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}:

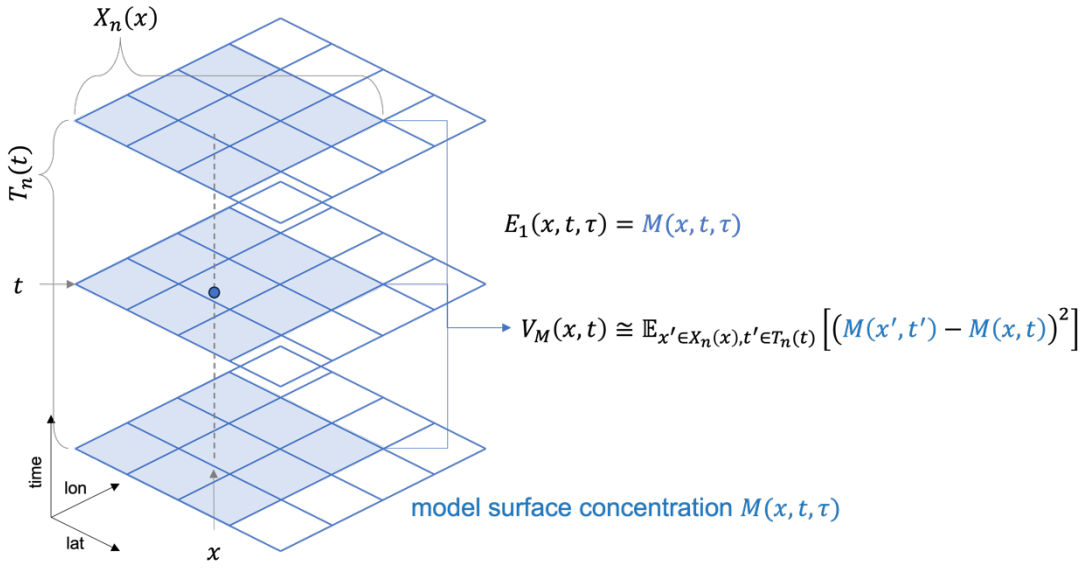
$$\boldsymbol{\zeta}, \xi, \mathbf{V}_{\boldsymbol{\zeta}}, V_{\xi}, \mathbf{V}_{\boldsymbol{\zeta\xi}}, V_{R,LCS} = \mathbb{LR}_{t' \in T_c(t), x' \in X_{LCS}}[E_3(x',t') \sim \mathbf{G}_{LCS}(x',t')]. \quad \text{(S1)}$$

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.

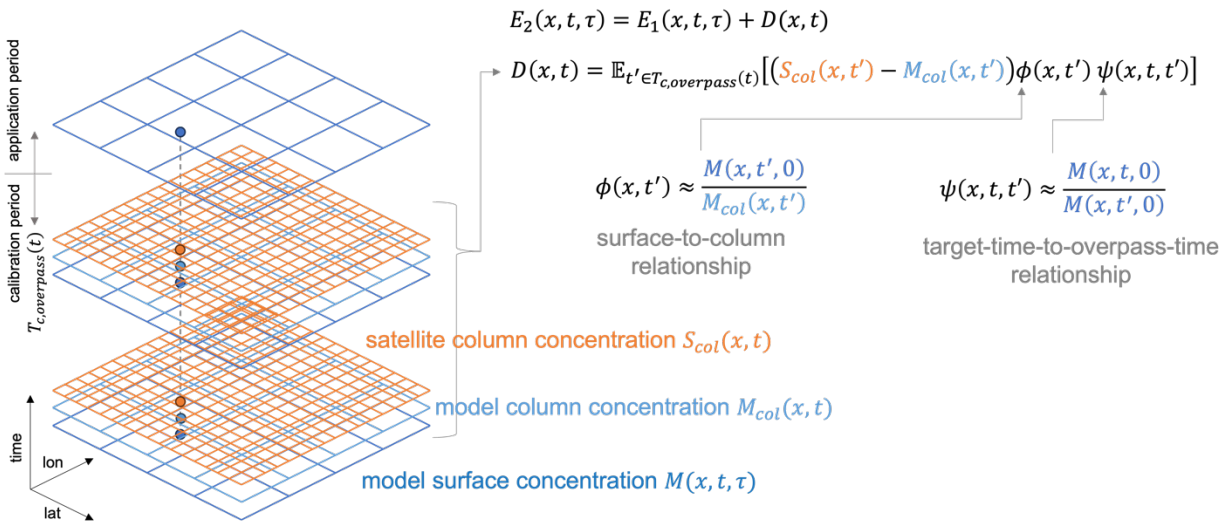
This approach is most suited to networks of LCS containing multiple devices with high inter-sensor precision and where the network is broadly distributed at a representative set of locations over the region of interest. In situations where inter-sensor precision is low, few LCS and no RGM are available,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



985

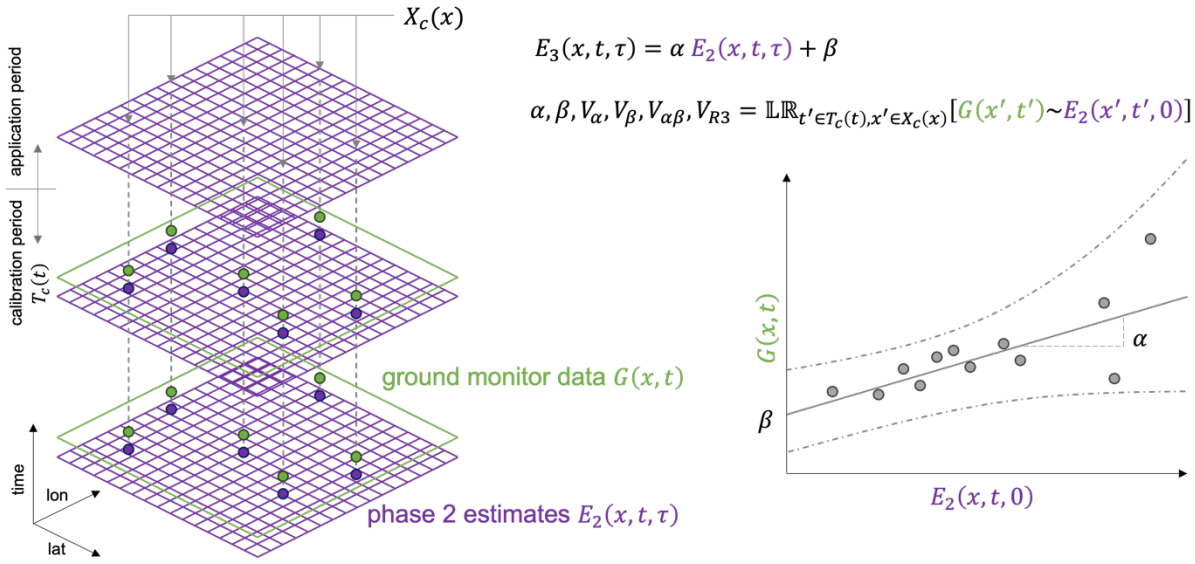
986 **Figure S1. Diagram of phase 1 of the data fusion process. Blue grids denote model grids in**
987 **space, with different layers denoting different timesteps. Shaded grids indicate the**
988 **neighborhood of the grid cell corresponding to location x and time t, used for estimation**
989 **of model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

990



991

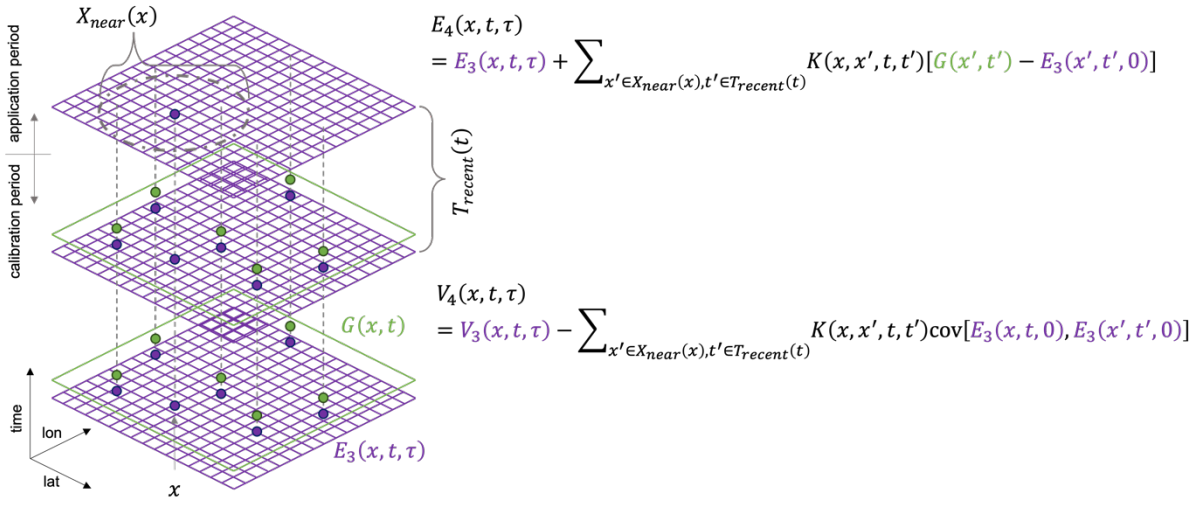
992 **Figure S2. Diagram of phase 2 of the data fusion process. Orange grids denote satellite**
993 **remote sensing data, with light blue grids corresponding to the analogous modeled column**
994 **quantity.**



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.**

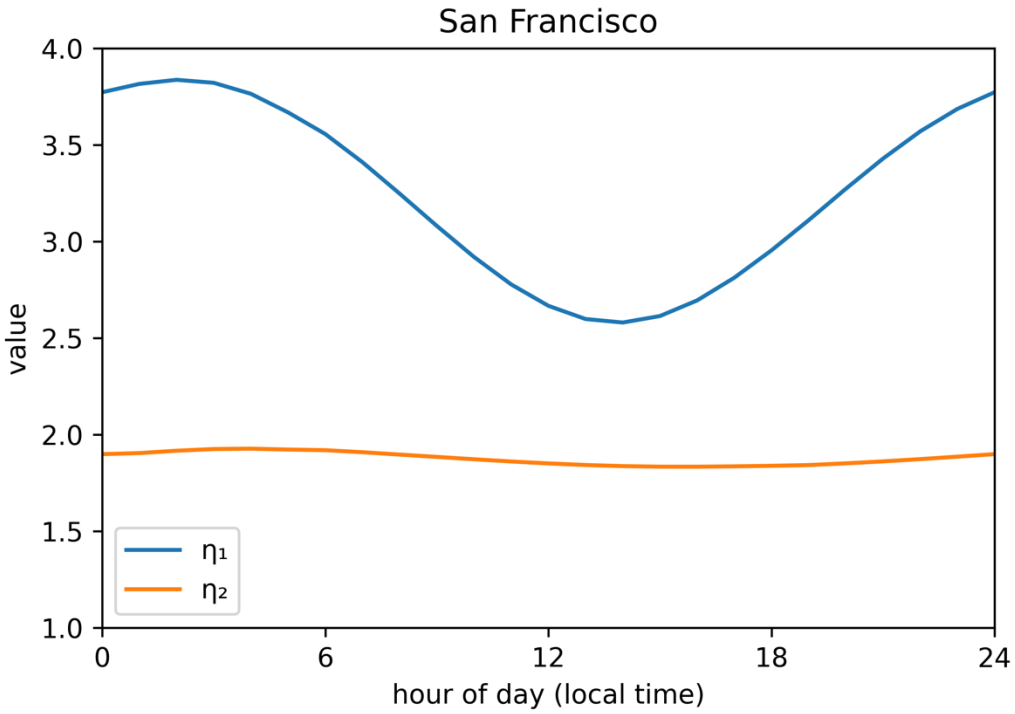


Figure S5. Empirically determined values for η_1 and η_2 used for San Francisco in this paper, as a function of hour of the day (presented in local time).

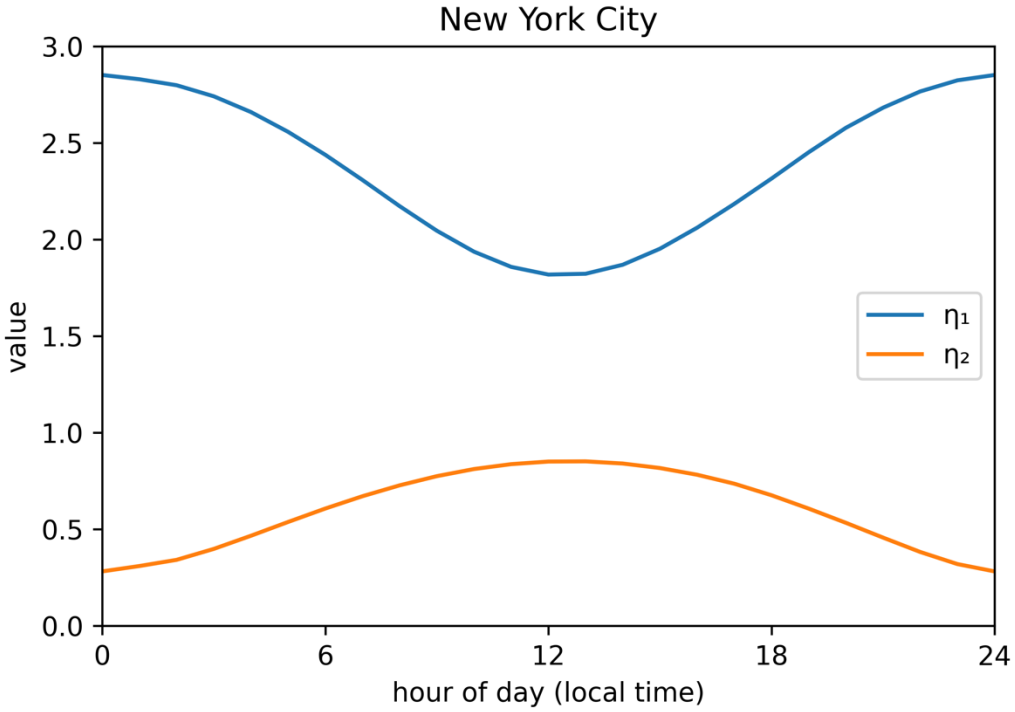


Figure S6. Empirically determined values for η_1 and η_2 used for New York City in this paper, as a function of hour of the day (presented in local time).

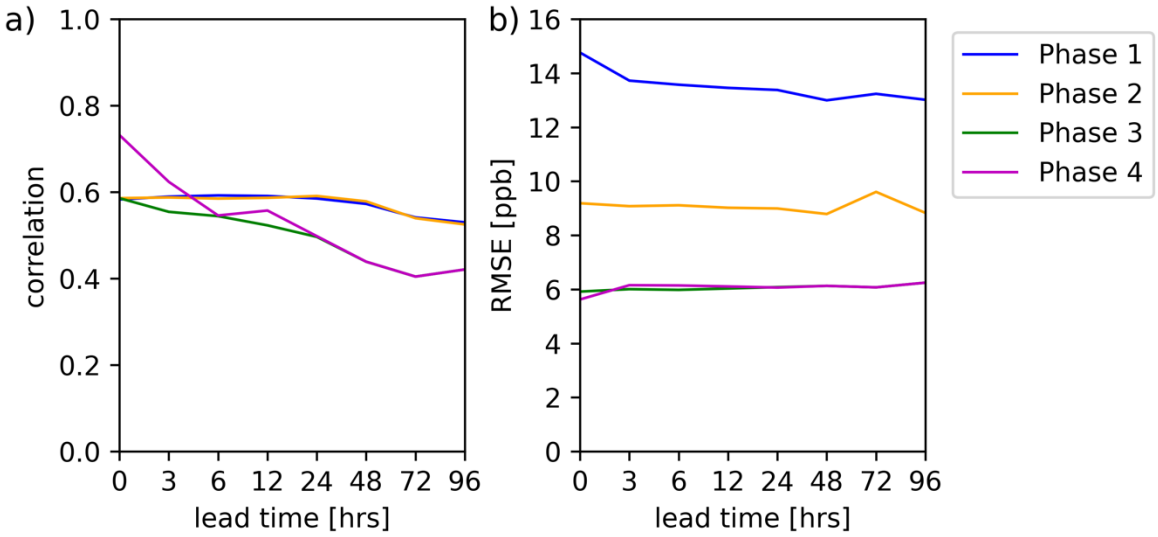


Figure S7. Summary performance metrics for the data fusion approach, evaluated for the San Francisco study region in September 2019 (same results as presented in Figure 2). Plots depict the Pearson correlation (a) and root mean square error (b) between the estimates of the various data fusion phases (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.

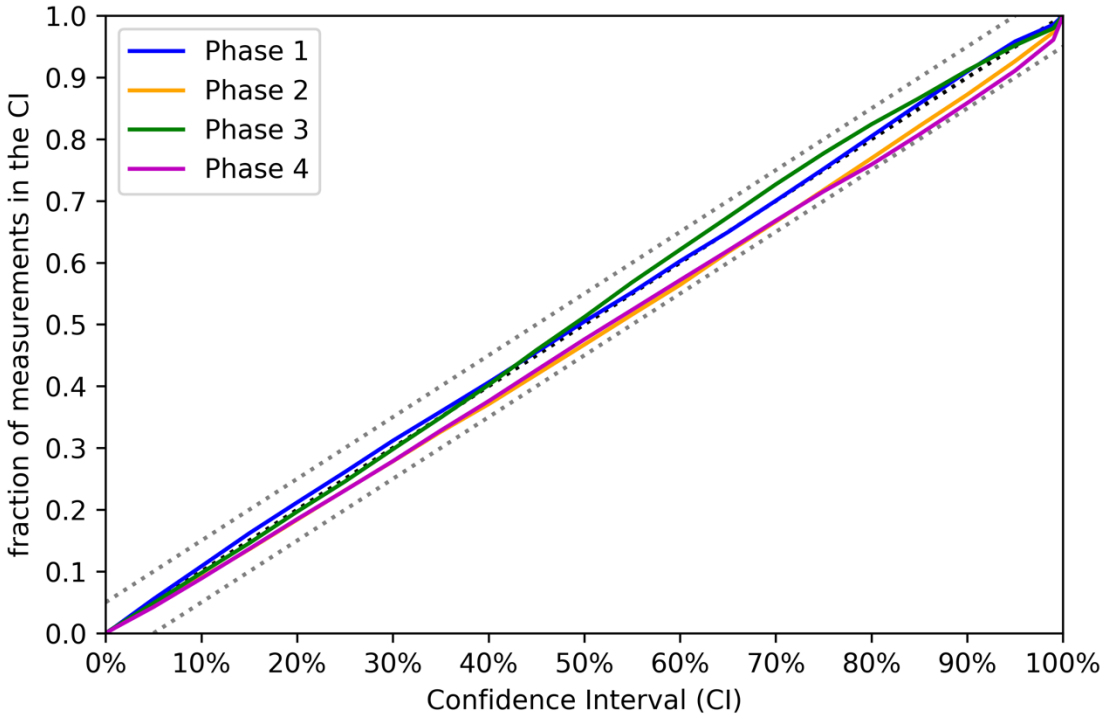


Figure S8. Assessment of CI coverage for different CI. The horizontal axis reports the nominal coverage of the CI, and the vertical axis reports the actual fraction of measurements falling within that CI. The assessment was conducted for zero lead time estimates in the San Francisco study region for 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.

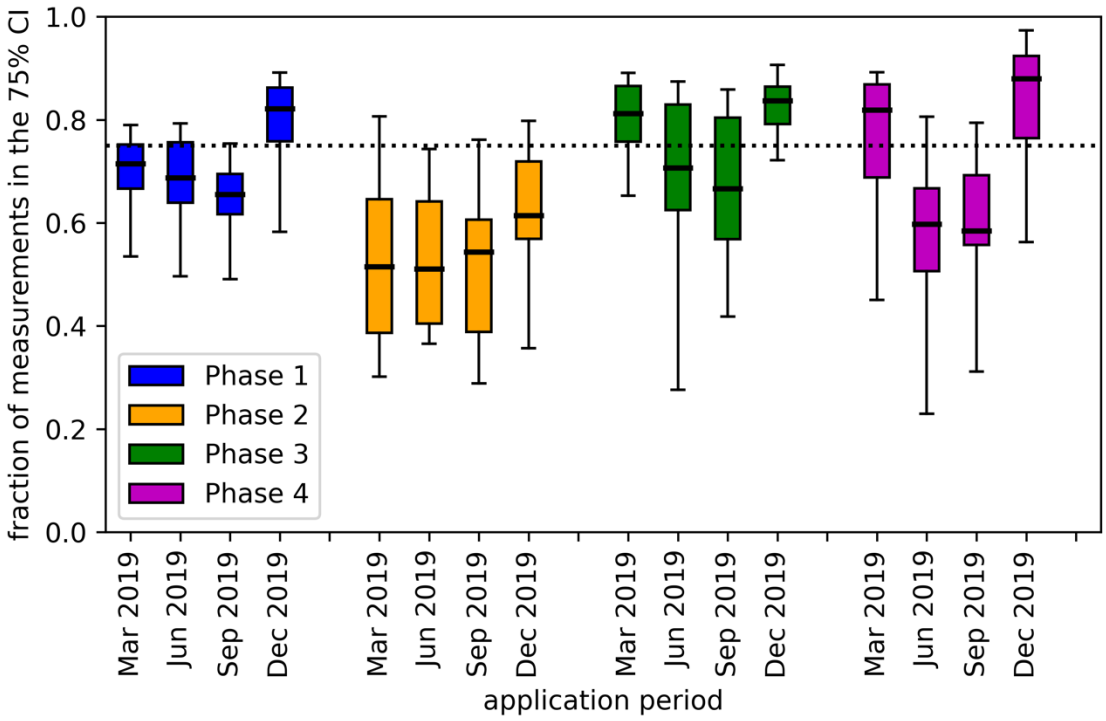


Figure S9. Fractions of measurements falling within the estimated 75 % CI for different phases of the data fusion process, with phases represented by different colors, presented for different application months. Box-and-whisker plots denote ranges of these fractions across active NO_2 monitor sites in 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.