physics*, *20*(12), 7231–7241. https://doi.org/10.5194/acp-20-7231-2020

Diao, M., Holloway, T., Choi, S., O'Neill, S. M., Al-Hamdan, M. Z., Van Donkelaar, A., Martin, R. V., Jin, X., Fiore, A. M., Henze, D. K., Lacey, F., Kinney, P. L., Freedman, F., Larkin, N. K., Zou, Y., Kelly, J. T., & Vaidyanathan, A. (2019). Methods, availability, and applications of PM $_{2.5}$ exposure estimates derived from ground measurements, satellite, and atmospheric models. *Journal of the Air & Waste Management Association*, *69*(12), 1391–1414. https://doi.org/10.1080/10962247.2019.1668498

Diggle, P. J. (2010). Historical Introduction. In A. E. Gelfand, M. Fuentes, P. Guttorp, & P. J. Diggle (Eds.), *Handbook of spatial statistics* (pp. 3–14). CRC Press.

Duncan, B. N., Malings, C. A., Knowland, K. E., Anderson, D. C., Prados, A. I., Keller, C. A., Cromar, K. R., Pawson, S., & Ensz, H. (2021). Augmenting the Standard Operating Procedures of Health and Air Quality Stakeholders With NASA Resources. *GeoHealth*, *5*(9). https://doi.org/10.1029/2021GH000451

Fanfarillo, A., Roozitalab, B., Hu, W., & Cervone, G. (2019). *Probabilistic Forecasting using Deep Generative Models*. https://doi.org/10.48550/ARXIV.1909.11865

Garaud, D., & Mallet, V. (2011). Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality. *Journal of Geophysical Research*, *116*(D19), D19304. https://doi.org/10.1029/2011JD015780

Garner, G. G., & Thompson, A. M. (2012). The Value of Air Quality Forecasting in the Mid-Atlantic Region. *Weather, Climate, and Society*, *4*(1), 69–79. https://doi.org/10.1175/WCAS-D-10-05010.1

Gilliam, R. C., Hogrefe, C., Godowitch, J. M., Napelenok, S., Mathur, R., & Rao, S. T. (2015). Impact of inherent meteorology uncertainty on air quality model predictions. *Journal of Geophysical Research: Atmospheres*, *120*(23). https://doi.org/10.1002/2015JD023674

Gilpin, S., Matsuo, T., & Cohn, S. E. (2023). A generalized, compactly supported correlation function for data assimilation applications. *Quarterly Journal of the Royal Meteorological Society*, *149*(754), 1953–1989. https://doi.org/10.1002/qj.4490

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., & Subramanian, R. (2021). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, *158*, 105833. https://doi.org/10.1016/j.jaerosci.2021.105833

Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environment International*, *143*, 105965. https://doi.org/10.1016/j.envint.2020.105965

788 Hassani, A., Schneider, P., Vogt, M., & Castell, N. (2023). Low-Cost Particulate Matter Sensors
789 for Monitoring Residential Wood Burning. *Environmental Science & Technology*, *57*(40), 15162–
790 15172. https://doi.org/10.1021/acs.est.3c03661

791 Hodoli, C. G., Coulon, F., & Mead, M. I. (2023). Source identification with high-temporal
792 resolution data from low-cost sensors using bivariate polar plots in urban areas of Ghana.
793 *Environmental Pollution*, *317*, 120448. https://doi.org/10.1016/j.envpol.2022.120448

794 Holloway, T., Miller, D., Anenberg, S., Diao, M., Duncan, B., Fiore, A. M., Henze, D. K., Hess, J.,
795 Kinney, P. L., Liu, Y., Neu, J. L., O'Neill, S. M., Odman, M. T., Pierce, R. B., Russell, A. G., Tong,
796 D., West, J. J., & Zondlo, M. A. (2021). Satellite Monitoring for Air Quality and Health. *Annual*
797 *Review of Biomedical Data Science*, *4*(1), 417–447. https://doi.org/10.1146/annurev-biodatasci-
798 110920-093120

799 Jang, J., Shin, S., Lee, H., & Moon, I.-C. (2020). Forecasting the Concentration of Particulate
800 Matter in the Seoul Metropolitan Area Using a Gaussian Process Model. *Sensors*, *20*(14), 3845.
801 https://doi.org/10.3390/s20143845

802 Judd, L. M., Al-Saadi, J. A., Janz, S. J., Kowalewski, M. G., Pierce, R. B., Szykman, J. J., Valin,
803 L. C., Swap, R., Cede, A., Mueller, M., Tiefengraber, M., Abuhassan, N., & Williams, D. (2019).
804 Evaluating the impact of spatial resolution on tropospheric NO2 column comparisons within urban
805 areas using high-resolution airborne data. *Atmospheric Measurement Techniques*, *12*(11), 6091–
806 6111. https://doi.org/10.5194/amt-12-6091-2019

807 Keller, C. A., Evans, M. J., Knowland, K. E., Hasenkopf, C. A., Modekurty, S., Lucchesi, R. A.,
808 Oda, T., Franca, B. B., Mandarino, F. C., Díaz Suárez, M. V., Ryan, R. G., Fakes, L. H., & Pawson,
809 S. (2020). *Global Impact of COVID-19 Restrictions on the Surface Concentrations of Nitrogen*
810 *Dioxide and Ozone* [Preprint]. Gases/Atmospheric Modelling/Troposphere/Chemistry (chemical
811 composition and reactions). https://doi.org/10.5194/acp-2020-685

812 Keller, C. A., Knowland, K. E., Duncan, B. N., Liu, J., Anderson, D. C., Das, S., Lucchesi, R. A.,
813 Lundgren, E. W., Nicely, J. M., Nielsen, E., Ott, L. E., Saunders, E., Strode, S. A., Wales, P. A.,
814 Jacob, D. J., & Pawson, S. (2021). Description of the NASA GEOS Composition Forecast
815 Modeling System GEOS-CF v1.0. *Journal of Advances in Modeling Earth Systems*, *13*(4).
816 https://doi.org/10.1029/2020MS002413

817 Kelp, M. M., Keller, C. A., Wargan, K., Karpowicz, B. M., & Jacob, D. J. (2023). Tropospheric
818 ozone data assimilation in the NASA GEOS Composition Forecast modeling system (GEOS-CF
819 v2.0) using satellite data for ozone vertical profiles (MLS), total ozone columns (OMI), and
820 thermal infrared radiances (AIRS, IASI). *Environmental Research Letters*, *18*(9), 094036.
821 https://doi.org/10.1088/1748-9326/acf0b7

822 Li, J., Zhang, H., Chao, C.-Y., Chien, C.-H., Wu, C.-Y., Luo, C. H., Chen, L.-J., & Biswas, P.
823 (2020). Integrating low-cost air quality sensor networks with fixed and satellite monitoring
824 systems to study ground-level PM2.5. *Atmospheric Environment*, *223*, 117293.
825 https://doi.org/10.1016/j.atmosenv.2020.117293

826 Li, Y., Martin, R. V., Li, C., Boys, B. L., van Donkelaar, A., Meng, J., & Pierce, J. R. (2023).
827 *Development and evaluation of processes affecting simulation of diel fine particulate matter*
828 *variation in the GEOS-Chem model* [Preprint]. Aerosols/Atmospheric Modelling and Data
829 Analysis/Troposphere/Chemistry (chemical composition and reactions).
830 https://doi.org/10.5194/egusphere-2023-704

Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O. L., Segers, A., & Heemink, A. W. (2021). Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá Valley, Colombia. *Atmosphere*, *12*(1), 91. https://doi.org/10.3390/atmos12010091

Malings, C. (2024). *Supporting Data for "Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results"* (1.0) [Python]. Zenodo. https://doi.org/10.5281/zenodo.10650853

Malings, C., Knowland, K. E., Keller, C. A., & Cohn, S. E. (2021). Sub-City Scale Hourly Air Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements. *Earth and Space Science*, *8*(7). https://doi.org/10.1029/2021EA001743

Martin, R. V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., & Dey, S. (2019). No one knows which city has the highest concentration of fine particulate matter. *Atmospheric Environment: X*, *3*, 100040. https://doi.org/10.1016/j.aeaoa.2019.100040

McFarlane, C., Isevulambire, P. K., Lumbuenamo, R. S., Ndinga, A. M. E., Dhammapala, R., Jin, X., McNeill, V. F., Malings, C., Subramanian, R., & Westervelt, D. M. (2021). First Measurements of Ambient PM2.5 in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of Congo Using Field-calibrated Low-cost Sensors. *Aerosol and Air Quality Research*, *21*. https://doi.org/10.4209/aaqr.200619

McFarlane, C., Raheja, G., Malings, C., Appoh, E. K. E., Hughes, A. F., & Westervelt, D. M. (2021). Application of Gaussian Mixture Regression for the Correction of Low Cost PM $_{2.5}$ Monitoring Data in Accra, Ghana. *ACS Earth and Space Chemistry*, acsearthspacechem.1c00217. https://doi.org/10.1021/acsearthspacechem.1c00217

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abegaz, K. H., Abolhassani, H., Aboyans, V., Abreu, L. G., Abrigo, M. R. M., Abualhasan, A., Abu-Raddad, L. J., Abushouk, A. I., Adabi, M., … Lim, S. S. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Raheja, G., Sabi, K., Sonla, H., Gbedjangni, E. K., McFarlane, C. M., Hodoli, C. G., & Westervelt, D. M. (2022). A Network of Field-Calibrated Low-Cost Sensor Measurements of $PM_{2.5}$ in Lomé, Togo, Over One to Two Years. *ACS Earth and Space Chemistry*, *6*(4), 1011–1021. https://doi.org/10.1021/acsearthspacechem.1c00391

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Riccio, A., & Chianese, E. (2024). Technical note: Accurate, reliable, and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method. *Atmospheric Chemistry and Physics*, *24*(3), 1673–1689. https://doi.org/10.5194/acp-24-1673-2024

Rose Eilenberg, S., Subramanian, R., Malings, C., Hauryliuk, A., Presto, A. A., & Robinson, A. L. (2020). Using a network of lower-cost monitors to identify the influence of modifiable factors driving spatial patterns in fine particulate matter concentrations in an urban environment. *Journal of Exposure Science & Environmental Epidemiology*. https://doi.org/10.1038/s41370-020-0255-x

Schneider, P., Vogt, M., Haugen, R., Hassani, A., Castell, N., Dauge, F. R., & Bartonova, A. (2023). Deployment and Evaluation of a Network of Open Low-Cost Air Quality Sensor Systems. *Atmosphere*, *14*(3), 540. https://doi.org/10.3390/atmos14030540

Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., & Hueglin, C. (2007). Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques. *Journal of Geophysical Research: Atmospheres*, *112*(D11), 2006JD007971. https://doi.org/10.1029/2006JD007971

Tanzer, R., Malings, C., Hauryliuk, A., Subramanian, R., & Presto, A. A. (2019). Demonstration of a Low-Cost Multi-Pollutant Network to Quantify Intra-Urban Spatial Variations in Air Pollutant Source Impacts and to Evaluate Environmental Justice. *International Journal of Environmental Research and Public Health*, *16*(14), 2523. https://doi.org/10.3390/ijerph16142523

US EPA. (2017). *Policy Assessment for the Review of the Primary National Ambient Air Quality Standards for Oxides of Nitrogen* (EPA-452/R-17-003). U.S. Environmental Protection Agency.

van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., & Martin, R. V. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology*, *55*(22), 15287–15300. https://doi.org/10.1021/acs.est.1c05309

van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2015). Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. *Environmental Health Perspectives*, *123*(2), 135–143. https://doi.org/10.1289/ehp.1408646

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., … Levelt, P. F. (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

Wang, P., Mihaylova, L., Chakraborty, R., Munir, S., Mayfield, M., Alam, K., Khokhar, M. F., Zheng, Z., Jiang, C., & Fang, H. (2021). A Gaussian Process Method with Uncertainty Quantification for Air Quality Monitoring. *Atmosphere*, *12*(10), 1344. https://doi.org/10.3390/atmos12101344

Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., & Zhu, T. (2023). First close insight into global daily gapless 1 km PM2.5 pollution, variability, and health impact. *Nature Communications*, *14*(1), 8349. https://doi.org/10.1038/s41467-023-43862-3

Zhang, H., Wang, J., García, L. C., Ge, C., Plessel, T., Szykman, J., Murphy, B., & Spero, T. L. (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

Zhang, H., Wang, J., García, L. C., Zhou, M., Ge, C., Plessel, T., Szykman, J., Levy, R. C., Murphy, B., & Spero, T. L. (2022). Improving Surface PM $_{2.5}$ Forecasts in the United States Using an Ensemble of Chemical Transport Model Outputs: 2. Bias Correction With Satellite Data for Rural

913 Areas. *Journal of Geophysical Research: Atmospheres*, *127*(1).
914 https://doi.org/10.1029/2021JD035563

915

*Journal of Geophysical Research: Machine Learning and Computation*

Supporting Information for

# Air Quality Estimation and Forecasting via Data Fusion with Uncertainty Quantification: Theoretical Framework and Preliminary Results

Carl Malings[1,2[0000-0002-2242-4328]], K. Emma Knowland[1,2[0000-0003-0837-8502]], Nathan Pavlovic[3[0000-0003-2127-3940]], Justin G. Coughlin[3[0000-0003-3882-3064]], Christoph Keller[1,2[0000-0002-0552-4298]], Stephen Cohn[2[0000-0001-8506-9354]], and Randall V. Martin[4[0000-0003-2632-8402]]

[1]Morgan State University, GESTAR II Cooperative Agreement, Baltimore, MD 21251, USA.

[2]NASA Goddard Space Flight Center Global Modeling & Assimilation Office, Greenbelt, MD 20771, USA.

[3]Sonoma Technology, Inc., Petaluma, CA 94954, USA.

[4]Washington University in St. Louis., St. Louis, MO 63130, USA.

## Contents of this file



## Introduction

This document provides supplemental supporting information for the manuscript indicated above. This includes a section (S1) detailing the handing of data from low-cost air quality sensors (LCS), as alluded to in Section 2.2.3. Additional results to supplement those presented in Section 3 are provided in Figure S5 through Figure S9. Diagrams of the various phases of the data fusion process are also illustrated in Figure S1 through Figure S4.

Note also that the data used to generate the results and figures presented here and are available in an online Zenodo archive (Malings, 2024), governed under a CC BY-NC License.

## Text S1. Details of the supplemental New York City case study example

For the supplemental study area of interest is the region surrounding New York City, New York, USA (defined as between 40°N and 42°N and between 73°W and 75°W). Data sources were the same as indicated in the paper for the San Francisco study area. Data from calendar year 2019 were included as potential inputs for calibration purposes.

## Text S2. Handling less reliable in-situ data from low-cost monitors

In the case of data from LCS, there are typically concerns associated with using the raw output data from these sensors. It is preferred that these data be calibrated to nearby RGM, with these calibrations usually being regionally specific, i.e., a single calibration approach is typically unsuitable beyond the region where it was developed (Giordano et al., 2021; McFarlane, Raheja, et al., 2021). Wherever possible, such regionally specific calibrations should be applied to LCS data before they are considered in this data fusion approach. However, due to the relative lack of RGM for conducting such calibration (a major motivation for data fusion approaches in the first place), such a local calibration may be lacking. In that case, the data fusion approach itself could be used to provide necessary data to conduct a crude regional calibration.

To address data from LCS with lower reliability and potentially large biases, we propose to apply a linear calibration approach, where data collected by LCS, $\mathbf{G}_{LCS}(x,t)$, provide the independent variable. The phase 3 estimates, $E_3(x,t)$, which include any RGM information in the area but not LCS information, provide the dependent variable. In regions lacking any RGM, the phase 2 estimate $E_2(x,t)$ may be used instead. As a vector quantity, $\mathbf{G}_{LCS}(x,t)$ may include important ancillary data such as temperature and humidity measurements, which are often important in calibrating LCS, together with measurements of the target pollutant. Regression is conducted considering a time interval $T_c$ and the set of discrete surface monitoring sites with LCS in the region $X_{LCS}$:

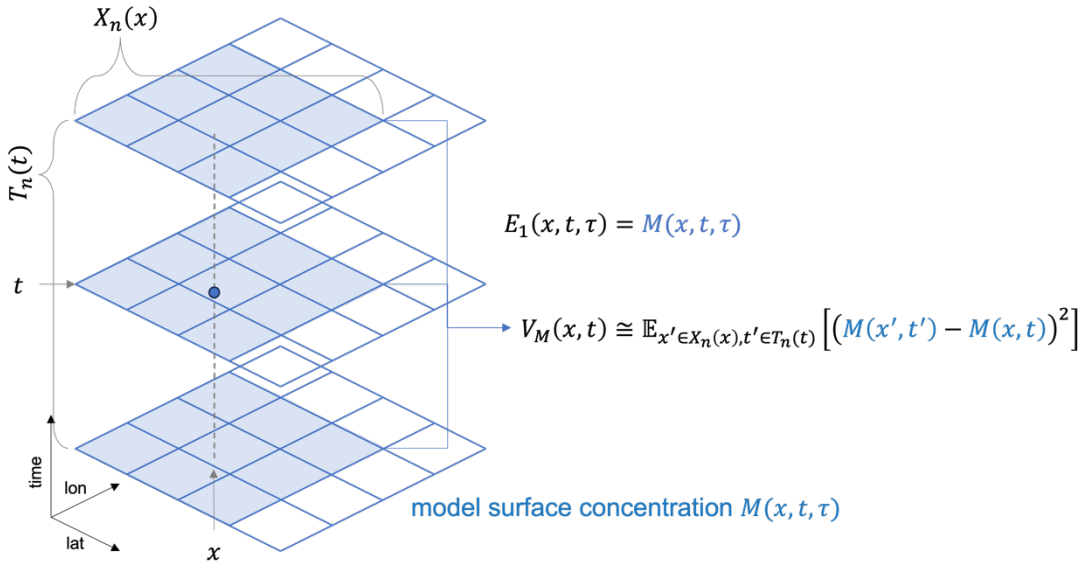
$$\boldsymbol{\zeta}, \xi, \mathbf{V}_{\boldsymbol{\zeta}}, V_{\xi}, \mathbf{V}_{\boldsymbol{\zeta\xi}}, V_{R,LCS} = \mathbb{LR}_{t' \in T_c(t), x' \in X_{LCS}}[E_3(x',t') \sim \mathbf{G}_{LCS}(x',t')]. \quad \text{(S1)}$$

The linear regression is then applied to the raw LCS data:

$$G_{LCS,calibrated}(x,t) = \boldsymbol{\zeta} \cdot \mathbf{G}_{LCS}(x,t) + \xi, \quad \text{(S2)}$$

where · denotes a dot product. The calibrated LCS data are then used in phase 4 to provide information for local updating of the estimates in their vicinities. In doing so, the relatively higher measurement uncertainties of these LCS should be considered when evaluating $K(x,x',t,t')$. These uncertainties can be quantified using the regression residual variance $V_{R,LCS}$. Note that since this calibration approach seeks to match, on a regional basis and for an extended calibration period, the LCS data to the phase 3 data fusion estimates, including these calibrated data back into the phase 3 estimation would be redundant. Once calibrated, however, individual LCS can provide valuable local and near-real-time information, and so including these data in phase 4 is potentially beneficial.

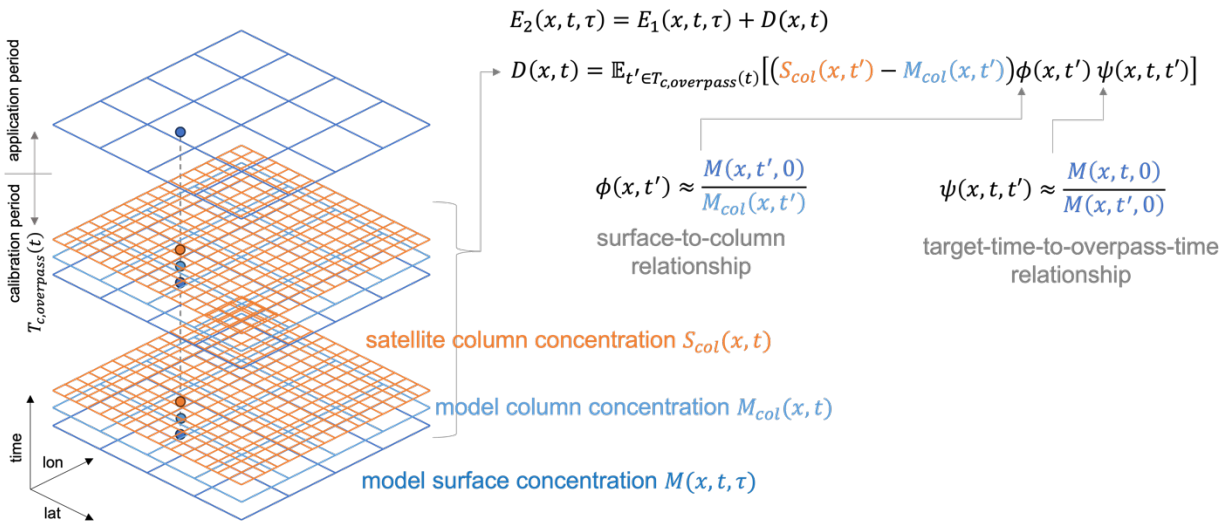
This approach is most suited to networks of LCS containing multiple devices with high inter-sensor precision and where the network is broadly distributed at a representative set of locations over the region of interest. In situations where inter-sensor precision is low, few LCS and no RGM are available, and/or where LCS deployments over-represent specific environments, especially near-source environments, this approach is likely to perform poorly.




985

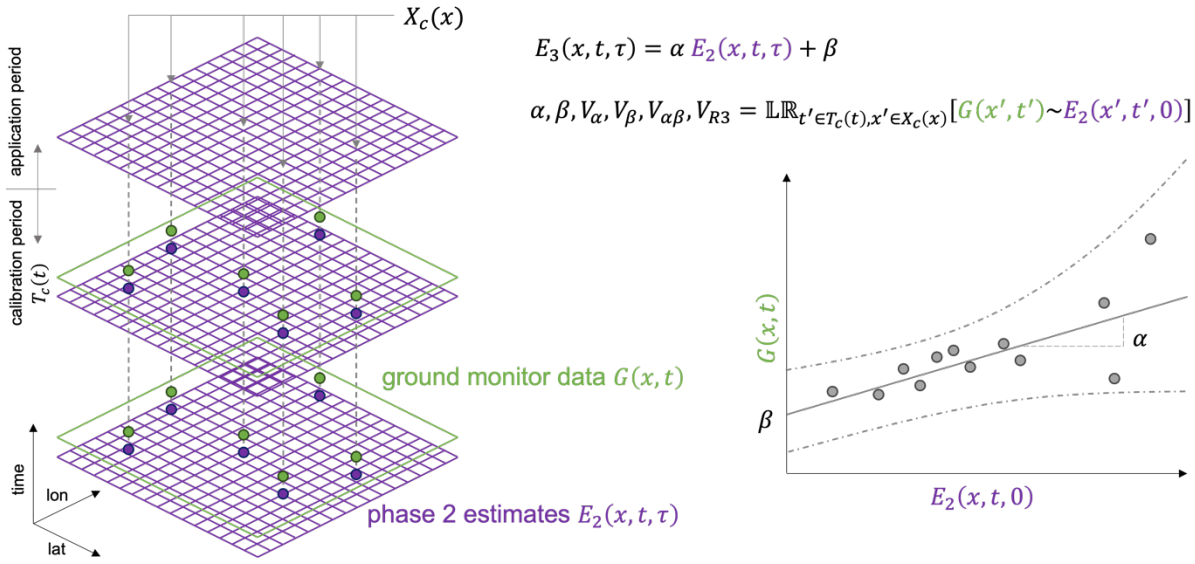
986 **Figure S1. Diagram of phase 1 of the data fusion process. Blue grids denote model grids in**
987 **space, with different layers denoting different timesteps. Shaded grids indicate the**
988 **neighborhood of the grid cell corresponding to location $x$ and time $t$, used for estimation**
989 **of model variability.**

990




991

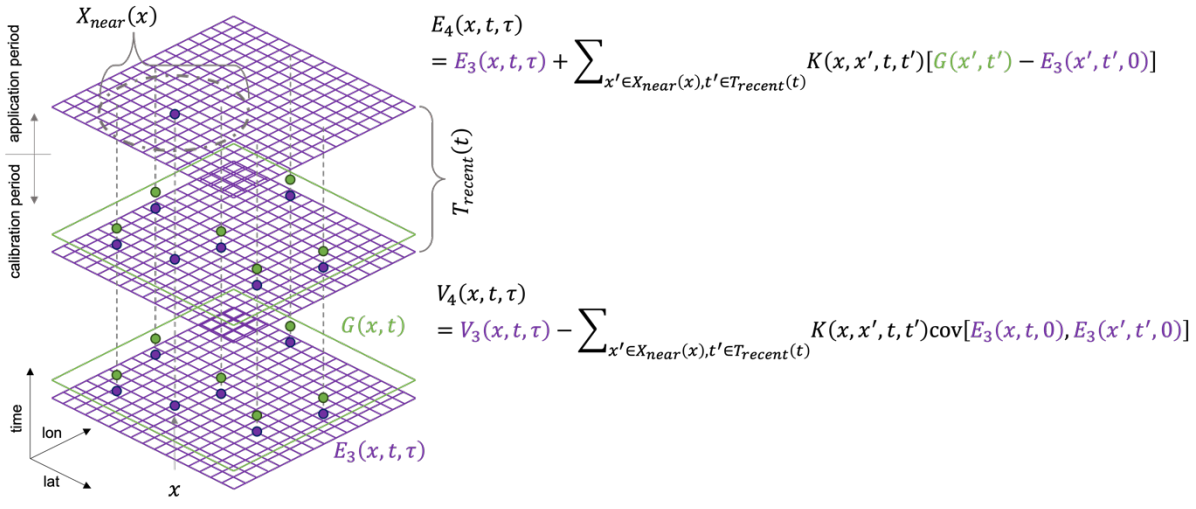
992 **Figure S2. Diagram of phase 2 of the data fusion process. Orange grids denote satellite**
993 **remote sensing data, with light blue grids corresponding to the analogous modeled column**
994 **quantity.**




995

996 **Figure S3. Diagram of phase 3 of the data fusion process. Purple grids correspond to the**
997 **phase 2 estimates. Green points indicate ground measurements at monitor sites $X_c(x)$**
998 **collected during calibration period $T_c(t)$. A conceptual illustration of the linear regression**
999 **process is provided on the right.**

1000




1001

1002 **Figure S4. Diagram of phase 4 of the data fusion process. The nearby region used for this**
1003 **phase, $X_{near}(x)$, is denoted with a grey ring. Recent times $T_{recent}(t)$ are considered to be the**
1004 **last timestep in the calibration period.**

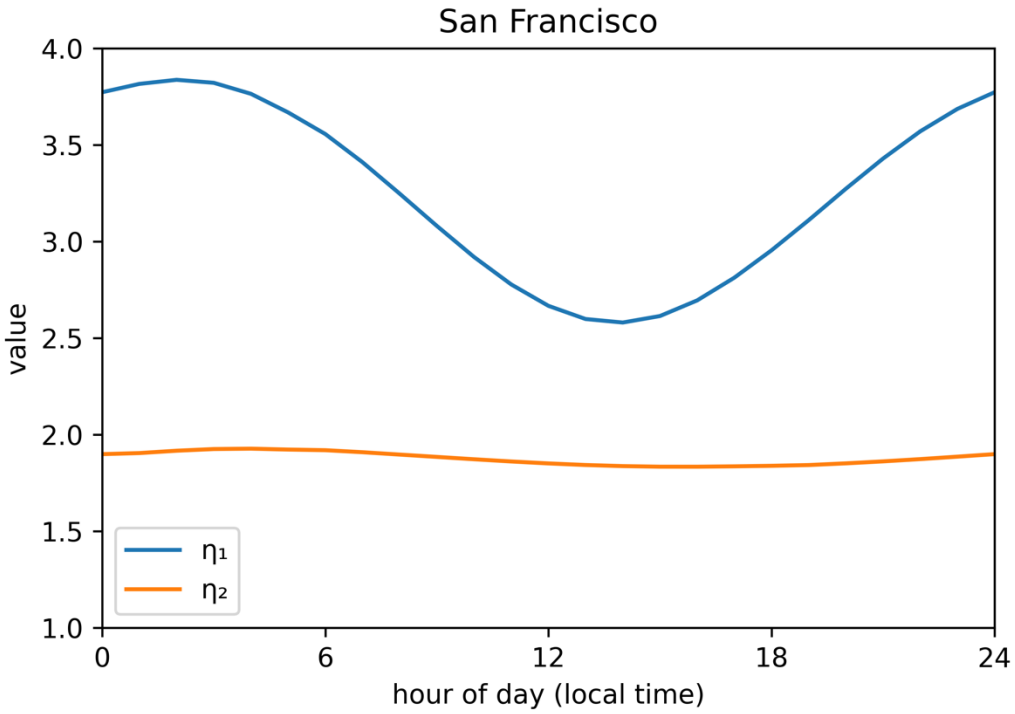



**Figure S5. Empirically determined values for $\eta_1$ and $\eta_2$ used for San Francisco in this paper, as a function of hour of the day (presented in local time).**

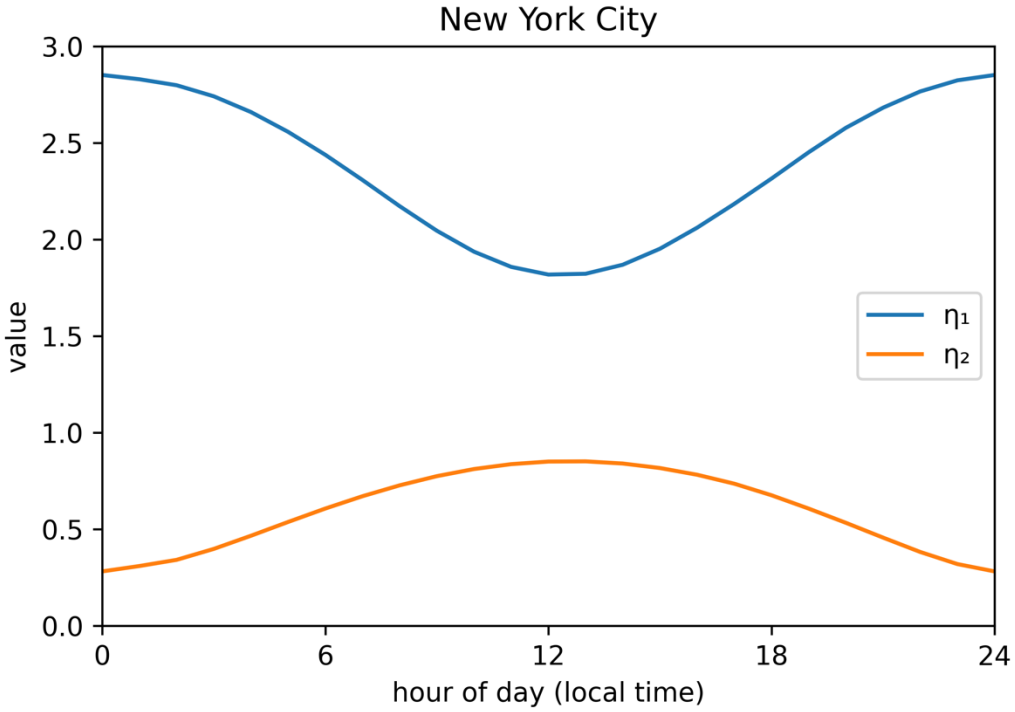



**Figure S6. Empirically determined values for $\eta_1$ and $\eta_2$ used for New York City in this paper, as a function of hour of the day (presented in local time).**

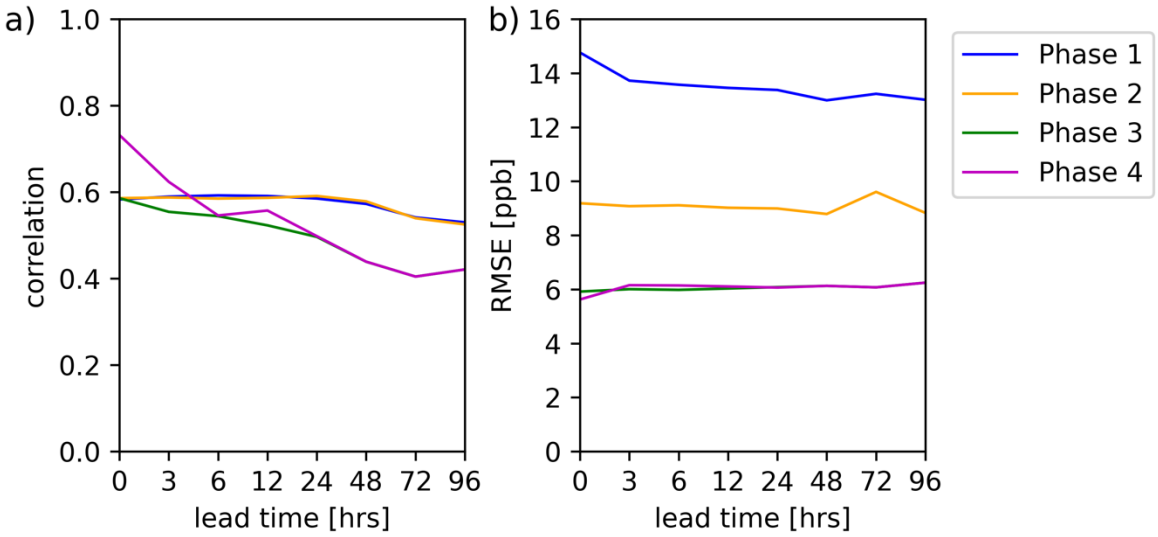



**Figure S7. Summary performance metrics for the data fusion approach, evaluated for the San Francisco study region in September 2019 (same results as presented in Figure 2). Plots depict the Pearson correlation (a) and root mean square error (b) between the estimates of the various data fusion phases (denoted by colors) as a function of the forecast lead time on the horizontal axis (note that the horizontal axis is not linearly scaled). The plotted values**

**depict the median value of the performance metrics assessed across the active monitor sites in the study region.**

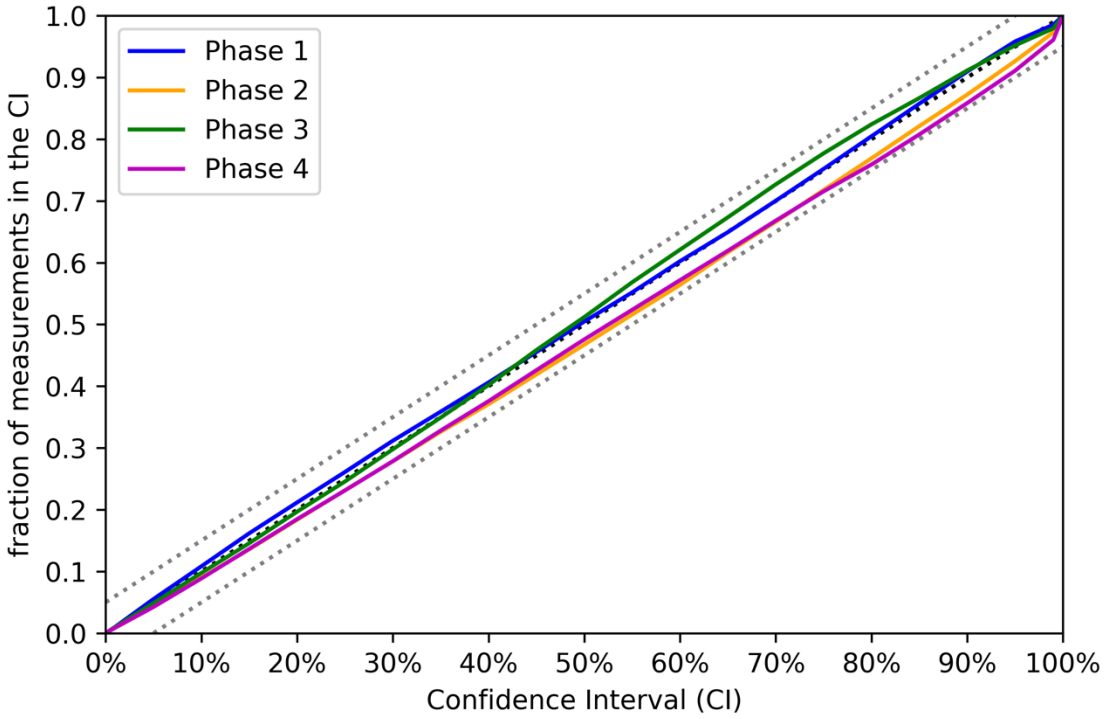



**Figure S8. Assessment of CI coverage for different CI. The horizontal axis reports the nominal coverage of the CI, and the vertical axis reports the actual fraction of measurements falling within that CI. The assessment was conducted for zero lead time estimates in the San Francisco study region for September 2019 (same results as presented in Figure 2). Coverage is assessed across all data simultaneously, i.e., the fraction of hourly measurements falling within the CI across all sites and all hours in the month is presented. Different colored lines represent different phases of the data fusion. The black dotted lines denote a one-to-one relationship (the ideal result), and grey dotted lines indicate results within 5 percentage points of this ideal.**

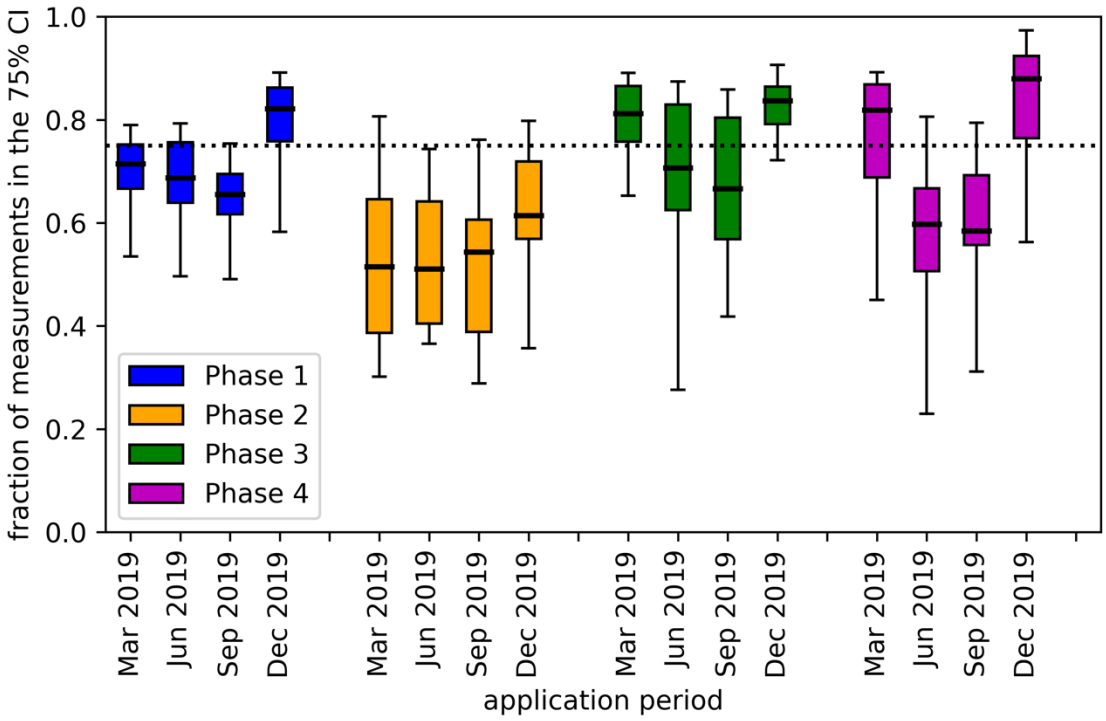



**Figure S9. Fractions of measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, presented for different application months. Box-and-whisker plots denote ranges of these fractions across active $NO_2$ monitor sites in New York City during that month, with the horizontal line in the box denoting the median, the box denoting the $25^{th}$-to-$75^{th}$-percentile range, and the whiskers denoting the full range. The horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling within the 75 % CI.**