

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

3 **Carl Malings**^{1,2[0000-0002-2242-4328]}, **K. Emma Knowland**^{1,2[0000-0003-0837-8502]}, **Nathan**
4 **Pavlovic**^{3[0000-0003-2127-3940]}, **Justin G. Coughlin**^{3[0000-0003-3882-3064]}, **Christoph Keller**^{1,2[0000-0002-}
5 ^{0552-4298]}, **Stephen Cohn**^{2[0000-0001-8506-9354]}, and **Randall V. Martin**^{4[0000-0003-2632-8402]}

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

7 ²NASA Goddard Space Flight Center, Global Modeling & Assimilation Office, Greenbelt, MD
8 20771, USA.

9 ³Sonoma Technology, Inc., Petaluma, CA 94954, USA.

10 ⁴Washington University in St. Louis., St. Louis, MO 63130, USA.

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

12 **Key Points:**

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

18 **Abstract**

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

32 **Plain Language Summary**

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

47 **1 Introduction**

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

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

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

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

96 Data fusion is an approach for bringing together various data sources. In contrast to data
97 assimilation, where observations are used to update the state of a model, data fusion combines
98 multiple data sources to produce a new data product, distinct from the inputs. A typical niche filled
99 by data fusion is “downscaling” of coarser-resolution regional or global CTM output to produce
100 more locally applicable outputs (Diao et al., 2019). A myriad of approaches using different inputs
101 and methodologies has been proposed. On a local scale, data fusion of a dispersion model and LCS
102 data has supported hourly PM₁₀ mapping in Nantes, France (Gressent et al., 2020). Regionally,
103 satellite information is commonly used to support data fusion approaches; fusion of satellite
104 aerosol optical depth (AOD), land use information, and meteorological data with surface
105 observations from RGM and LCS allowed for daily 1-km resolution estimation of PM_{2.5} over

106 California, USA (Bi et al., 2020). Satellite AOD, RGM, and LCS data were similarly combined
107 for PM_{2.5} mapping over Taiwan (J. Li et al., 2020). Globally, data fusion approaches are used to
108 create yearly, monthly, or daily average surface PM_{2.5} and constituent estimates (van Donkelaar et
109 al., 2015, 2021; Wei et al., 2023). These estimates support analysis of the global impacts of air
110 quality (Murray et al., 2020). For forecasting applications, i.e., prediction of surface concentrations
111 in advance, bias correction for an ensemble of CTM was performed using surface RGM
112 observations in both urban and rural areas to improve hourly PM_{2.5} forecasting over the USA
113 (Zhang et al., 2020, 2022). CTM, satellite and RGM data are combined to improve hourly NO₂
114 forecasts at sub-city scale (Malings et al., 2021). Machine learning methods have also been used
115 for bias-correction of global CTM to produce daily PM_{2.5} forecasts at 1-km resolution for
116 applications at sub-city scale (Keller et al., 2020; Duncan et al., 2021; Bi et al., 2022). These studies
117 demonstrate the wide applicability and flexibility of data fusion to incorporate models with various
118 observational datasets.

119 In contrast to deterministic methods, probabilistic estimates and forecasts for air quality
120 may be better suited to the needs of air quality managers and policy makers. For example, in a
121 decision-focused analysis of ozone forecasting based on public health protection, it was found that
122 single deterministic forecasts may produce less robust results compared to the use of multiple
123 forecasts or an ensemble of forecasts for guiding air quality decision-making (Balashov et al.,
124 2017; Garner & Thompson, 2012). This was because the ensemble forecasts more readily allowed
125 for choosing actions which would be robust under a range of outcomes, i.e., robust under
126 uncertainty. For global data fusion estimates of monthly PM_{2.5}, uncertainty quantification also
127 supports analyzing the impact of this uncertainty on global health and epidemiological assessments
128 (van Donkelaar et al., 2021). Several recent efforts have aimed at the quantification of uncertainty
129 in air quality estimation and forecasting. Most of these approaches make use of ensembles of
130 deterministic models (Garaud & Mallet, 2011; Gilliam et al., 2015; Riccio & Chianese, 2024) or
131 machine learning methods, e.g., using generative models to produce a simulated ensemble
132 (Fanfarillo et al., 2019). Data fusion approaches making use of geostatistical methods, especially
133 Gaussian process or kriging approaches, have inherent capabilities to constrain estimates and
134 quantify uncertainties for air quality estimation and forecasting (Wang et al., 2021). Kriging is
135 referred to as “objective analysis” or “optimum interpolation” in the early numerical weather
136 prediction literature (Diggle, 2010, p. 8). A major barrier to the wider use of probabilistic forecasts
137 in air quality applications has been the difficulty associated with the interpretation of probabilistic
138 forecasts by decision-makers and effectively communicating these to the public. Recent work has
139 aimed at addressing these issues by explicitly analyzing different interpretation strategies
140 corresponding with different desired outcomes (Balashov et al., 2023).

141 This paper presents a framework for combining CTM output, satellite remote sensing data,
142 and in-situ measurements from a combination of RGM and LCS via a data fusion approach to
143 support air quality estimation and/or forecasting. This framework includes explicit quantification
144 of uncertainties associated with outputs from each stage, i.e., as each additional dataset is added.
145 This paper aims at presenting a simple, generalizable method for data fusion with uncertainty
146 quantification which can be implemented for near-real-time applications, with more limited
147 computational requirements than a full data assimilation approach. We demonstrate this framework
148 with a case study, focusing on estimation and forecasting of nitrogen dioxide in two US cities (San
149 Francisco and New York City) in 2019. Nitrogen dioxide (NO₂), a regulated pollutant in the US
150 (US EPA, 2017), represents a useful test case since it is known to vary on fine spatial scales in
151 urban areas, which may not be captured even in high-resolution satellite datasets (e.g., Judd et al.,

152 2019). The ability to characterize this variability is an informative illustration of the capabilities
153 of the proposed framework. The development of analysis tools and data products which combine
154 multiple sources of air quality information, alongside methods to express confidence in or
155 quantification of uncertainties in these products, has been suggested as a key need of air quality
156 managers worldwide (Duncan et al., 2021). The methods presented in this paper are being
157 implemented as part of a NASA-funded project to develop such tools for air quality data managers.

158 **2 Methods**

159 2.1 Input datasets

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

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

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

179 This paper presents a case study focused on San Francisco, California, USA (defined as
180 between 37° N and 39° N and between 121° W and 123° W). Data for the month of September
181 2019 were used for the primary analysis; additional data from calendar year 2019 were also
182 included as potential inputs for calibration purposes and for additional analysis presented in
183 Section 3.3. An additional case study focused on New York City, New York, USA is also presented
184 in the supplemental materials, described in supplemental text S1. These locations were selected
185 due to their relatively high density of RGM for NO_2 , as well as for comparability with previous
186 related work (Malings et al., 2021). Ground monitoring data for hourly NO_2 were obtained from
187 the US EPA’s RGM network. Relevant data were queried using the [Air Quality System API](#).

188 2.2 Data fusion approach and uncertainty quantification

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

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

198 *2.2.1 Phase 1: model-based estimation and uncertainty*

199 This data fusion approach starts with air quality estimate and forecast model outputs. Let
 200 $M(x, t, \tau)$ denote the estimated surface concentration of a given pollutant applicable at location x
 201 and time t produced by an air quality model (the GEOS-CF model in the current work). The
 202 forecasting lead-time is denoted by τ . If target time t is in the future, lead-time τ will be the
 203 difference between t and when the model forecast was initialized. If t is in the past, then $\tau = 0$,
 204 and the latest available model output covering time t is used. Lead-time τ may not always be
 205 explicitly noted for notational convenience; when it is omitted, assume $\tau = 0$. The phase 1 estimate
 206 is simply the relevant model output:

$$207 \quad E_1(x, t, \tau) = M(x, t, \tau). \quad (1)$$

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

214 To better inform end-users on the uncertainty in data fusion estimates, we also aim to
 215 quantify the uncertainty of $E_1(x, t, \tau)$ in terms of the expected mean square error of the estimate
 216 with respect to the true concentration. We denote this uncertainty as $V_1(x, t, \tau)$. We estimate this
 217 uncertainty as the sum of four components, where independence between the components is
 218 assumed. These components are the uncertainty in the forecast due solely to its lead-time,
 219 $V_{F1}(x, t, \tau)$, the uncertainty due to local variability in the air quality model output, $V_M(x, t)$, the
 220 uncertainty due to potential bias in the air quality model, $V_{B1}(x, t)$, and the uncertainty due to the
 221 representational error of the model, $V_{R1}(x, t)$, due to its relatively coarse spatial resolution. Thus:

$$222 \quad V_1(x, t, \tau) = V_{F1}(x, t, \tau) + V_M(x, t) + V_{B1}(x, t) + V_{R1}(x, t). \quad (2)$$

223 Model-based uncertainties $V_{F1}(x, t, \tau)$ and $V_M(x, t)$ are estimated empirically using model
 224 outputs. $V_{F1}(x, t, \tau)$ is estimated using the mean square difference of past model forecasts at lead-
 225 time τ and estimates at lead-time 0 for location x . This is evaluated over a set of times denoted
 226 $T_{c,t.o.d.}(t)$, representing times during a calibration period in the recent past, e.g., the prior week, at
 227 the same time-of-day (t.o.d.) as the time of interest t . This is meant to account for potential
 228 systematic differences in forecasting capabilities at different times of the day due to diel cycles or
 229 initialization times.

$$230 \quad V_{F1}(x, t, \tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)} \left[\left(M(x, t', \tau) - M(x, t', 0) \right)^2 \right], \quad (3)$$

231 where $\mathbb{E}_i[\cdot]$ denotes the expected value, i.e., the mean, of the expression in brackets with respect
 232 to indexing parameter i . Note that $V_{F1}(x, t, 0) = 0$ by design, and so this term can be ignored for
 233 $\tau = 0$.

234 $V_M(x, t)$ is estimated as the expected square difference of model outputs in the immediate
 235 vicinity of location x and time t , i.e., the mean square difference of the model outputs in the grid
 236 cells immediately surrounding it in space and time:

$$237 \quad V_M(x, t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[(M(x', t') - M(x, t))^2 \right], \quad (4)$$

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

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

255 *2.2.2 Phase 2: model downscaling with satellite data*

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

$$260 \quad E_2(x, t, \tau) = E_1(x, t, \tau) + D(x, t), \quad (5)$$

261 where:

$$262 \quad D(x, t) = \mathbb{E}_{t' \in T_{c,overpass}(t)} \left[(S_{col}(x, t') - E_{1,col}(x, t')) \phi(x, t') \psi(x, t, t') \right]. \quad (6)$$

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

$$272 \quad \phi(x, t) \cong \frac{M(x, t, 0)}{M_{col}(x, t, 0)}. \quad (7)$$

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

$$276 \quad \psi(x, t, t') \cong \frac{M(x, t, 0)}{M(x, t', 0)}. \quad (8)$$

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

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

288 Similar to phase 1, the uncertainty of the phase 2 estimate, $V_2(x, t, \tau)$, is estimated as the
 289 sum of the uncertainty due to forecast lead-time, $V_{F2}(x, t, \tau)$, the local variability of the model,
 290 $V_M(x, t)$, the variance in the satellite-informed sub-grid difference pattern, $V_D(x, t)$, twice the co-
 291 variance of the model and sub-grid difference pattern, $V_{MD}(x, t)$, the uncertainty due to the
 292 potential bias in the model-and-satellite-derived surface concentration estimates, $V_{B2}(x, t)$, and the
 293 uncertainty due to the representational error of the model-and-satellite-derived surface
 294 concentration estimates, $V_{R2}(x, t)$:

$$295 \quad V_2(x, t, \tau) = V_{F2}(x, t, \tau) + V_M(x, t) + V_D(x, t) + 2V_{MD}(x, t) + V_{B2}(x, t) + V_{R2}(x, t). \quad (9)$$

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

$$299 \quad V_{F2}(x, t, \tau) \cong \mathbb{E}_{t' \in T_{c,t.o.d.}(t)} \left[\left(E_2(x, t', \tau) - E_2(x, t', 0) \right)^2 \right]. \quad (10)$$

300 $V_D(x, t)$ and $V_{MD}(x, t)$ can be estimated with the empirical variance and co-variance of
 301 relevant terms involved in computation of the satellite-informed sub-grid difference pattern:

$$302 \quad V_D(x, t) \cong \mathbb{V}_{t' \in T_{c,overpass}(t)} \left[\left(S_{col}(x, t') - E_{1,col}(x, t') \right) \phi(x, t') \psi(x, t, t') \right], \quad (11)$$

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

$$304 \quad V_{MD}(x, t) \cong \mathbb{E}_{x' \in X_n(x), t' \in T_n(t)} \left[\left(E_1(x', t') - E_1(x, t) \right) \left(D(x', t') - D(x, t) \right) \right]. \quad (12)$$

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

310 Comparing $V_1(x, t, \tau)$ with $V_2(x, t, \tau)$, and assuming a zero lead-time such that forecast-
 311 related uncertainty can be ignored, we can establish some constraints on the bias and
 312 representational error from phase 1 using phase 2 results. Due to the inclusion of satellite data in
 313 phase 2 compared to phase 1, we might assume that $V_2(x, t, \tau)$ will be less than or equal to
 314 $V_1(x, t, \tau)$ generally. Thus:

$$315 \quad V_{B1}(x, t) + V_{R1}(x, t) \geq V_D(x, t) + 2V_{MD}(x, t) + V_{B2}(x, t) + V_{R2}(x, t). \quad (13)$$

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

324 *2.2.3 Phase 3: linear correction with reliable surface measurements*

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

$$328 \quad E_3(x, t, \tau) = \alpha E_2(x, t, \tau) + \beta. \quad (14)$$

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

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

$$335 \quad \alpha, \beta, V_\alpha, V_\beta, V_{\alpha\beta}, V_{R3} = \mathbb{L}\mathbb{R}_{t' \in T_c(t), x' \in X_c(x)}[G(x', t') \sim E_2(x', t', 0)], \quad (15)$$

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

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

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

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

$$355 \quad V_3(x, t, \tau) = V_{F3}(x, t, \tau) + \alpha^2[V_M(x, t) + V_D(x, t) + 2V_{MD}(x, t)] + V_\alpha E_2(x, t)^2 + \\ 356 \quad 2V_{\alpha\beta} E_2(x, t) + V_\beta + V_{R3}. \quad (16)$$

357 Now that in-situ data have been included, systematic bias due to the misrepresentation of
 358 the surface air quality due to model and satellite information only, as well as representational issues
 359 due to the limited spatial resolutions of the model and satellite data with respect to specific points
 360 represented in the surface data, are considered to be captured in terms related to regression
 361 coefficient variance and residual variance. However, practical limitations on the availability of
 362 surface air quality measurement sites, as well as the tendencies of such sites to be clustered in
 363 high-population-density areas, might mean that there are some residual biases which are not fully
 364 captured in this formulation. In other words, by necessity, the data fusion process will be tailored
 365 towards better representing locations where surface monitors already exist, and the above
 366 formulation for phase 3 uncertainty will tend to be more appropriate in those types of areas, rather
 367 than, e.g., more rural areas which are not covered by surface-based monitors. Furthermore, biases
 368 in the in-situ data will not be accounted for, e.g., the known sensitivity of NO₂ monitors to other
 369 species (e.g., Steinbacher et al., 2007).

370 Comparing the phase 2 and 3 variance estimates, assuming zero lead-time, and assuming
 371 that inclusion of surface information will tend to decrease phase 3 uncertainty with respect to phase
 372 2, we can establish that:

$$373 \quad V_{B2}(x, t) + V_{R2}(x, t) \geq (\alpha^2 - 1)[V_M(x, t) + V_D(x, t) + 2V_{MD}(x, t)] + V_\alpha E_2(x, t)^2 + \\ 374 \quad 2V_{\alpha\beta} E_2(x, t) + V_\beta + V_{R3}. \quad (17)$$

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

379 *2.2.4 Phase 4: updating with recent, nearby in-situ data*

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

$$383 \quad E_4(x, t, \tau) = E_3(x, t, \tau) + \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x, x', t, t') [G(x', t') - E_3(x', t')], \\ 384 \quad (18)$$

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

391 e.g., the difference in space and time between the two sets of coordinates (Rasmussen & Williams,
 392 2006). Recent work has proposed the use of Gaussian process regression for interpolating air
 393 quality data in space and/or time based on sparse measurements, and have proposed using square
 394 exponential, Matérn, and periodic kernel functions for this purpose for different pollutants of
 395 interest (Jang et al., 2020; Malings et al., 2021; Wang et al., 2021). The approach used here to
 396 determine appropriate kernel functions and parameters is described in (Malings et al., 2021, section
 397 3.7). Equation 18 combines equations 11 and 14 of Malings et al. (2021), using a more generic
 398 notation of the kernel. A schematic for this phase is provided in Supplemental Figure S4.

399 Spatio-temporal kriging also quantifies the resulting uncertainty reduction:

$$400 \quad V_4(x, t, \tau) = V_3(x, t, \tau) - \sum_{x' \in X_{near}(x), t' \in T_{recent}(t)} K(x, x', t, t') \text{cov}[E_3(x', t'), E_3(x, t)],$$

401 (19)

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

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

410 *2.2.5 Quantifying uncertainties in phases 1 and 2*

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

$$419 \quad V_{B1}(x, t) + V_{R1}(x, t) \cong \eta_1^2(t \bmod 24\text{h}) \mathbb{E}_{t' \in T_c(t)} V_M(x, t'), \quad (20)$$

$$420 \quad V_{B2}(x, t) + V_{R2}(x, t) \cong \eta_2^2(t \bmod 24\text{h}) \mathbb{E}_{t' \in T_c(t)} [V_M(x, t') + V_D(x, t') + 2V_{MD}(x, t')].$$

421 (21)

422 In these estimates, the unquantified portions of the uncertainty are related to the quantified
 423 performance via empirically determined factors η_1 for phase 1 and η_2 for phase 2. These factors
 424 are assumed to vary as a function of time-of-day, based on observations for how relationships
 425 between different portions of the quantified uncertainty varied over the calibration period
 426 investigated here. Empirically determined values of η_1 and η_2 for San Francisco are presented
 427 Supplemental Figure S5; values for New York City are presented in Supplemental Figure S6.

428 This proposed approach has important limitations. Most notably, it relies on proceeding to
 429 phase 3 of the data fusion approach. In regions without ground-based monitoring, or where only a
 430 small number of ground-based monitors are available, the results from phase 3 of the data fusion
 431 approach will be unavailable or highly unreliable. Empirically determined values of η_1 and η_2

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

435 2.3 Confidence interval determination

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

$$446 \quad \mu_p(x, t, \tau) = \log \left[\frac{E_p(x, t, \tau)}{\sqrt{1 + \frac{V_p(x, t, \tau)}{E_p(x, t, \tau)^2}}} \right], \quad (22)$$

$$447 \quad \sigma_p(x, t, \tau) = \sqrt{\log \left[1 + \frac{V_p(x, t, \tau)}{E_p(x, t, \tau)^2} \right]}. \quad (23)$$

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

$$449 \quad F_p(x, t, \tau) \sim \text{LN} \left(\mu_p(x, t, \tau), \sigma_p(x, t, \tau) \right). \quad (24)$$

450 where $\text{LN}(\mu, \sigma)$ denotes a lognormal distribution with mean μ and standard deviation σ for the
 451 associated normal distribution. This distribution can be used to determine a CI for the quantity of
 452 interest. For example, the 75 % confidence range is defined with a lower bound, representing the
 453 12.5th percentile of the lognormal distribution, and an upper bound, representing the 87.5th
 454 percentile of the lognormal distribution.

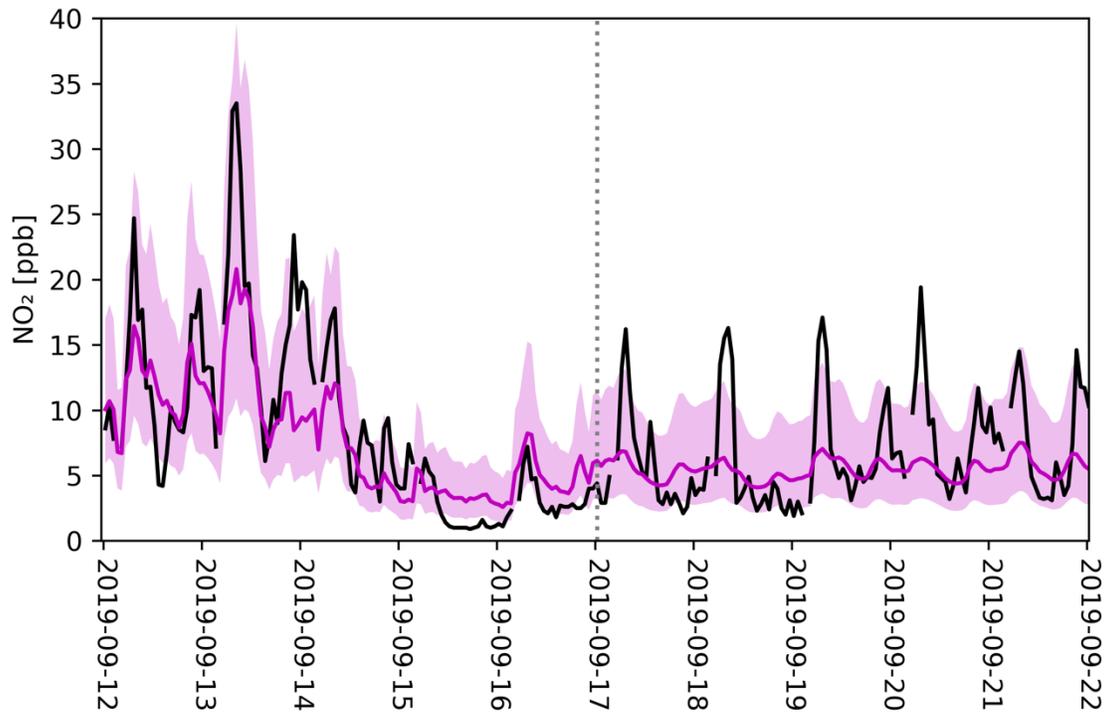
455 The lognormal distribution assumption is of course an approximation of the true
 456 distribution of the quantity of interest. Therefore, the CI determined as described above would not
 457 necessarily correspond to the actual CI for the quantity of interest, even if the mean and variance
 458 were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

460 3 Results

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

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

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



494

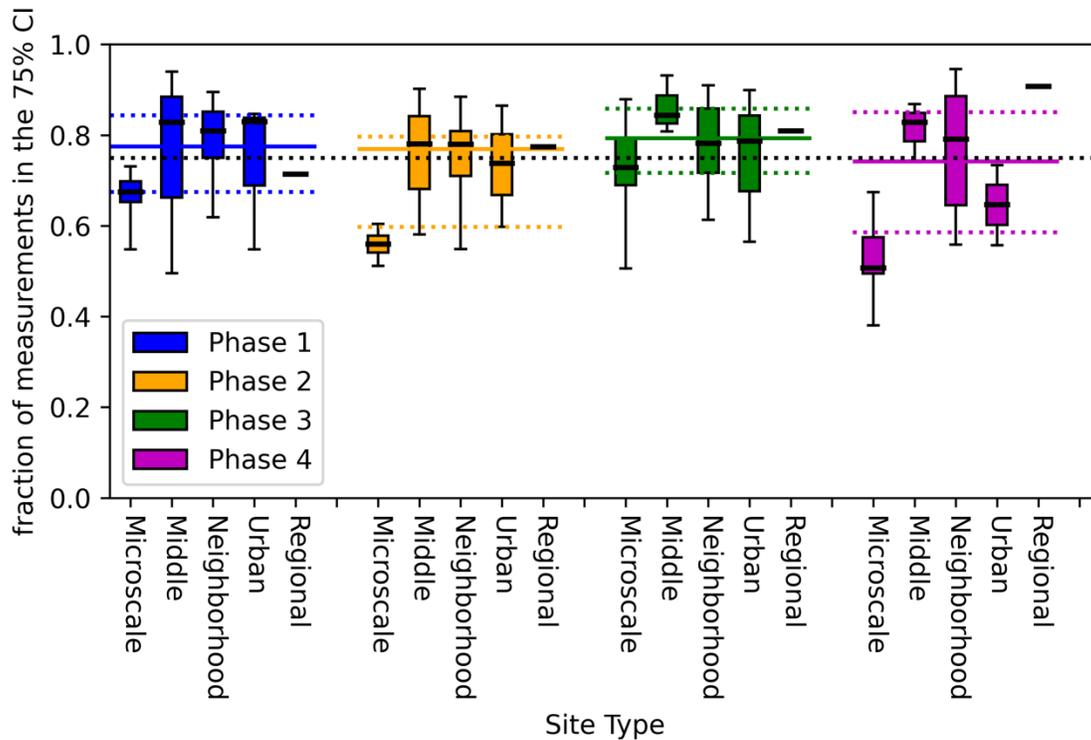
495 **Figure 1. Representative example of probabilistic estimates and forecasts for hourly surface-**
 496 **level NO₂ concentrations at the Redwood City monitor site (AQS ID 06-081-1001) in San**
 497 **Francisco, between September 12 and 22, 2019 local time. The black line indicates the**
 498 **reported concentrations from the regulatory monitor, i.e., the true concentration. The**
 499 **colored line indicates the mean estimated concentration from phase 4 of the data fusion**
 500 **process, $E_4(x, t)$. The colored shaded areas denote the 50 % CI for the estimates. Estimates**
 501 **are presented with zero lead time up to midnight on September 17th, denoted with a vertical**
 502 **dotted line. Beyond this, forecasts with an increasing lead time are presented.**

503

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

504

To investigate the accuracy of the assessed uncertainties in the data fusion, the fraction of
 505 actual measurements falling within the estimated 75 % CI across different phases of the data fusion
 506 approach is presented in Figure 2. This analysis considers all NO₂ monitor sites operating during
 507 September 2019 in the San Francisco study region, a total of 25 sites. The fraction of measurements
 508 falling within the 75 % CI is calculated for each site and considering the estimates for each phase
 509 of the data fusion process. Total uncertainties for phases 1 and 2 are estimated as outlined in section
 510 2.2.5. Horizontal colored solid and dotted lines indicate the median, 25th percentile, and 75th
 511 percentile values of these fractions across all sites for each phase. Furthermore, sites are divided
 512 into types based on their assumed scale of spatial representativity, which is assessed for each
 513 monitoring site by US EPA. The five site types are microscale (0-0.1 km; 5 sites), middle (0.1-0.5
 514 km; 3 sites), neighborhood (0.5-4 km; 13 sites), urban (4-50 km; 3 sites) and regional (50+ km; 1
 515 site), as defined in [40 CFR Part 58](#). By investigating the capacity of the data fusion system to
 516 capture uncertainties at different spatial scales in this way, its benefits and limitations can be better
 517 understood.



518

519 **Figure 2. Assessment of the fraction of actual measurements falling within the estimated 75**
 520 **% CI for different phases of the data fusion process, with phases represented by different**
 521 **colors. The analysis represents data from 25 active NO₂ ground monitoring sites in the San**
 522 **Francisco study region for September 2019. A horizontal dotted line across the figure**
 523 **indicates the goal, i.e., 75 % of measurements falling within the 75 % CI. For each ground**
 524 **monitor site, the fraction of measurements at that site falling within the 75 % CI is calculated.**
 525 **For each phase, a solid horizontal line in the corresponding color indicates the median of these**
 526 **fractions across sites, and two horizontal dotted colored lines indicate the 25th percentile and**
 527 **75th percentile of these fractions across sites. Furthermore, monitoring sites are divided into**
 528 **different site types. The spread in fraction of measurements falling within the 75 % CI for**
 529 **each site type is indicated with a box-and-whisker plot. In each box-and-whisker plot, the**
 530 **horizontal line inside the box denotes the median, the box denotes the 25th-to-75th-percentile**
 531 **range, and the whiskers denote the full range.**

532 Overall, for all phases of the data fusion process, the estimated 75 % CI captures roughly
 533 75 % of measured data. Performance is most consistent for phases 1 and 3, which have the smallest
 534 inter-quartile spreads in fraction of measurements falling within the 75 % CI. Focusing on phase
 535 1, where only model outputs are considered, performance is consistent across most site types.
 536 There is a slight bias towards underconfidence, i.e., more measurements falling within the 75 %
 537 CI than expected. For microscale sites, however, estimates are systematically overconfident, with
 538 fewer measurements falling within the 75 % CI than expected. Considering the native spatial
 539 resolution of the model, better representation of uncertainties at urban and regional scales is to be
 540 expected. There is a lack of information at this stage to make informed assessments of confidence
 541 at finer spatial scales. This manifests in the results with a slightly larger spread in performance for
 542 middle scale sites and the overconfidence noted for microscale sites.

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

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

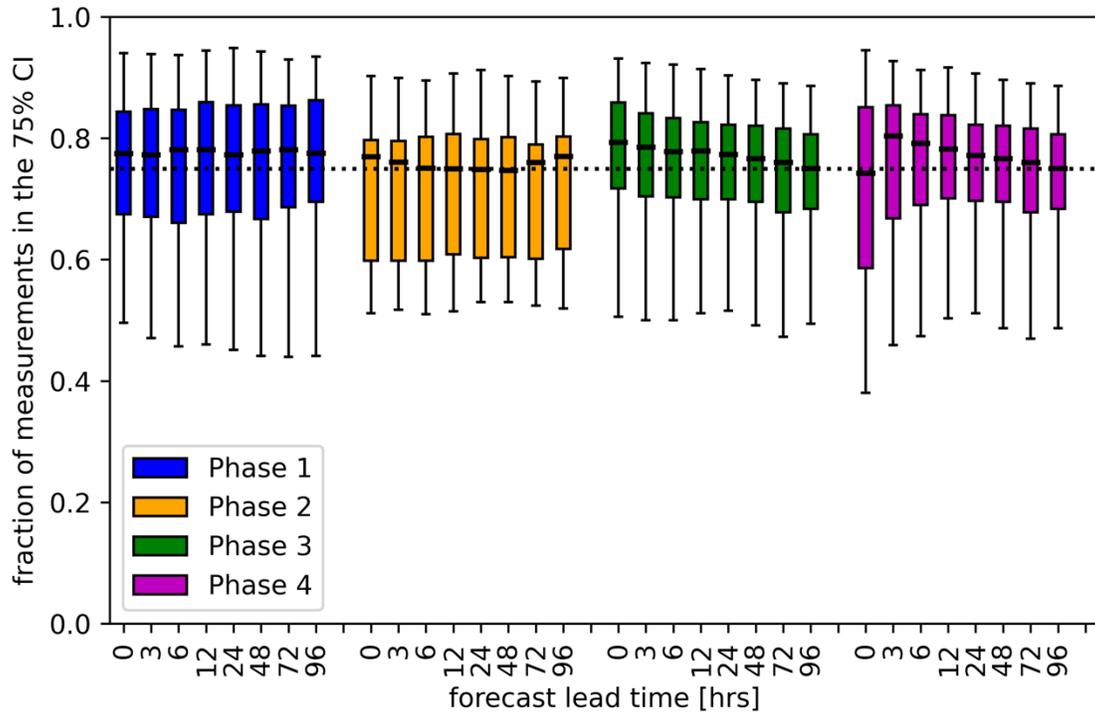
562 In phase 4, while uncertainty estimates seem to be most accurate in the median, the spread
563 in performance has increased. Microscale sites are again exhibiting systematic overconfidence,
564 along with urban scale sites, while middle scale and regional sites are underconfident. With only a
565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

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

584 3.2 Assessment of confidence interval coverage for different forecast lead times

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

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



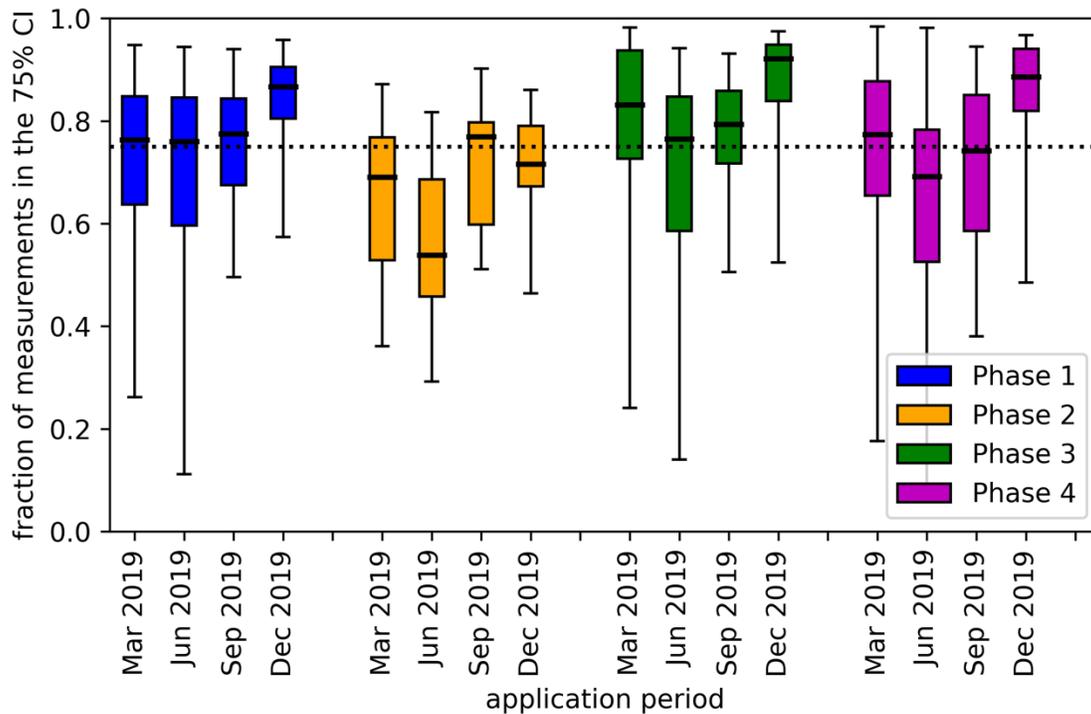
589

590 **Figure 3. Assessment of the fraction of actual measurements falling within the estimated 75**
 591 **% CI for different phases of the data fusion process, with phases represented by different**
 592 **colors, as a function of forecasting lead time, in hours. The analysis represents data from 25**
 593 **active NO₂ ground monitoring sites in the San Francisco study region for September 2019. A**
 594 **horizontal dotted line across the figure indicates the goal, i.e., 75 % of measurements falling**
 595 **within the 75 % CI. For each ground monitor site, the fraction of measurements at that site**
 596 **falling within the 75 % CI is calculated. The box-and-whisker plots denote the ranges of these**
 597 **fractions across sites, with the horizontal line in the box denoting the median, the box**
 598 **denoting the 25th-to-75th-percentile range, and the whiskers denoting the full range.**

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

610 3.3 Assessment of confidence interval coverage across different times of year

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



619
 620 **Figure 4. Fractions of measurements falling within the estimated 75 % CI for different**
 621 **phases of the data fusion process, with phases represented by different colors, presented for**
 622 **different application months. Box-and-whisker plots denote ranges of these fractions across**
 623 **active NO₂ monitor sites in San Francisco during that month, with the horizontal line in the**
 624 **box denoting the median, the box denoting the 25th-to-75th-percentile range, and the whiskers**
 625 **denoting the full range. The horizontal dotted line across the figure indicates the goal, i.e., 75**
 626 **% of measurements falling within the 75 % CI.**

627 A similar assessment was conducted for the region of New York City, as discussed in the
 628 supplemental materials. Results for CI coverage at zero forecast lead time in March 2019, June
 629 2019, September 2019, and December 2019 are shown in Supplemental Figure S9. Similar
 630 variability in performance for different phases in different months is observed as was noted above.
 631 Underconfidence in December 2019 seems to be more extreme, especially in phase 1, than in the
 632 case of San Francisco. Overconfidence in phase 2 also appears to be more severe. Again, monthly
 633 or seasonal differences in relevant parameters, especially the factors η_1 and η_2 calculated for the
 634 domain and kriging spatial and temporal scales associated with phase 4, might be influencing this.

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

640 **4 Conclusions**

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

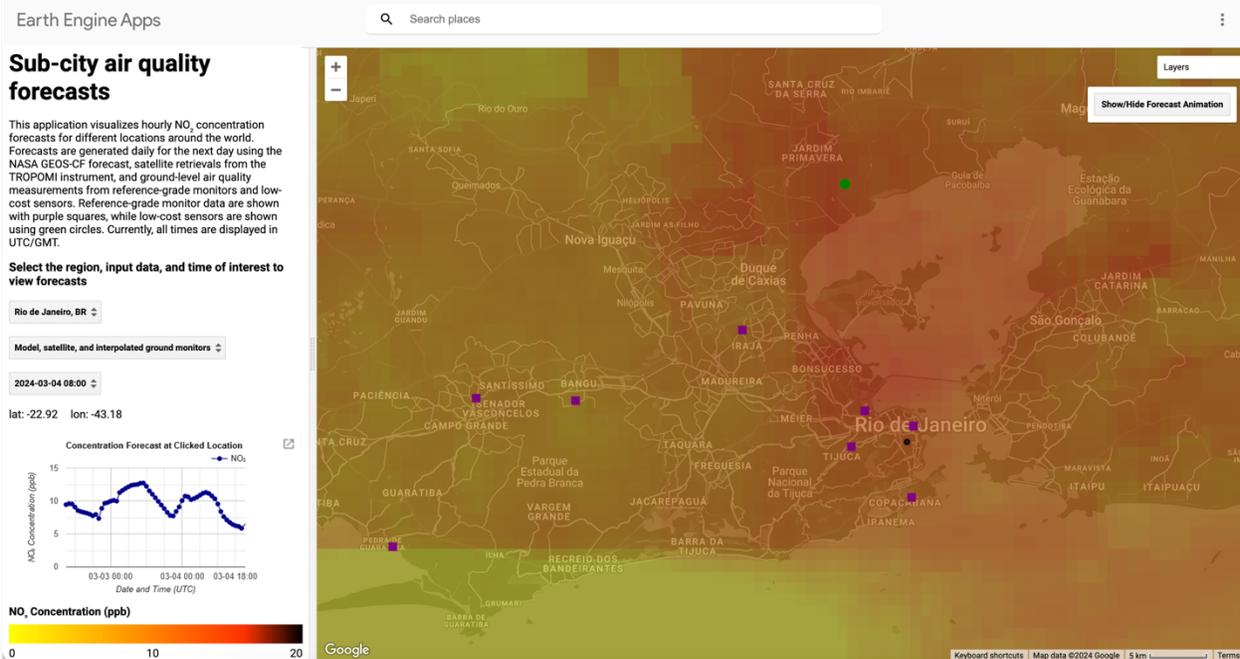
649 The uncertainty quantification was found to be least accurate overall for microscale sites,
650 which are most impacted by hyperlocal sources. In the San Francisco case study, these sites were
651 adjacent to highways, which are most heavily impacted by NO₂ pollution. This finding is useful to
652 convey to any user of this system, i.e., that results may not be reliable within about 100 meters of
653 a major source like a highway or other intense combustion activity. Similar limitations are likely,
654 should the method be applied to other constituents measured near their respective sources.

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

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

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

681 Nevertheless, the framework established here presents a reasonable prior CI for the
682 estimates and forecasts of the proposed data fusion system, and this fact supports effective and
683 appropriate interpretation of its output by users. For example, these uncertainty estimates might be
684 applied with respect to a given regulatory pollutant threshold to estimate the probability of
685 exceeding that threshold. Such information could support air quality management decision-
686 making. In an ongoing project supported by the NASA Health and Air Quality Applied Sciences
687 Program, the authors are implementing the data fusion and uncertainty quantification scheme
688 presented here in an online application via the [Google Earth Engine](#) platform. It is hoped that this
689 application will present a useful tool for local air quality managers to visualize sub-city-scale
690 atmospheric composition and variability using a combination of model, satellite, and in-situ data.
691 This project is being conducted in collaboration with local environmental managers in the USA,
692 Brazil, and Senegal. An example prototype for this tool is presented in Figure 5. As part of this
693 project, the framework will also be extended to other relevant pollutants, primarily PM_{2.5} and O₃.



694
695 **Figure 5. Screenshot of an application currently under development which will implement**
696 **the data fusion framework presented here, including uncertainty quantification, via the**
697 **[Google Earth Engine](#) platform. This application will enable air quality managers to access**
698 **and visualize estimates and forecasts of relevant air quality parameters such as NO₂, O₃,**
699 **PM_{2.5}, along with associated expressions of confidence. Example outputs are presented for**
700 **the city of Rio de Janeiro, Brazil, one of the partners for this project.**

701 Acknowledgements

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

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

713 **Open Research**

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

718 **Author Contributions**

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

725 **References**

- 726 Anenberg, S. C., Bindl, M., Brauer, M., Castillo, J. J., Cavalieri, S., Duncan, B. N., Fiore, A. M.,
727 Fuller, R., Goldberg, D. L., Henze, D. K., Hess, J., Holloway, T., James, P., Jin, X., Kheirbek, I.,
728 Kinney, P. L., Liu, Y., Moheg, A., Patz, J., ... West, J. J. (2020). Using Satellites to Track
729 Indicators of Global Air Pollution and Climate Change Impacts: Lessons Learned From a NASA-
730 Supported Science-Stakeholder Collaborative. *GeoHealth*, 4(7).
731 <https://doi.org/10.1029/2020GH000270>
- 732 Balashov, N. V., Huff, A. K., & Thompson, A. M. (2023). Interpretation of Probabilistic Surface
733 Ozone Forecasts: A Case Study for Philadelphia. *Weather and Forecasting*, 38(10), 1895–1906.
734 <https://doi.org/10.1175/WAF-D-22-0185.1>
- 735 Balashov, N. V., Thompson, A. M., & Young, G. S. (2017). Probabilistic Forecasting of Surface
736 Ozone with a Novel Statistical Approach. *Journal of Applied Meteorology and Climatology*, 56(2),
737 297–316. <https://doi.org/10.1175/JAMC-D-16-0110.1>
- 738 Bi, J., Knowland, K. E., Keller, C. A., & Liu, Y. (2022). Combining Machine Learning and
739 Numerical Simulation for High-Resolution PM_{2.5} Concentration Forecast. *Environmental Science*
740 *& Technology*, 56(3), 1544–1556. <https://doi.org/10.1021/acs.est.1c05578>
- 741 Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating Low-Cost Sensor Measurements
742 into High-Resolution PM_{2.5} Modeling at a Large Spatial Scale. *Environmental Science &*
743 *Technology*, 54(4), 2152–2162. <https://doi.org/10.1021/acs.est.9b06046>
- 744 Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness,
745 A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov,

- 746 A., Carnevale, C., Grell, G., & Seigneur, C. (2015). Data assimilation in atmospheric chemistry
 747 models: Current status and future prospects for coupled chemistry meteorology models.
 748 *Atmospheric Chemistry and Physics*, 15(10), 5325–5358. [https://doi.org/10.5194/acp-15-5325-](https://doi.org/10.5194/acp-15-5325-2015)
 749 2015
- 750 Cooper, M. J., Martin, R. V., Henze, D. K., & Jones, D. B. A. (2020). Effects of a priori profile
 751 shape assumptions on comparisons between satellite NO₂ columns and
 752 model simulations. *Atmospheric Chemistry and Physics*, 20(12), 7231–7241.
 753 <https://doi.org/10.5194/acp-20-7231-2020>
- 754 Diao, M., Holloway, T., Choi, S., O’Neill, S. M., Al-Hamdan, M. Z., Van Donkelaar, A., Martin,
 755 R. V., Jin, X., Fiore, A. M., Henze, D. K., Lacey, F., Kinney, P. L., Freedman, F., Larkin, N. K.,
 756 Zou, Y., Kelly, J. T., & Vaidyanathan, A. (2019). Methods, availability, and applications of PM_{2.5}
 757 exposure estimates derived from ground measurements, satellite, and atmospheric models. *Journal*
 758 *of the Air & Waste Management Association*, 69(12), 1391–1414.
 759 <https://doi.org/10.1080/10962247.2019.1668498>
- 760 Diggle, P. J. (2010). Historical Introduction. In A. E. Gelfand, M. Fuentes, P. Guttorp, & P. J.
 761 Diggle (Eds.), *Handbook of spatial statistics* (pp. 3–14). CRC Press.
- 762 Duncan, B. N., Malings, C. A., Knowland, K. E., Anderson, D. C., Prados, A. I., Keller, C. A.,
 763 Cromar, K. R., Pawson, S., & Ensz, H. (2021). Augmenting the Standard Operating Procedures of
 764 Health and Air Quality Stakeholders With NASA Resources. *GeoHealth*, 5(9).
 765 <https://doi.org/10.1029/2021GH000451>
- 766 Fanfarillo, A., Roozitalab, B., Hu, W., & Cervone, G. (2019). *Probabilistic Forecasting using Deep*
 767 *Generative Models*. <https://doi.org/10.48550/ARXIV.1909.11865>
- 768 Garaud, D., & Mallet, V. (2011). Automatic calibration of an ensemble for uncertainty estimation
 769 and probabilistic forecast: Application to air quality. *Journal of Geophysical Research*, 116(D19),
 770 D19304. <https://doi.org/10.1029/2011JD015780>
- 771 Garner, G. G., & Thompson, A. M. (2012). The Value of Air Quality Forecasting in the Mid-
 772 Atlantic Region. *Weather, Climate, and Society*, 4(1), 69–79. [https://doi.org/10.1175/WCAS-D-](https://doi.org/10.1175/WCAS-D-10-05010.1)
 773 10-05010.1
- 774 Gilliam, R. C., Hogrefe, C., Godowitch, J. M., Napelenok, S., Mathur, R., & Rao, S. T. (2015).
 775 Impact of inherent meteorology uncertainty on air quality model predictions. *Journal of*
 776 *Geophysical Research: Atmospheres*, 120(23). <https://doi.org/10.1002/2015JD023674>
- 777 Gilpin, S., Matsuo, T., & Cohn, S. E. (2023). A generalized, compactly supported correlation
 778 function for data assimilation applications. *Quarterly Journal of the Royal Meteorological Society*,
 779 149(754), 1953–1989. <https://doi.org/10.1002/qj.4490>
- 780 Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M.,
 781 Beekmann, M., & Subramanian, R. (2021). From low-cost sensors to high-quality data: A
 782 summary of challenges and best practices for effectively calibrating low-cost particulate matter
 783 mass sensors. *Journal of Aerosol Science*, 158, 105833.
 784 <https://doi.org/10.1016/j.jaerosci.2021.105833>
- 785 Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality
 786 mapping using low-cost sensor observations: Feasibility and added-value. *Environment*
 787 *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-](https://doi.org/10.1146/annurev-biodatasci-110920-093120)
 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 PM_{2.5}. *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>

- 831 Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O. L., Segers, A., & Heemink, A. W. (2021).
 832 Urban Air Quality Modeling Using Low-Cost Sensor Network and Data Assimilation in the Aburrá
 833 Valley, Colombia. *Atmosphere*, 12(1), 91. <https://doi.org/10.3390/atmos12010091>
- 834 Malings, C. (2024). *Supporting Data for “Air Quality Estimation and Forecasting via Data Fusion*
 835 *with Uncertainty Quantification: Theoretical Framework and Preliminary Results”* (1.0) [Python].
 836 Zenodo. <https://doi.org/10.5281/zenodo.10650853>
- 837 Malings, C., Knowland, K. E., Keller, C. A., & Cohn, S. E. (2021). Sub-City Scale Hourly Air
 838 Quality Forecasting by Combining Models, Satellite Observations, and Ground Measurements.
 839 *Earth and Space Science*, 8(7). <https://doi.org/10.1029/2021EA001743>
- 840 Martin, R. V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., & Dey, S. (2019). No one
 841 knows which city has the highest concentration of fine particulate matter. *Atmospheric*
 842 *Environment: X*, 3, 100040. <https://doi.org/10.1016/j.aeaoa.2019.100040>
- 843 McFarlane, C., Isevulambire, P. K., Lumbuenamo, R. S., Ndinga, A. M. E., Dhammapala, R., Jin,
 844 X., McNeill, V. F., Malings, C., Subramanian, R., & Westervelt, D. M. (2021). First Measurements
 845 of Ambient PM_{2.5} in Kinshasa, Democratic Republic of Congo and Brazzaville, Republic of
 846 Congo Using Field-calibrated Low-cost Sensors. *Aerosol and Air Quality Research*, 21.
 847 <https://doi.org/10.4209/aaqr.200619>
- 848 McFarlane, C., Raheja, G., Malings, C., Appoh, E. K. E., Hughes, A. F., & Westervelt, D. M.
 849 (2021). Application of Gaussian Mixture Regression for the Correction of Low Cost PM_{2.5}
 850 Monitoring Data in Accra, Ghana. *ACS Earth and Space Chemistry*, acsearthspacechem.1c00217.
 851 <https://doi.org/10.1021/acsearthspacechem.1c00217>
- 852 Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M.,
 853 Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abegaz, K. H., Abolhassani, H.,
 854 Aboyans, V., Abreu, L. G., Abrigo, M. R. M., Abualhasan, A., Abu-Raddad, L. J., Abushouk, A. I.,
 855 Adabi, M., ... Lim, S. S. (2020). Global burden of 87 risk factors in 204 countries and territories,
 856 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*,
 857 396(10258), 1223–1249. [https://doi.org/10.1016/S0140-6736\(20\)30752-2](https://doi.org/10.1016/S0140-6736(20)30752-2)
- 858 Raheja, G., Sabi, K., Sonla, H., Gbedjangni, E. K., McFarlane, C. M., Hodoli, C. G., & Westervelt,
 859 D. M. (2022). A Network of Field-Calibrated Low-Cost Sensor Measurements of PM_{2.5} in Lomé,
 860 Togo, Over One to Two Years. *ACS Earth and Space Chemistry*, 6(4), 1011–1021.
 861 <https://doi.org/10.1021/acsearthspacechem.1c00391>
- 862 Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT
 863 Press.
- 864 Riccio, A., & Chianese, E. (2024). Technical note: Accurate, reliable, and high-resolution air
 865 quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel
 866 statistical post-processing method. *Atmospheric Chemistry and Physics*, 24(3), 1673–1689.
 867 <https://doi.org/10.5194/acp-24-1673-2024>
- 868 Rose Eilenberg, S., Subramanian, R., Malings, C., Haurlyliuk, A., Presto, A. A., & Robinson, A. L.
 869 (2020). Using a network of lower-cost monitors to identify the influence of modifiable factors
 870 driving spatial patterns in fine particulate matter concentrations in an urban environment. *Journal*
 871 *of Exposure Science & Environmental Epidemiology*. <https://doi.org/10.1038/s41370-020-0255-x>

- 872 Schneider, P., Vogt, M., Haugen, R., Hassani, A., Castell, N., Dauge, F. R., & Bartonova, A. (2023).
 873 Deployment and Evaluation of a Network of Open Low-Cost Air Quality Sensor Systems.
 874 *Atmosphere*, 14(3), 540. <https://doi.org/10.3390/atmos14030540>
- 875 Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C.,
 876 Prevot, A. S. H., & Hueglin, C. (2007). Nitrogen oxide measurements at rural sites in Switzerland:
 877 Bias of conventional measurement techniques. *Journal of Geophysical Research: Atmospheres*,
 878 112(D11), 2006JD007971. <https://doi.org/10.1029/2006JD007971>
- 879 Tanzer, R., Malings, C., Hauryliuk, A., Subramanian, R., & Presto, A. A. (2019). Demonstration
 880 of a Low-Cost Multi-Pollutant Network to Quantify Intra-Urban Spatial Variations in Air Pollutant
 881 Source Impacts and to Evaluate Environmental Justice. *International Journal of Environmental*
 882 *Research and Public Health*, 16(14), 2523. <https://doi.org/10.3390/ijerph16142523>
- 883 US EPA. (2017). *Policy Assessment for the Review of the Primary National Ambient Air Quality*
 884 *Standards for Oxides of Nitrogen* (EPA-452/R-17-003). U.S. Environmental Protection Agency.
- 885 van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C.,
 886 Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., & Martin, R.
 887 V. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty.
 888 *Environmental Science & Technology*, 55(22), 15287–15300.
 889 <https://doi.org/10.1021/acs.est.1c05309>
- 890 van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2015). Use of Satellite Observations
 891 for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter.
 892 *Environmental Health Perspectives*, 123(2), 135–143. <https://doi.org/10.1289/ehp.1408646>
- 893 Veeffkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J.,
 894 de Haan, J. F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R.,
 895 Tol, P., Ingmann, P., Voors, R., Kruizinga, B., ... Levelt, P. F. (2012). TROPOMI on the ESA
 896 Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for
 897 climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120, 70–83.
 898 <https://doi.org/10.1016/j.rse.2011.09.027>
- 899 Wang, P., Mihaylova, L., Chakraborty, R., Munir, S., Mayfield, M., Alam, K., Khokhar, M. F.,
 900 Zheng, Z., Jiang, C., & Fang, H. (2021). A Gaussian Process Method with Uncertainty
 901 Quantification for Air Quality Monitoring. *Atmosphere*, 12(10), 1344.
 902 <https://doi.org/10.3390/atmos12101344>
- 903 Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., & Zhu,
 904 T. (2023). First close insight into global daily gapless 1 km PM_{2.5} pollution, variability, and health
 905 impact. *Nature Communications*, 14(1), 8349. <https://doi.org/10.1038/s41467-023-43862-3>
- 906 Zhang, H., Wang, J., García, L. C., Ge, C., Plessel, T., Szykman, J., Murphy, B., & Spero, T. L.
 907 (2020). Improving Surface PM_{2.5} Forecasts in the United States Using an Ensemble of Chemical
 908 Transport Model Outputs: 1. Bias Correction With Surface Observations in Nonrural Areas.
 909 *Journal of Geophysical Research: Atmospheres*, 125(14). <https://doi.org/10.1029/2019JD032293>
- 910 Zhang, H., Wang, J., García, L. C., Zhou, M., Ge, C., Plessel, T., Szykman, J., Levy, R. C., Murphy,
 911 B., & Spero, T. L. (2022). Improving Surface PM_{2.5} Forecasts in the United States Using an
 912 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