Gaussian Process filtering for calibration of low-cost air-pollution sensor network data

Claire Heffernan\textsuperscript{1}, Roger Peng \textsuperscript{1}, Drew R. Gentner \textsuperscript{2}, Kirsten Koehler \textsuperscript{3}, Abhirup Datta\textsuperscript{1*}

\textsuperscript{1}Department of Biostatistics, Johns Hopkins University
\textsuperscript{2}Department of Chemical & Environmental Engineering, Yale University
\textsuperscript{3}Department of Environmental Health and Engineering, Johns Hopkins University

*Corresponding author: abhidatta@jhu.edu

ABSTRACT. Low-cost air pollution sensors, offering hyper-local characterization of pollutant concentrations, are becoming increasingly prevalent in environmental and public health research. However, low-cost air pollution data can be noisy, biased by environmental conditions, and usually need to be field-calibrated by co-locating low-cost sensors with reference-grade instruments. We show, theoretically and empirically, that the common of procedure of regression-based calibration using co-located data systematically underestimates high air-pollution concentrations, which are critical to diagnose from a health perspective. Current calibration practices also often fail to utilize the spatial correlation in pollutant concentrations. We propose a novel spatial filtering approach to co-location-based calibration of low-cost networks that mitigates the underestimation issue by using an inverse regression and incorporates spatial correlation by second-stage modeling of the true pollutant concentrations using a conditional Gaussian Process. Our approach works with one or more co-located sites in the network and is dynamic, leveraging spatial correlation with the latest available reference data. Through extensive simulations, we demonstrate how the spatial filtering substantially improves estimation of pollutant concentrations, and measures peak concentrations with greater accuracy. We apply the methodology for calibration of a low-cost PM2.5 network in Baltimore, Maryland, and diagnose air-pollution peaks that are missed by the regression-calibration.

Keywords: spatial statistics, Gaussian Process, Bayesian, air pollution, low-cost sensors

1 Introduction

Air pollution is regulated nationally in the United States using reference-grade instruments that conform to measurement standards like the Federal Reference Method (FRM) or Federal Equivalent Method (FEM) set by the Environmental Protection Agency (EPA) [United States Environmental Protection Agency 2021a]. However, regulatory networks offer sparse geographical coverage [Apte et al. 2017] and hence the gold standard data cannot assess disparities in air-quality at fine scale spatial resolution.
To fill the knowledge gap left by the regulatory monitoring, local networks of low-cost air-pollution sensors are being increasingly installed in many areas, including Los Angeles (Lu, 2021), Salt Lake City (Chadwick et al., 2021), Denver (Considine et al., 2021), Berkeley (Kim et al., 2018) and the San Francisco Bay area (Apte et al., 2017). These sensors are orders of magnitude cheaper than the high-precision, high accuracy regulatory devices. They can thus be deployed in larger numbers, creating dense monitoring networks that have high spatial resolution, which allows for neighborhood-level estimates of air pollution concentrations. They also have the potential to make higher time resolution measurements, up to several times each second, as opposed to the hourly or daily data offered by the some regulatory devices. The hyper-local, high-frequency characterization of exposures from low-cost sensors promise insights on air pollution and its health impacts at spatio-temporal scales beyond the scope of the sparse regulatory networks.

The data abundance of low-cost sensors comes at the expense of data quality. Sensor data quality depends on many variables including the manufacturer, sensor type, meteorological factors like relative humidity and temperature, the chemical composition of particulates, time since installation, and cross-sensitivity to other pollutants, among others. Even sensors from the same manufacturer can perform differently under varying ambient conditions. The correlation between measurements from low-cost sensors of different types and reference devices varies from 0.01 to 0.77 in a study by the United States Environmental Protection Agency (2021b). For this reason, raw data from these low-cost networks is not an accurate representation of the pollutant surface.

To enhance the data accuracy of low-cost sensor networks, field-calibration is often used, co-locating some of the sensors in the network with one or more high-quality reference instruments in the network domain (Zimmerman et al., 2018; Topalović et al., 2019; Datta et al., 2020). The paired time-series of co-located reference and low-cost measurements is used to train a regression model which subsequently calibrates data from other sensors in the network. Different regression approaches to calibration include multiple linear regression (Bigi et al., 2018; Bi et al., 2020; Ardon-Dryer et al., 2020; Si et al., 2020; Barkjohn et al., 2021; Datta et al., 2020; Romero et al., 2020), random effects models (Nordio et al., 2013), land-use regression (Clougherty et al., 2013; Larson et al., 2009), and machine learning methods like random forests (Lim et al., 2019; Zimmerman et al., 2018), neural networks (Topalović et al., 2019) and boosting (Johnson et al., 2018).

Regression-based field-calibration approaches reduce the bias in the low-cost sensor measurements but suffer from two major limitations.

1. High air pollution exposures disproportionately affect health, and we will show theoretically and empirically that linear regression-calibration systematically underestimates high levels of exposure.

2. Air pollution concentrations exhibit similarity across space, but regression-calibration is typically applied separately to each sensor in the network. This practice does not leverage this correlation – neither among the low-cost data from different sites in the network, nor between the low-cost data and concurrent measurements from nearby reference devices which could inform estimation of the true pollution concentrations at the network sites.
In this paper, we propose a novel spatial filtering approach to co-location based field-calibration of low-cost sensor networks that mitigates the aforementioned shortcomings of regression-calibration. We first address the underestimation of regression-calibration by considering an inverse regression model, where the low-cost measurement is regressed on the true pollutant concentration and other covariates. We motivate this change by making connections of the forward and inverse regression models to Berkson and classical measurement errors, respectively. Low-cost data is a noisy observation of the latent true pollutant concentration at the same location. Hence, the inverse regression, modeling classical measurement error, is more appropriate. We prove that the underestimation issue is not present in the inverse model.

To leverage spatial correlation among pollutant concentrations, we then extend the inverse regression to a spatial filtering method for spatially informed predictions of the true pollutant concentrations from the low-cost data. The inverse regression is the observation model part of the filter. As the true pollutant concentration is now the independent variable (covariate) in the inverse regression, we can seamlessly add a second-stage model for the true pollutant concentrations that captures the spatial correlation. Since the true pollutant surface is partially observed at the reference sites, we use a conditional Gaussian Process spatial model to incorporate this information. This corresponds to the state-transition part of the filter. Our method thus filters the low-cost data over space given the gold standard measurements at reference sites, which results in a smooth estimated pollutant surface. Another advantage is that our method leverages the current data from the reference devices for dictating the spatial state-transition model, resulting in a dynamic calibration while other calibration methods only use the reference data for a fixed time-window for training a static regression model.

Another novelty of our approach is the filtering in space. Filtering approaches have been used to predict pollutant concentrations in the past, but the filtering has always been over time, not space (Van der Wal and Janssen 2000; Wu et al. 2020; Tang et al. 2013; June et al. 2021). The high-frequency low-cost data offer the opportunity to characterize ultra-short-term fluctuations of the pollutant concentrations, and filtering in time using lower frequency reference measurements will smooth these out, leading to loss of information. Filtering across space conditions the analysis on the information available on the true pollutant concentrations at the reference sites. Thus the unique setting of co-located calibration of low-cost networks mandates filtering in space instead of time.

We offer both a frequentist and a Bayesian implementation of the spatial filtering. The advantage of the Bayesian implementation is that the uncertainty of the model parameter estimation is propagated in the filtering, while in the faster frequentist implementation, only the parameter estimates from a preliminary step are plugged into the filtering step. Extensive numerical studies using simulated data were used to evaluate the method. We see that across a wide range of scenarios, compared to regression-calibration, spatial filtering offers consistently improved overall Root mean squared error (RMSE) and better identification of high pollution events. We apply the filtering method to calibrate PM$_{2.5}$ data from a low-cost sensor network data in Baltimore. The spatial filtering performs much better than regression-calibration in identifying high pollution days and is used to create maps of PM$_{2.5}$ concentrations in the city.
2 SEARCH low-cost PM$_{2.5}$ network in Baltimore

We first illustrate the underestimation of the regression-calibration model using data from a network of low-cost air pollution sensors (Buehler et al., 2021) in Baltimore, Maryland, installed by the Solutions to Energy, Air, Climate, and Health (SEARCH) Center. Within Baltimore City limits, there is only one regulatory site (at Oldtown) managed by the Maryland Department of Energy (MDE) that measures hourly PM$_{2.5}$ in the city using a reference monitor. Additionally, there is one reference device on the outskirts of the city at the Essex site, which measures PM$_{2.5}$ every 6 days. This regulatory PM$_{2.5}$ data from only two sites is not sufficient to provide insight about local fluctuations in air quality, which is important to understand environmental injustice within the city. The SEARCH network has sensors at more than 30 locations that measure multiple pollutants including PM$_{2.5}$ as well as relative humidity (RH) and temperature (T). Low-cost sensors have been co-located with the reference instruments at both Oldtown and Essex (Figure 1).

Previous work (Datta et al., 2020) used this field-co-location data from Oldtown to estimate a gain-offset model (Balzano and Nowak, 2007) for regression-calibration of the low-cost PM$_{2.5}$ data:

$$E(x(s,t)) = o(s,t) + g(s,t)y(s,t),$$

where $o(s,t) = \phi'z(s,t)$ and $g(s,t) = \gamma'z(s,t)$, (1)

where $x(s,t)$ denotes the reference PM$_{2.5}$ data at location $s$ and time $t$ and $y(s,t)$ is the corresponding low-cost data. The gain $g$ (multiplicative bias) and offset $o$ (additive bias) were modeled as linear functions of the covariates $z(s,t)$ (RH, T, a weekend indicator, and a daylight indicator). The model parameters were estimated on a training window using least squares. Predictions from this calibration model were substantially more accurate and precise than the raw or lab-corrected low-cost data.

While the overall performance of the regression-calibration model was satisfactory, producing RMSE at par with reported numbers for other studies using similar calibrated low-cost PM$_{2.5}$ data, the performance of this model specifically during windows of high pollutant concentrations was not considered. Figure 1 (right) presents comparisons of the model predictions in December 2019 for the SEARCH low-cost sensor at Oldtown compared to the reference instrument at that site. The regression-calibration model clearly underestimates when the true PM$_{2.5}$ reaches unhealthy levels on December 23, 2019. The predicted concentrations from the regression-calibration model (~20µg/m$^3$) is nearly half of the true concentration (~40µg/m$^3$).

The Air Quality Index (AQI) is a classification of the concentrations of different pollutants in the air. In the United States, for PM$_{2.5}$, the AQI classifies PM$_{2.5}$ concentrations of ≤12µg/m$^3$ as Good, 12.1 − 35.4µg/m$^3$ as Moderate, and ≥35.5µg/m$^3$ as Unhealthy. A metric for evaluating the health-relevant performance of a calibration method is how often moderate or unhealthy observations are misclassified according to these cutoffs. Table 1 shows that 36% of the moderate instances, as measured by the reference instruments, are incorrectly predicted by the regression classification model as being good, and all of the truly unhealthy timepoints are incorrectly predicted as being moderate. This example shows the misclassification of PM$_{2.5}$ concentrations by the regression-calibration model.
Figure 1: SEARCH low-cost network: (Left) Map of the SEARCH network in Baltimore. The blue site is the co-located site at Oldtown which has a reference instrument for PM$_{2.5}$ from the Maryland Department of Energy (MDE). The red sites are non-co-located monitors. The purple site is at Essex outside of the city and has a reference instrument that measure PM$_{2.5}$ every 6 days and co-located low-cost sensors. This device will be used to validate calibration methods. (Right) Daily PM$_{2.5}$ time series in December 2019 of data from Oldtown MDE reference instrument and regression-calibrated data from the SEARCH low-cost sensor co-located at Oldtown.

For high values of true PM$_{2.5}$ concentrations. As high levels of exposure affect health disproportionately, it is critical for calibration techniques for low-cost sensors to be aware of this asymmetry in risks of exposure misclassification and be able to accurately identify days high air pollution events.

Table 1: Misclassification rates according to AQI classification of the regression-calibration model, December 2019. US AQI Classifications are: Good PM$_{2.5}$ is less than 12µg/m$^3$, Moderate is 12.1 – 35.4µg/m$^3$, Unhealthy is 35.5µg/m$^3$ or more.

| Prediction Classification | Good (%) | Moderate (%) | Unhealthy (%) | Sample Size |
|--------------------------|----------|--------------|---------------|-------------|
| True PM$_{2.5}$          |          |              |               |             |
| Good (≤ 12µg/m$^3$)      | 88       | 12           | 0             | 484         |
| Moderate (12.1 – 35.4µg/m$^3$) | 36       | 64           | 0             | 181         |
| Unhealthy (≥ 35.5µg/m$^3$) | 0        | 100          | 0             | 28          |
3 Methods

3.1 Low-cost air pollution networks

A schematic of a general low-cost air pollution network is shown in Figure 2. The locations in the schematic can be split into four sets. The blue sites have a reference device and a co-located low-cost sensor at that location, and will be referred to as Set A. At least one site must be in Set A to learn the biases in the low-cost data based on co-location. The red sites, Set B, only have low-cost sensors. This set is typically numerous. Set C is the green sites, where there is a reference device and no low-cost sensor. This set will typically have very few sites as the reference network is sparse, and it can even be empty if there is a low-cost sensor placed at every reference site in the area. Lastly, in addition to predicting the true pollutant concentrations at the sites of low-cost sensors, an additional goal is to predict pollutant concentrations at a dense grid of locations which are then interpolated to create maps of pollutant concentrations for the area. Set D represents such a grid of locations, denoted by the black crosses, that have no sensors but where we would like to predict pollutant concentrations.

![Schematic of a network of reference and low-cost sensors](image)

Figure 2: Schematic of a low-cost air-pollution network in an area with a 6 reference devices (blue and green sites) and 20 low-cost sensors (blue and red sites), where the blue sites are co-located sites. The grid of crosses indicates locations at which a PM$_{2.5}$ prediction is desired but there are no instruments.
3.2 Regression calibration

The gain-offset model \( \text{(1)} \) subsumes a large class of regression-calibration models. For example, if no covariates \( z \) are considered, it reduces to the most basic calibration model

\[
E(x(s, t)) = \beta_0 + \beta_1 y(s, t)
\]  

(2)

with constant gain \( \beta_1 \) and offset \( \beta_0 \). In many sensor calibration problems, (2) is popularly used (Miskell et al., 2018; Balzano and Nowak, 2007; Zheng et al., 2019). The offset \( \beta_0 \) and gain \( \beta_1 \) are often modeled as device-specific parameters in settings offering time and resources to co-locate each low-cost sensor with a reference device. Such co-location of each sensor takes a significant amount of time and considerably delays deployment of the network, so it is not always feasible. Hence, available covariates are commonly leveraged to capture variability in the biases of the sensors installed at different times, locations and ambient settings. A popular choice is calibration using the multiple linear regression (MLR) model

\[
E(x(s, t)) = \beta_0 + \beta_1 y(s, t) + \beta_2 z(s, t)
\]

where \( z \) is the set of covariates. MLR is widely used for calibrating low-cost air-pollution sensors (Bigi et al., 2018; Bi et al., 2020; Ardun-Dryer et al., 2020; Si et al., 2020; Barkjohn et al., 2021; Romero et al., 2020). The choice of covariates depends on the network design, type of pollutant and will typically include meteorological variables, daily, weekly or seasonal periodicity variables, time since installation, land-use variables, etc. MLR is also a special case of the gain-offset model \( \text{(1)} \) with a constant gain \( \beta_1 \) and the offset being a linear model of the covariates.

We consider the gain-offset model for regression-calibration in its most general form \( \text{(1)} \) as it subsumes all the aforementioned special cases and allows to also model the gain as a function of covariates. Rewriting \( \text{(1)} \) as

\[
E(x(s, t)) = \beta_0 + \beta_1 y(s, t) + \beta_2 z(s, t) + \beta_3 z(s, t)y(s, t)
\]  

(3)

we note that it corresponds to a regression model which allows for interaction of the low-cost data \( y \) with each covariate in \( z \). The model coefficients \( \beta_0, \beta_1, \beta_2, \beta_3 \) can be fit using least squares on the data from the co-located sites (Set A in Figure 2) and predictions can be made across the entire network (Set B).

This regression-calibration model suffers from two major limitations: (a) the predictions \( \hat{x} \) underestimate \( x \) when \( x \) is large, and (b) the model does not consider the underlying spatial structure of air pollution. We illustrate the underestimation issue here and discuss issue (b) in Section 3.4.

**Proposition 1.** Consider a data generation process relating low-cost pollutant measurements and the true pollutant values given by Equation \( \text{(3)} \) and with i.i.d. errors \( \epsilon \). Assume that the covariates low-cost measurements are bounded and that \( \text{Var}(\epsilon) < \infty \). Then the bias \( \hat{x} - x \) of the predictions \( \hat{x} \) from regression calibration is asymptotically negatively correlated with the true pollutant concentration \( x \).
The proof is given in the Appendix 1. The negative correlation implies that for higher concentrations of the true pollutant \( x \), the bias \( \hat{x} - x \) will be negative and thus the regression-calibration estimate \( \hat{x} \) tends to underestimate the pollution concentration \( x \), as observed in Figure 1 and Table 1. This is problematic since we want to be sure not to miss high pollution events. Proposition 1 is a general result for linear models and shows that regression-calibration is inherently flawed for accurately identifying peaks in pollutant concentrations even when the underlying data generation process is same as the model fitted. The linear assumption is on the parameters and not on the functional form of the covariates, and hence the result can be valid even for some non-linear regression functions (e.g., the gain-offset model (1) includes interactions between the low-cost data and the covariates).

### 3.3 Inverse regression model

To address the underestimation issue, we investigate using an inverse model, where the true pollutant concentration is used as an independent variable and the response is the low-cost sensor pollutant reading:

\[
y(s, t) = \beta_0 + \beta_1 x(s, t) + \beta_2 z(s, t) + \beta_3 x(s, t) z(s, t) + \epsilon(s, t)
\]  

(4)

This is equivalent to the inverse gain-offset model

\[
E(y(s, t)) = o(s, t) + g(s, t) z(s, t).
\]  

(5)

with linear offset \( \beta_0 + \beta_2 z(s, t) \) and gain \( \beta_1 + \beta_3 z(s, t) \).

The inverse model is more interpretable as the gains and offsets can be viewed as biases of the low-cost data from the true pollutant concentrations. In fact, it is easy to see that the forward (1) and inverse (5) gain-offset models correspond to respectively the Berkson and classical measurement errors for the low-cost data (Fuller, 1987). In air pollution modeling, the Berkson error model is suitable when the observed values are aggregations over geographical areas (Zeger et al., 2000). However, in the case of low-cost data at co-located sites, there is no such aggregation involved. The classical error, which assumes that the observed low-cost measurements are more noisy than the underlying true pollutant concentrations at the same location, is a more appropriate representation.

We first show that the underestimation issue of regression-calibration persists even under the classical error model for the data generation process.

**Proposition 2.** Consider the classical error model (4) for the low-cost data, with i.i.d. errors and no other covariates. Then the bias \( \hat{x} - x \) of the predictions from a regression-calibration model (3) fitted to this data is asymptotically negatively correlated with the true pollutant concentration \( x \).

The proof is included in the Appendix 2. Propositions 1 and 2 prove that under both models of measurement error for the low-cost data, the regression-calibration residuals will be negatively correlated with the true pollutant concentration, leading to underestimation when the true concentration levels are high. The classical measurement error is
a more organic model for the low-cost data, as argued above. Hence, we propose using the inverse regression model (4) for fitting the low-cost data. Once the model is fit, for a given value $y$ of the low-cost data, we can predict the true pollutant concentrations by simply inverting the regression equation:

$$\hat{x} = \frac{y - \hat{\beta}_0 - \hat{\beta}_2'z}{\hat{\beta}_1 + \hat{\beta}_3'z}$$

(6)

We show that the inverse regression model does not suffer from the underestimation issue.

**Proposition 3.** Consider the classical error data generation process (4) for the low-cost pollutant measurements given the true pollutant values, and assume i.i.d. errors. Also, assume that the variables $x$ and $z$ are bounded, $\epsilon \sim N(0, \tau^2)$, $\beta_1 + \beta_3'z$ are bounded away from zero (i.e., there exists some $a > 0$ such that $P(|\beta_1 + \beta_3'z| > a) = 1$), and that $X'X/n$ converges in probability to a positive definite matrix, where $X$ is the matrix of independent variables. Then the bias of the predictions (6) from the inverse regression model is asymptotically uncorrelated with the true pollutant concentrations.

The proof is in Appendix . Thus, if the inverse model is fit, the residuals are uncorrelated with the value of the true pollutant concentration and predictions will not suffer from the overestimation issue when the concentration is high. The assumption of $\beta_1 + \beta_3'z$ lying away from zero is necessary since for covariates values lying on or near this hyperplane, (6) involves division by a near-zero quantity that can result in some predictions of unrealistically high magnitude. In practice, however, this assumption may be nearly violated, and in finite samples there can be instability in the predictions from this inverse model. In the next Section we discuss how to mitigate this issue in practice via spatial shrinkage by augmenting the inverse-regression model with a second-stage spatial model for the true pollutant surface to complete a spatial filtering algorithm.

### 3.4 Gaussian Process Filtering

The second major limitation of regression-calibration is that the spatial structure of air pollution is not used by the approach. Subsequent to training the regression model on the co-located data, the model calibrates the low-cost data at each site in the network independently. Leveraging the spatial structure in the low-cost data across the network sites can potentially improve quality of the calibration.

A related issue is the static nature of the calibration equation owing to not using concurrent reference data available. To estimate the regression-calibration model, of the reference instruments in Sets A and C (Figure 2), only the data from Set A (the co-located sites) is used for a fixed training window $W$. When calibrating the low-cost network data for a subsequent time $t$, data from the reference sites in set A or C will usually be available for that time. This data is not utilized in regression-calibration, despite carrying valuable information. The low-cost data at co-located or nearby network sites are likely to be correlated with this concurrent reference data and a dynamic calibra-
tion approach leveraging this time-varying correlation will better capture true pollutant concentrations.

We extend the inverse regression model to a novel a spatial filtering approach that accommodates both types of spatial correlation — among the low-cost data at different locations (sites A and B), and between the low-cost data and the reference data from sites A and C. We consider a two stage model. The first stage is the inverse regression model (4). Unlike the regression-calibration, the inverse regression has the true pollutant concentration $x$ as the independent variable. This allows a second-stage geospatial model for $x$ to capture the spatial correlation in true pollution concentrations. We propose a second-stage Gaussian process (GP) model for the pollutant concentrations $x_i(\cdot) \sim GP(\mu_t, C_t)$ where $x_t = \{x(s,t) : s \in D\}$ is the pollutant surface over the spatial domain $D$ at time $t$, $\mu_t$ is the surface mean, and $C_t$ is the GP covariance function such that $C_t(s_i, s_j) = Cov(x(s_i, t), x(s_j, t))$. GPs are widely used to model smooth spatial surfaces owing to the convenient representation of finite GP realizations as multivariate normal distributions which facilitates predictions at new locations (kriging) via simple conditional normal distributions. The mean function $\mu_t$ can be modeled using covariates if there are sufficient number of reference sites (sites A and C). Otherwise, it can simply be modeled as a time-specific constant, as we do here. We use the exponential covariance function where $C_t(s_i, s_j) = \sigma_t^2 \exp(-\phi_t||s_i - s_j||)$. However, any other choice like the Matérn or squared-exponential functions can also be used for $C_t$.

Let $S_B$ be the coordinates of the n non-co-located low-cost sites in Set B for which $y(s, t)$ is observed but $x(s, t)$ is not, and $S_{AUC}$ be the coordinates of the p sites in Sets A and C where $x(s, t)$ is known. Our goal is to infer on the true pollutant concentrations $x(S_B, t)$ in $S_B$ based on knowledge of the true pollutant concentrations $x(S_{AUC}, t)$ at $S_{AUC}$ and the low-cost data $y(S_B, t)$ at $S_B$. At each timepoint $t$, kriging using the GP model implies the following conditional distribution for the true pollutant $x(S_B, t)$ at the non-co-located sites $S_B$.

\[
x(S_B, t)|x(S_{AUC}, t) \sim N(\mu_t, \Sigma_t)
\]

\[
\mu_t = \mu_1 + C_{t, B, AUC}C_{t, AUC, AUC}^{-1}(x(S_{AUC}, t) - \mu_1)
\]

\[
\Sigma_t = C_{t, B, B} - C_{t, B, AUC}C_{t, AUC, AUC}^{-1}C_{t, AUC, B}
\]

where $C_{t, i,j} = Cov(x(s_i, t), x(s_j, t))$.

Equation (7) differs from the common geospatial models where the entire $x(s, t)$ surface is latent and an unconditional GP prior is used. In the setting of low-cost networks, the latent surface of true pollutant concentrations is partially observed at the reference sites $A \cup C$. Hence, (7) is a conditional GP prior for the unobserved part of the surface given the available knowledge of the surface from realizations at $A \cup C$.

Equations (4) and (7) complete the specification of our spatial filter to obtain predictions of true pollutant concentrations based on all available low-cost and reference data. Our two-stage model can be perceived as a spatial analog of Kalman-filtering (Kalman, 1960). In Kalman-filters or other filtering approaches, a stochastic process is observed at one or few time-points which dictates the evolution at a future time. This temporal evolution, based on partial realization of the stochastic process, is used to filter noisy observations at future time-points. In low-cost networks, at each time point $t$, the stochastic process
(pollutant surface) is over space. The low-cost data at the network sites $A \cup B$ are the noisy observations, and the reference data at the small set of locations $A \cup C$ are the partial realizations of the process that informs about the true pollutant concentrations at the network sites $B$ owing to spatial correlation in pollutant concentrations. Thus, equation (7) is the state-transition model dictating the spatial evolution of the partially observed surface $x(s, t)$, while Equation (4) is the observation model for the noisy low-cost data $y(s, t)$. Together, these equations form a filtering setup for calibration and smoothing of low-cost networks, where the quantity of interest is $x(S_B, t) | y(S_B, t), x(S_{A \cup C}, t)$.

To predict $x(S_B, t)$, we first write the observation model (4) for a vector of locations $y(S_B, t)$. Let $Z(S_B, t)$ be the $n \times n_{\text{cov}}$ matrix of covariates at time $t$ and $\tau^2$ be the variance of the normally distributed errors $\epsilon$. We then have the observation model

$$y(S_B, t) \sim N \left( \beta_0 1 + Z(S_B, t) \beta_2 + (\beta_1 I + \text{diag}(Z(S_B, t) \beta_3)) x(S_B, t), \tau^2 I \right).$$  \hspace{1cm} (8)

Note that during estimation of the observation model based on co-located data for a fixed time window $\mathcal{W}$, $x$ is known at the co-location sites $A$ and the unknown quantities are the parameters $\beta_i$'s and $\tau^2$. However, at the filtering stage at a time $t$, the pre-estimated parameters $\beta_i$'s and $\tau^2$ are now assumed known and the unknowns in Equation (8) are the true pollutant concentrations $x(S_B, t)$.

Letting $H(S_B, t) = \beta_1 I + \text{diag}(Z(S_B, t) \beta_3)$, we can rewrite the observation model as

$$y(S_B, t) \sim N \left( \beta_0 1 + Z(S_B, t) \beta_2 + H(S_B, t) x(S_B, t), \tau^2 I \right).$$

Now we transform the observations $y(s, t)$ to $u(s, t) = y(s, t) - \beta_0 - \beta_2 z(s, t)$ and get

$$u(S_B, t) = y(S_B, t) - \beta_0 1 - Z(S_B, t) \beta_2 \sim N \left( H(S_B, t) x(S_B, t), \tau^2 I \right).$$ \hspace{1cm} (9)

The transformed observations $u$ can be considered as the measurements in a Kalman filter model, with the observation model defined by $H(S_B, t)$. The Kalman filter equations can be fit to the two stage model given by (7) and (9) to get

$$\hat{x}(S_B, t) = \hat{\mu}_t \quad \text{(Predict)}$$

$$x_{\text{update}}(S_B, t) = \left( \Sigma_t^{-1} + \frac{1}{\tau^2} H(S_B, t)^2 \right)^{-1} \left( \Sigma_t^{-1} \hat{x}(S_B, t) + \frac{1}{\tau^2} H(S_B, t) u(S_B, t) \right) \quad \text{(Update)}$$ \hspace{1cm} (10)

where $\hat{\mu}_t$ and $\Sigma_t$ are defined as in (7). The schematic in Figure 3 summarizes the entire process. Our approach jointly predicts the pollutant concentration $x(S_B, t)$ at all the non-co-located sites given the known true pollutant values $x(S_{A \cup C}, t)$ at $S_{A \cup C}$ and the observed low-cost data $y(S_B, t)$ at $S_B$. An initial update (‘predict step’) of $x(S_B, t) | x(S_{A \cup C}, t)$ is dictated by the state-transition model, i.e., the conditional GP distribution in (7). The final update of $x(S_B, t)$ given the low-cost data $y(S_B, t)$ is analogous to the Kalman update step, and gives the network-wide calibrated and smoothed estimate of the pollutant surface. The filtering relies on pre-estimation of the observation model and the spatial parameters which will be discussed in Section 5.5.
Unlike regression-calibration, the resulting calibration equation (10) from spatial filtering is dynamic in nature. The Kalman-update (10) is a weighted sum of the kriging prediction (7) and the measurement from the inverse regression model (4), with the weights $\Sigma_t$ and $H_t$ being time-specific and estimated from the data. Thus the calibration is informed by the current concentrations of true pollutants at the reference sites.

The conditional GP model, using all available reference data, simultaneously incorporates the spatial correlation between the low-cost data and the reference data (via the kriging prediction $\tilde{\mu}_t$) and the correlation among the low-cost sites (via the kriging covariance $\Sigma_t$). This effectuates a spatially smooth estimate of $x(S_B, t)$ unlike regression-calibration which treats data from each site independently. Also, leveraging of the spatial information is essentially a spatial shrinkage method that mitigates the instability issue of the naive predictions from the inverse regression (6). Unstable predictions correspond to low-precision (near-zero diagonal entries in the $H$ matrix) and will be naturally downweighted in (10).

Finally, we note that in our approach, the filtering is over space at each timepoint, unlike most Kalman filter applications for spatio-temporal air-pollution data that filter over time. We discuss this difference in more details in Discussion.

### 3.4.1 Predicting on a grid of locations

The spatial filtering approach offers a coherent way to obtain joint predictions of the true pollutant concentrations at any arbitrary set of locations with neither reference or low-cost sensors (Set D in Figure 2) to create smooth maps of the pollutant concentrations in the area.

The joint posterior likelihood of the unknown true pollutant concentrations conditional on the observed data can be rewritten as:

$$ p(x(S_B), x(S_D)|x(S_A), x(S_C), y(S_B)) $$

$$ = p(x(S_D)|x(S_A), x(S_B), x(S_C), y(S_B)) \times p(x(S_B)|x(S_A), x(S_C), y(S_B)) $$

$$ = p(x(S_D)|x(S_A), x(S_B), x(S_C)) \times p(x(S_B)|x(S_A), x(S_C), y(S_B)) $$
where the time-index $t$ is omitted and the final equality comes from the fact that $y(S_B)$ and $x(S_D)$ are independent conditional on $x(S_B)$ as there are no observations in Set D. In this expression, the first term is a conditional normal distribution that can be obtained from kriging, since $x(S_B)$ is a Gaussian Process. The second term is the posterior normal distribution already available from the update step of the previous section (Equation 10). This facilitates straightforward prediction of $x(S_D)$ conditional on the observed data. We provide the details in the Supplement.

3.5 Implementation

The filtering update in Equation (10) assumes that the parameters of the observation model and the state-transition model are known. In practice, these parameters are unknown and will need to be estimated in addition to inferring the true values $x(S_B,t)$. The parameters $\beta_i$’s and $\tau^2$ of the observation model can be estimated over a training period $W$ where both $x(S_A,t)$ and $y(S_A,t)$ are measured at the co-located sites $S_A$. Least squares can be used to estimate the coefficients and the observation model variance. Since there is typically abundant co-located data consisting of hourly co-located time-series for several weeks to months, these parameters can be estimated with high precision and the estimates can be plugged into the filtering updates.

The parameters of the GP model, $\mu_t$, $\sigma^2_t$, and $\phi_t$, are allowed to be time-varying to capture dynamic spatial correlation in the air pollution surface. They need to be estimated at each timepoint $t$ using all available data $x(S_{A,C},t)$ and $y(S_B,t)$ for the time-point. As the total number of sites with either a reference or a low-cost sensor will be relatively small, these estimates may have high variability and the Kalman-updates may be sensitive on the decision to propagate or not propagate this parameter uncertainty. We explore the impact of this choice by offering both a frequentist and a Bayesian implementation of the filtering method.

3.5.1 Frequentist

Under the frequentist implementation of the filtering approach, the observation model can be used to predict an initial value of the true $x$ from the observed $y$ using Equation 6. With an exponential covariance function, we then have $\hat{x}_{init} \approx GP(\mu_t, (\sigma_t \exp(-\phi_t ||s_i - s_j||))_{i,j})$ where at the reference sites (A and C) $\hat{x}_{init}$ denotes the observed true $x$ and at the low-cost network sites (B) $\hat{x}_{init}$ denotes the initial predicted value of $x$ from Equation 6. Maximum likelihood can be used to estimate the spatial parameters from $\hat{x}_{init}(S_B,t)$. There are several packages in R that estimate the MLE of a Gaussian process. We use the SpatialTools package (French, 2018). Once the estimates $\hat{\mu}_t$, $\hat{\sigma}^2_t$, and $\hat{\phi}_t$ are obtained, Equation (10) can be used to estimate $\hat{x}(S_B,t)$ on Set B with the estimated covariance matrix and observation model, or Equation S2 of the Supplement can be used if predictions on both Sets B and D are desired. This method does not propagate the uncertainty of the spatial parameters in to the Kalman updates.
3.5.2 Bayesian

Our spatial filtering, like Kalman-filter or other filtering approaches, has an inherent Bayesian flavor as the update step in (3) can be viewed as the posterior mean of \( x(S_B, t) \mid x(S_A \cup C, t), y(S_B, t) \) given the conditional GP prior (7) and the low-cost observations modeled as (8). Augmenting these two equations with additional priors for the remaining spatial hyper-parameters, we can use a Bayesian model to jointly estimate \( \hat{x} \) and the GP model parameters, and propagate uncertainty in the spatial parameter estimation into the estimate of \( x \). We still estimate the observation model and plug in its parameters before the Bayesian estimation. As mentioned before, these parameters will be estimated with high-precision given abundant co-located data, hence the associated uncertainty is negligible. However, if desired, these parameters can also be estimated jointly in a fully Bayesian setup that propagates all uncertainty.

The main advantage of the Bayesian formulation is that the division in (6) is not explicitly performed unlike the frequentist implementation where it is used to obtain the initial estimate of \( x \). This means that the predictions that have highly inflated variances will not be directly used to estimate the spatial parameters, and are naturally down-weighted in the Bayesian framework owing to incorporation of the prediction uncertainty. The main disadvantage of the Bayesian framework is that it is more computationally expensive and time consuming than the frequentist implementation.

4 Simulation studies

We conduct several simulation studies to evaluate the performance of the proposed method compared to the original regression-calibration model. We consider settings of varying degrees of underlying spatial correlation for the true air pollution surface, different low-cost and reference network designs and sample sizes, as well as various forms of model misspecification. For each simulation experiment, the \( p \) co-located and \( n \) non-co-located sites are sampled from the unit square, 1,000 timepoints of pollutant concentrations are generated to estimate the observation/regression models, and the methods are evaluated on 100 future timepoints. At each timepoint, a vector \( x_t \) of true pollution concentrations at a set of locations is generated from a spatial model, whose specifications differ across simulations. The low-cost data are generated from an observation model with coefficient values obtained from fitting model (4) on the Baltimore SEARCH network \( \text{PM}_{2.5} \) data in August and September 2019. The observation model variance is chosen to be \( \tau^2 = 2 \). We generate data using an observation model with covariates, except when investigating the effect of covariate misspecification, i.e., of including extraneous covariates and of missing covariates. Covariates are sampled from the following distributions:

\[
RH \sim U(24, 76), T \sim U(17, 45), weekend \sim Bern\left(\frac{2}{7}\right), daylight \sim Bern\left(\frac{2}{3}\right)
\]

where \( U \) denotes the uniform distribution and \( Bern(r) \) denotes the Bernoulli distribution.

The ranges for temperature and relative humidity correspond to the respective ranges of...
them as measured by the SEARCH network in Baltimore in August and September of 2019.

The methods compared are the regression-calibration model (RegCal) and the Bayesian implementation of the Gaussian Process filtering method (GP Filter). To evaluate the performance of the method we use root mean square error (RMSE) and false negative rate (FNR) of AQI classification for PM$_{2.5}$ based on the latent true air pollution surface. FNR is defined as the proportion of observations where the true pollutant concentration at a site is a Moderate or Unhealthy AQI (PM$_{2.5}$ ≥ 12) but the prediction from the low-cost sensor data is a Good classification (PM$_{2.5}$ < 12). 50 replicate datasets are simulated for each simulation setting, and the results are averaged over these replicate datasets.

4.1 Simulation 1a: Correctly specified model

We begin by assuming a correctly specified model with 1 co-located site and 50 non-co-located sites (similar design as the Baltimore SEARCH network). We let $\phi = 3 \sqrt{2}$ and $\mu = 7$, and $\sigma^2 \in \{5, 10, 15, 20\}$. Different values of $\sigma^2$ are used to evaluate the performance of the model with different spatial noise to random noise ratios ($\sigma^2/\tau^2$) in the low-cost data.

The results of the correctly specified simulations are shown in Figure 4. The first panel shows the RMSE (averaged over 50 datasets) of the methods for this setting. For all values of $\sigma^2$, GP Filtering has $\sim 20\%$ lower RMSE than regression-calibration. As the spatial variance increases, the RMSE of both methods increases. This is expected as with higher spatial variability, the predictions at each site have higher uncertainty. The top right panel shows how the RMSE changes as a function of the distance from the reference instrument across all the datasets, for one choice of $\sigma^2 (= 15)$. We see that for the filtering approach the RMSE decreases for sensors closer to the reference site. This is because the filtering model, via use of the conditional GP (7), accounts for the underlying spatial correlation in the true pollution surface. So the closer sites will have predicted pollutant concentrations closer to the reference value. Even at further distances, the RMSE for the filtering method is substantially lower than that for the regression-calibration. The RMSE by distance plots for all other $\sigma^2$ values are included in the supplemental materials and reveal similar trends (Figure S1).

The bottom left panel plots the FNR of inaccurately classifying moderate or unhealthy AQI days as good. Once again, the GP Filter has consistently smaller FNR than RegCal. As $\sigma^2$ increases, the FNR decreases in all methods. Additionally, the bottom right panel shows that the prediction interval estimates for both methods have coverage probabilities close to 95%. However, the GP Filter has narrower interval widths than RegCal, showing that it offers improved precision for the predictions.

The GP Filter can also be implemented in a frequentist setting, where the spatial parameters are estimated and then considered fixed and Equation 10 is used directly, as discussed in Section 3.5. While the performance of the two implementations are very similar, the frequentist approach results in much lower computation time (Table 2) and
Figure 4: Results for setting 1a, averaged over 50 datasets: (Top Left) RMSE for a correctly specified model. Results for four $\sigma^2$ values are shown. (Top Right) RMSE by distance of site from the co-located reference site for $\sigma^2 = 15$. (Bottom Left) False negative rate. (Bottom Right) Coverage probability and average confidence interval width. 95% probability denoted by the black line.

might be the pragmatic choice for large datasets. However, low-cost networks typically do not have very large sample sizes so the computation time for the Bayesian model will not generally be excessive. Also, for larger networks, computations for the Bayesian implementation can be expedited by replacing the Gaussian Process priors by the scalable Nearest Neighbor Gaussian Processes (Datta et al., 2016; Finley et al., 2019).

4.2 Simulation 1b: Correctly specified model with more sensors

The previous simulation setup emulated the Baltimore SEARCH low-cost network design with around 50 low-cost sites and only one reference site. We conduct additional simulations with other network designs where we increase the numbers of both the low-cost sites and the reference sites. We first consider $p = 5$ reference sites with co-located low-cost sensors, and keep all other settings and parameter choices the same from setting 1a. Figure 5 shows the differences in model comparison metrics when increasing the number of reference sites from 1 to 5. We focus on the performance on the co-located sites in sites within 0.1 units of any additional reference site and the results are averages over 50 replicate datasets. We see that the GP filter benefits from more reference data. As
Figure 5: Results of setting 1b. (Top) Comparison of the change in performance of the methods when increasing the number of co-located sites from 1 to 5, with percent difference in RMSE in top-left and percent difference in FNR in top-right. (Bottom) Comparison of the change in performance of the methods when increasing the number of low-cost sensors from 50 to 200 non-co-located sites with percent difference in RMSE in bottom-left and percent difference in FNR in bottom-right.

the number of reference sites increases from 1 to 5, for sites near the reference sites, the RMSE of the filtering method decreases by 5 – 12% for all setups as shown in the top left panel. The FNR of the filtering method decreases by up to 10% in the sites around the reference sites, as shown in the top right panel. The RMSE or FNR of the regression calibration remains roughly unchanged with increase in number of reference sites, as the method does not leverage any of the increased spatial information available from more reference data.

We then consider a network with only 1 reference site for co-location but increase the number of non-co-located low-cost network sites from $n = 50$ to 200. We see that with an increase in density of deployment of the low-cost network, the RMSE of the filtering method consistently decreases by 10 – 18% (bottom left panel). This is because with more total sensors, the spatial parameters are better estimated (Figure S2). The FNR decreases by 10-13% when 200 non-co-located sensors are used (bottom right panel). Once again the regression-calibration does not benefit from increased density of low-cost sensors.
4.3 Simulation 2: Misspecified observation model

We now consider situations where the covariates used in the observation model are incorrectly specified. We consider both the case where the set of covariates is under or over specified. The results when \( \sigma^2 = 15 \) are shown in Figure 6. We see that for the setting of underparametrization, i.e., when covariates used in the true data generation mechanism is not used in fitting, the RMSE is expectedly higher for the misspecified model than the correctly specified one. However, the RMSE is still lower for the misspecified GP Filter method compared to the corresponding misspecified regression-calibration. With overparametrization, i.e., when there are no covariates in the true model but they are included in the model fitted, adding covariates results in only a tiny increase in RMSE. The RMSE and FNR for other \( \sigma^2 \) values are included in the supplemental materials (Figures S3, S4). The FNR also increases when true covariates are omitted from the observation model.

4.4 Simulation 3: Misspecified state-transition model

We now simulate data from a misspecified latent air pollution model. Instead of generating the true pollution surface from a Gaussian Process, we generate a fixed smooth spatial surface as follows. Two point sources are randomly selected in the unit square at locations \( s_{1}^{*}, s_{2}^{*} \). At each timepoint, the emission from each source \( (p_i(t), i = 1, 2) \) are sampled from a \( U(2, 9) \) distribution. We then let the PM\( _{2.5} \) surface be defined as a distance-based kernel weighted sum of the two emissions, i.e., \( x(s, t) = \sum_{i=1}^{2} p_i(t) \exp\{\|s - s_{i}^{*}\|^2/(2\gamma)\} \), where \( \gamma \) is a scale parameter for how slow the concentrations decay around the source. We use \( \gamma \in \{0.1, 0.4, 0.7, 1\} \). One example simulated pollutant surface is shown in Figure 7 (left).

The RMSE and FNR of these simulations are similar to the correctly specified model and are included in the Supplement (Figures S5, S6). The GP Filter performs substantially
better than the regression-calibration, even when the true data is not generated using a GP. This is not surprising as GP are widely used as a non-parametric technique to estimate smooth functions or surfaces and have established theoretical guarantees about accurate surface estimation (Choi and Schervish, 2005; van der Vaart et al., 2008). Figure 7 shows maps of the true pollution surface and the predictions from the two methods in one dataset at a single timepoint. The GP Filter captures the peaks of the true pollution surface, whereas the regression-calibration method underestimates most of the higher pollutant concentrations. We also see that the RMSE over all 100 validation timepoints in that dataset is significantly lower for the GP Filter (0.61) than the regression-calibration method (1.23). Therefore, even in the case of a misspecified model, the GP Filter is a considerable improvement over regression-calibration both in overall RMSE and for capturing peaks in air pollution.

5 Analysis of Baltimore SEARCH low-cost PM$_{2.5}$ network data

The SEARCH network in Baltimore had sensors at 34 locations in December 2019, which are shown in Figure 1. The PM$_{2.5}$ sensor is a Plantower A003, which has been used in many low-cost networks (Feenstra et al., 2019; Malings et al., 2020; Magi et al., 2020). These are optical sensors that measure pollution concentrations based on the scattering of light. The measurements from these types of sensors are known to depend on meteorological conditions, such as relative humidity and temperature. There are two sensors
co-located with a reference device at the Oldtown location, shown in blue. The reference device is an FEM Beta Attenuation Monitor (BAM) that measures hourly PM$_{2.5}$.

We compare the performance of the GP Filter and the regression-calibration used in Datta et al. (2020) on PM$_{2.5}$ data from the SEARCH network (RegCal). The same set of covariates are used as considered previously, i.e., RH, T, a weekend indicator, and a daylight indicator. The regression-calibration model and the observation model of the filtering method are trained on 614 hourly observations from November 2019. The methods are tested on observations from December 2019. Since filtering uses the latest available reference data from the co-location site (Oldtown), the Oldtown data cannot be used for evaluation. Instead, we test the models using data from the MDE device at Essex, the purple site in Figure 1, which is not used for training any model. The Essex site collects PM$_{2.5}$ measurements every 6 days, so 5 observations are available for testing in December. At both the Oldtown and Essex sites, there are two low-cost sensors at the same location, data from which are averaged to create a single low-cost time-series for each location.

We use the 5 daily observations from the Essex site to assess the calibration methods. The RMSE of the two methods are shown in Figure 8a. The RMSE of the GP Filter is much lower than the regression-calibration. For a more in-depth understanding of the performance, we also look at the predictions from the methods at each individual date in Figure 8b. The figure shows that the GP Filter prediction is always closer to the true Essex value, and the prediction intervals always capture the true values. The baseline PM$_{2.5}$ concentrations are better estimated by the GP Filter. The regression-calibration point estimate considerably underestimates true PM$_{2.5}$ concentration on December 23, 2019 which was in the unhealthy AQI range. The filtering method does not suffer from such substantial underprediction. The prediction intervals from regression-calibration are wider than the GP Filter intervals, showing higher uncertainty about the true PM$_{2.5}$ concentration. Despite this, the intervals do not always capture the true value.

Figure 9 shows maps of the predicted PM$_{2.5}$ surfaces on December 23, 2019. On this day, the Oldtown and Essex MDE PM$_{2.5}$ concentrations both correspond to an unhealthy AQI, and the regression-calibration has much lower estimates than the GP Filter methods over the entire city. The map produced by GP Filter matches the observed PM$_{2.5}$ at Oldtown and Essex more closely than the RegCal.

To further investigate the underprediction issue, we look at the predictions from the two methods at the SEARCH site closest to the Oldtown reference instrument, so that the reference time series from Oldtown can be used to approximate the true PM$_{2.5}$ time series at that location. We use this site instead of the reference site of Oldtown itself for comparison since the prediction from the GP Filter at Oldtown agrees exactly with the known true PM$_{2.5}$ concentration measured at that location, due to the exact interpolation property of kriging. Figure 10 (left) shows that the filtering method results in predictions that are much closer to the (nearby) reference, especially where the reference concentrations spike around December 23.

A pseudo-residual plot is used for model diagnostics in Figure 10 (right), where the true PM$_{2.5}$ concentration used for all sensors is the Oldtown concentration since that is the only site where the true PM$_{2.5}$ concentration is observed. Only sensors within
Figure 8: Out-of-sample comparison of GP filtering and regression-calibration for the SEARCH low-cost network: (Left) RMSE at the Essex site in December 2019. (Right) Predictions and 95% confidence intervals at Essex. The horizontal black lines are the true measurements each day.

Figure 9: Interpolated maps of predictions across SEARCH network on December 23. The first two panels show the prediction surface from the GP Filter and RegCal. The last two panels show the prediction surface if the measurement at the Essex or Oldtown reference device is assumed to be the PM$_{2.5}$ concentration across the city. The two squares denote the reference measurements at the Oldtown (center) and Essex (east) sites. The circles denote the calibrated low-cost PM$_{2.5}$ data from each method at the sites in the SEARCH network.
Figure 10: (Left) Daily time series in December 2019. The GP Filter and RegCal predictions are at the closest sensor to the reference site at Oldtown. The reference time series is the PM$_{2.5}$ recorded by the reference-grade BAM at the Oldtown site. (Right) Pseudo-residuals (prediction – Oldtown) for SEARCH sensors within 5km of the Oldtown MDE device at timepoints where the Oldtown PM$_{2.5}$ AQI level is moderate or unhealthy (>12µg/m$^3$).

5km of Oldtown are considered so that the Oldtown reference measurement is a reasonable substitute for the true value. The residuals from the GP filter generally does not exhibit strong correlation with the true pollutant concentrations except when the true concentrations are very high (∼50µg/m$^3$) where there is some underestimation. The RegCal residuals exhibit strong negative correlation with the true pollution concentrations throughout its range. The residuals are negative for moderate or unhealthy PM$_{2.5}$ concentrations, thereby considerably underestimating them. This behavior of systematic underestimation is consistent with the theoretical results of Propositions 2 and 1.

6 Discussion

The promise of low-cost air pollution networks is indisputable, owing to their cost-efficacy and spatio-temporal richness of their output. However, low-cost air pollution data can be highly biased and variable and any responsible use of the data in scientific studies mandates thorough quality control and evaluation of the data. While field-calibration using regression based on co-located reference devices has become the state-of-the-art for low-cost data correction, we show that this practice does not pay attention to a) accuracy of predicting high pollution events which leads to underestimating peaks in air pollution, and b) the spatial correlation in pollution concentrations across an area.

We present a simple but novel dynamic calibration approach via a spatial filtering that uses inverse-regression to mitigate the under-estimation issue and a conditional Gaussian Process model that leverages spatial correlation of the low-cost data across the network sites as well as correlation with the latest available reference data in the area. Our filtering approach works with as few as one co-location site. Our simulations showed that even a network of 50 non-co-located sites and 1 co-located site can provide enough spatial information to make better joint calibration and predictions than if each sensor is considered individually.
In fact, across all the simulation settings and in the data analysis on the SEARCH network, it is evident that both implementations of the Gaussian process filtering method significantly outperform a linear regression-calibration method. The RMSE and the FNR for identifying high pollutant concentrations are both lower in the GP Filter, and the method is robust to various forms of misspecification.

Our spatial filtering method filters in space and is notably different from existing filtering approaches for spatio-temporal data like the spatio-temporal filter or kriged Kalman-filter (Mardia et al., 1998; Sahu and Mardia, 2005), all of which filter in time. The high-frequency low-cost sensor data offer the opportunity to characterize ultra-short-term fluctuations of the pollutant concentrations, and filtering in time using lower frequency reference measurements will smooth these out, leading to loss of information. On the other hand, filtering across space is necessary to smoothly interpolate the data beyond the network locations to create continuous pollutant maps. The available reference data dictates the state-transition model (7) across space by modeling the spatial correlation. Thus the unique setting of co-located calibration mandates filtering in space instead of time.

Recently (Zheng et al., 2019) used GPs to smooth and calibrate data from a low-cost network. However, their calibration equation still used a forward regression with the low-cost data as the independent variable and the true concentrations as the response. Hence this does not mitigate the underestimation issue and implicitly assumes that the true pollutant surface to be noisier than the low-cost surface. Also, the approach required the presence of many (∼20) reference instruments in the area to capture the spatial structure to estimate device-specific calibration equations. This is unrealistic in many applications (e.g., Baltimore has only one continuous reference PM$_{2.5}$ measurement).

Our spatial filtering approach is more parsimonious in terms of resource needs and can be applied with as few as one reference site in the region which facilitates both training of the observation model as well as positing the conditional GP model for spatial smoothing. It is important to note that both the regression-calibration and our model assume that the relationship between the true pollutant concentration and the low-cost measurement is the same across all sensors, so the same observation model trained at one site can be used across the network. This assumption may not be valid in all gas sensor models, where there is a large amount of unit-to-unit variability. This unit-specific effect is not included in our model, so it may not be directly applicable to gas sensor networks. A calibration approach tailored to gas measurements has been proposed (Kim et al., 2018). This method uses gas cross-sensitivities, regional concentrations measured by reference monitors (which need not be co-located), co-emitted gases, ozone uniformity over space, and chemical conservation equations to calibrate a low-cost network of gas sensors. This approach takes the chemistry of pollutants into account and calibrates each one according to its particular behavior, but does not impose an explicit spatial model for air pollutant as our method does or investigate underprediction of high concentrations.

We identify multiple possible extensions of this method that can be considered for future research. Currently, a Gaussian error term is assumed. Generalizations of the Kalman filter that allow for non-Gaussian distributions of the error (Wüthrich et al., 2016) can be incorporated into the method. Also, the gains and offsets used in the observation model
are modeled as linear functions of the covariates. Non-linear calibration models have also been considered for calibration of low-cost networks (Topalović et al., 2019; Lim et al., 2019; Zimmerman et al., 2018; Johnson et al., 2018). More flexible non-linear observation models can be developed for the filtering to potentially improve fitting complex variable relationships.

Proofs of Propositions

In the Appendix, Propositions 1 and 3 are only proved in the case without covariates. When covariates are present, the proofs of these propositions are more technical and are provided in the supplementary materials file.

.1 Proposition 1 Proof

Proof. We first prove the result for the case without covariates (model (2)). Consider data \((y_i, x_i), i = 1, \ldots, n\). If the true model is \(x_i = \beta_0 + \beta_1 y_i + \epsilon_i\), and linear regression is used to estimate coefficients \(\hat{\beta}_0, \hat{\beta}_1\), then

\[
\hat{\text{Cov}}(\hat{x} - x, x) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \sum_{i=1}^{n} x_i^2 \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 y_i) (\beta_0 + \beta_1 y_i + \epsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 y_i) (\beta_0 + \beta_1 y_i + \epsilon_i)
\]

\[
= \beta_1 \hat{\beta}_1 \left( \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \bar{y}^2 \right) + \hat{\beta}_1 \left( \frac{1}{n} \sum_{i=1}^{n} y_i \epsilon_i - \bar{y} \bar{\epsilon} \right)
\]

where \(s_x^2, s_y^2\) are the sample variance of \(x\) and and \(y\) and \(s_{y,\epsilon}\) is the sample covariance of \(y\) and \(\epsilon\). This quantity asymptotes to \(\beta_1^2 \text{Var}(Y) + \beta_1 \text{Cov}(Y, \epsilon) - \text{Var}(X)\) by convergence of sample variances and covariances to their population analogs, consistency of \(\hat{\beta}_1\) to \(\beta_1\), and Slutsky’s theorem. Finally, noting that \(\hat{\beta}_1 = \text{Cov}(X, Y) / \text{Var}(Y)\) and \(\text{Cov}(Y, \epsilon) = 0\) we have

\[
\lim \hat{\text{Cov}}(\hat{x} - x, x) = \frac{\text{Cov}(X, Y)^2}{\text{Var}(Y)} - \text{Var}(X) < 0
\]

by the Cauchy-Schwartz inequality. Therefore, the bias \(\hat{x} - x\) is negatively correlated with the true response \(x\).

The general case with covariates in the model is proved in the supplement. 

\[\square\]
.2 Proposition 2 Proof

Proof. Let \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \) be the true classical error model and we fit the regression-calibration model \( x_i = \hat{\alpha}_0 + \hat{\alpha}_1 y_i \) using least squares. Then

\begin{align*}
\hat{\text{Cov}}(\hat{x} - x, x) &= \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\hat{\alpha}_0 + \hat{\alpha}_1 y_i) x_i - (\hat{\alpha}_0 + \hat{\alpha}_1 \bar{y}) \bar{x} - s_x^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (\hat{\alpha}_0 + \hat{\alpha}_1 (\beta_0 + \beta_1 x_i + \epsilon_i)) x_i - (\hat{\alpha}_0 + \hat{\alpha}_1 (\beta_0 + \beta_1 \bar{x} + \bar{\epsilon})) \bar{x} - s_x^2 \\
&= \hat{\alpha}_1 \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \hat{\alpha}_1 \frac{1}{n} \sum_{i=1}^{n} x_i \epsilon_i - (\hat{\alpha}_1 \beta_1 \bar{x}^2 + \hat{\alpha}_1 \bar{x} \bar{\epsilon}) - s_x^2 \\
&= \frac{s_{xy}}{s_y^2} \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 + \frac{s_{xy}}{s_y^2} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \epsilon_i - \bar{x} \bar{\epsilon} \right) - s_x^2 \\
&= \frac{s_{xy}}{s_y^2} \beta_1 s_x^2 + \frac{s_{xy}}{s_y^2} s_{x,\epsilon} - s_x^2 \\
\rightarrow \quad &\frac{\text{Cov}(X,Y)}{\text{Var}(Y)} \beta_1 \text{Var}(X) + \frac{\text{Cov}(X,Y)}{\text{Var}(Y)} \text{Cov}(X,\epsilon) - \text{Var}(X) \\
\text{by consistency and Slutsky’s} \\
= \frac{\text{Cov}(X,Y)^2}{\text{Var}(Y)} - \text{Var}(X) \quad \text{since } \beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \text{ and } \text{Cov}(X,\epsilon) = 0 \\
< 0 \quad \text{by the Cauchy-Schwartz inequality.}
\end{align*}

\( \square \)

.3 Proposition 3 Proof

Proof. We first consider a true model with no covariates, \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \) where we only need the assumption \( \beta_1 \neq 0 \). If the estimated coefficients from least squares are
\[ \hat{\beta}_0, \hat{\beta}_1, \text{ then} \]

\[ \text{Cov}(\hat{x} - \mathbf{x}, \mathbf{x}) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i) x_i - \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i) \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(\beta_0 + \beta_1 x_i + \epsilon_i) - \hat{\beta}_0}{\hat{\beta}_1} - x_i \right) x_i - \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(\beta_0 + \beta_1 x_i + \epsilon_i) - \hat{\beta}_0}{\hat{\beta}_1} - x_i \right) \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ = \frac{1}{\hat{\beta}_1} \left( (\beta_1 - \hat{\beta}_1) \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\beta_1 - \hat{\beta}_1) \bar{x}^2 + \frac{1}{n} \sum_{i=1}^{n} x_i \epsilon_i - \bar{\epsilon} \bar{x} \right) \]

\[ \rightarrow 0. \]

The limiting result holds by by consistency of the least squares estimates and since the errors \( \epsilon \) have \( E[\epsilon] = 0 \) and \( X \perp \epsilon \).

The general case of a model with covariates is proved in the supplement.

\[ \square \]

Acknowledgments

The authors would like to thank Colby Buehler (Yale) and Misti Levy Zamora (U. Conn.) for their contributions to the SEARCH network deployment.

Funding

CH was partially supported by the Fonds de recherche du Québec - Nature et Technologies bourse de maîtrise B1X, and partially supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2139757. AD, RP, and KK were partially supported by National Institute of Environmental Health Sciences (NIEHS) grant R01 ES033739. AD was partially supported by National Science Foundation (NSF) Division of Mathematical Sciences grant DMS-1915803. KK, DRG, RP, AD and CH acknowledge support from the assistance agreement no. RD835871 awarded by the U.S. Environmental Protection Agency to Yale University. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the agency. The EPA does not endorse any products or commercial services mentioned in this publication. DRG acknowledges HKF Technology (a Kindwell Company) for also supporting the sensor development.

References

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C., and Hamburg, S. P. (2017). High-resolution air pollution mapping with Google street view cars: exploiting big data. *Environmental science & technology*, 51(12):6999–7008.
Ardon-Dryer, K., Dryer, Y., Williams, J. N., and Moghimi, N. (2020). Measurements of pm 2.5 with purpleair under atmospheric conditions. *Atmospheric Measurement Techniques*, 13(10):5441–5458.

Balzano, L. and Nowak, R. (2007). Blind Calibration of Sensor Networks.

Barkjohn, K. K., Gantt, B., and Clements, A. L. (2021). Development and application of a united states-wide correction for pm 2.5 data collected with the purpleair sensor. *Atmospheric Measurement Techniques*, 14(6):4617–4637.

Bi, J., Wildani, A., Chang, H. H., and Liu, Y. (2020). Incorporating low-cost sensor measurements into high-resolution pm2. 5 modeling at a large spatial scale. *Environmental science & technology*, 54(4):2152–2162.

Bigi, A., Mueller, M., Grange, S. K., Ghermandi, G., and Hueglin, C. (2018). Performance of no, no 2 low cost sensors and three calibration approaches within a real world application. *Atmospheric Measurement Techniques*, 11(6):3717–3735.

Buehler, C., Xiong, F., Zamora, M. L., Skog, K. M., Koorman-Glaser, J., Colton, S., McNamara, M., Ryan, K., Redlich, C., Bartos, M., et al. (2021). Stationary and portable multipollutant monitors for high-spatiotemporal-resolution air quality studies including online calibration. *Atmospheric Measurement Techniques*, 14(2):995–1013.

Chadwick, E., Le, K., Pei, Z., Sayahi, T., Rapp, C., Butterfield, A., and Kelly, K. (2021). Understanding the effect of COVID-19 on particle pollution using a low-cost sensor network. *Journal of aerosol science*, 155:105766.

Choi, T. and Schervish, M. J. (2005). Posterior consistency in nonparametric regression problems under gaussian process priors.

Clougherty, J. E., Kheirbek, I., Eisl, H. M., Ross, Z., Pezeshki, G., Gorczyński, J. E., Johnson, S., Markowitz, S., Kass, D., and Matte, T. (2013). Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York City Community Air Survey (NYCCAS). *Journal of exposure science & environmental epidemiology*, 23(3):232–240.

Considine, E. M., Reid, C. E., Ogletree, M. R., and Dye, T. (2021). Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network. *Environmental Pollution*, 268:115833.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association Theory and Methods*, 111(514):800–812.

Datta, A., Saha, A., Zamora, M. L., Buehler, C., Hao, L., Xiong, F., Gentner, D. R., and Koehler, K. (2020). Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. *Atmospheric Environment*, 242:117761.

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Der Boghossian, B., Cocker, D., and Polidori, A. (2019). Performance evaluation of twelve low-cost pm2. 5 sensors at an ambient air monitoring site. *Atmospheric Environment*, 216:116946.
Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, pages 1–14.

French, J. (2018). *SpatialTools: Tools for Spatial Data Analysis*. R package version 1.0.4.

Fuller, W. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Johnson, N. E., Bonczak, B., and Kontokosta, C. E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmospheric Environment*, 184:9–16.

June, N., Vaughan, J., Lee, Y., and Lamb, B. K. (2021). Operational bias correction for PM2.5 using the AIRPACT air quality forecast system in the Pacific Northwest. *Journal of the Air & Waste Management Association*, 71(4):515–527.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Kim, J., Shusterman, A. A., Lieschke, K. J., Newman, C., and Cohen, R. C. (2018). The Berkeley atmospheric CO2 observation network: Field calibration and evaluation of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 11(4):1937–1946.

Larson, T., Henderson, S. B., and Brauer, M. (2009). Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. *Environmental science & technology*, 43(13):4672–4678.

Lim, C. C., Kim, H., Vilcassim, M. R., Thurston, G. D., Gordon, T., Chen, L.-C., Lee, K., Heimbinder, M., and Kim, S.-Y. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment international*, 131:105022.

Lu, Y. (2021). Beyond air pollution at home: Assessment of personal exposure to PM2.5 using activity-based travel demand model and low-cost air sensor network data. *Environmental Research*, page 111549.

Magi, B. I., Cupini, C., Francis, J., Green, M., and Hauser, C. (2020). Evaluation of pm2.5 measured in an urban setting using a low-cost optical particle counter and a federal equivalent method beta attenuation monitor. *Aerosol Science and Technology*, 54(2):147–159.

Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Robinson, A. L., Presto, A. A., and Subramanian, R. (2020). Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Science and Technology*, 54(2):160–174.

Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998). The kriged kalman filter. *Test*, 7(2):217–282.

Miskell, G., Salmond, J. A., and Williams, D. E. (2018). A solution to the Problem of Calibration of Low-Cost Air Quality Measurement Sensors in Networks. *ACS Sensors*. 
Nordio, F., Kloog, I., Coull, B. A., Chudnovsky, A., Grillo, P., Bertazzi, P. A., Baccarelli, A. A., and Schwartz, J. (2013). Estimating spatio-temporal resolved PM10 aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy. *Atmospheric environment*, 74:227–236.

Romero, Y., Velásquez, R. M. A., and Noel, J. (2020). Development of a multiple regression model to calibrate a low-cost sensor considering reference measurements and meteorological parameters. *Environmental Monitoring and Assessment*, 192(8):1–11.

Sahu, S. K. and Mardia, K. V. (2005). A bayesian kriged kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):223–244.

Si, M., Xiong, Y., Du, S., and Du, K. (2020). Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. *Atmospheric Measurement Techniques*, 13(4):1693–1707.

Tang, X., Zhu, J., Wang, Z., Wang, M., Gbaguidi, A., Li, J., Shao, M., Tang, G., and Ji, D. (2013). Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter. *Atmospheric Environment*, 81:676–686.

Topalović, D. B., Davidović, M. D., Jovanović, M., Bartonova, A., Ristovski, Z., and Jovašević-Stojanović, M. (2019). In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches. *Atmospheric Environment*.

United States Environmental Protection Agency (2021a). Air Data: Air Quality Data Collected at Outdoor Monitors Across the US.

United States Environmental Protection Agency (2021b). Evaluation of Emerging Air Sensor Performance.

van der Vaart, A. W., van Zanten, J. H., et al. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.

Van der Wal, J. and Janssen, L. (2000). Analysis of spatial and temporal variations of PM10 concentrations in the Netherlands using Kalman filtering. *Atmospheric Environment*, 34(22):3675–3687.

Wu, H., Tang, X., Wang, Z., Wu, L., Li, J., Wang, W., Yang, W., and Zhu, J. (2020). High-spatiotemporal-resolution inverse estimation of CO and NOx emission reductions during emission control periods with a modified ensemble Kalman filter. *Atmospheric Environment*, 236:117631.

Wüthrich, M., Cifuentes, C. G., Trimpe, S., Meier, F., Bohg, J., Issac, J., and Schaal, S. (2016). Robust gaussian filtering using a pseudo measurement. In *2016 American Control Conference (ACC)*, pages 3606–3613. IEEE.
Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental health perspectives*, 108(5):419–426.

Zheng, T., Bergin, M. H., Sutaria, R., Tripathi, S. N., Caldow, R., and Carlson, D. E. (2019). Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in delhi. *Atmospheric Measurement Techniques*, 12(9).

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313.
Supplement to “Gaussian Process filtering for calibration of low-cost air-pollution sensor network data”

Claire Heffernan, Roger Peng, Drew R. Gentner, Kirsten Koehler, Abhirup Datta

S1 Supplemental proofs

S1.1 Proposition 1 Proof with covariates

We proved proposition 1 in the case where the regression-calibration model does not contain covariates in the appendix. Now, we consider the case with covariates in the model.

Proof. For the general case, i.e., when the regression-calibration model contains covariates, as in Equation 3, the true model can be written as $x_i = \beta_0 + \beta_1 y_i + \beta_2 z_i + \varepsilon_i$, where $v_i = (1, z_i')$ and $\beta_0 = (\beta_0, \beta_2')$, $\beta_1 = (\beta_1, \beta_2')$.

Assume that the covariates $v_i$ and the measured pollutant concentrations $y_i$ are bounded above. Also assume $\text{Var}(\varepsilon) < \infty$. Then

$$
\hat{\text{Cov}}(\hat{x} - x, x) = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)x_i - \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)\frac{1}{n} \sum_{i=1}^{n} x_i
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} (v_i'\hat{\beta}_0 + v_i'\hat{\beta}_1 y_i - v_i'\beta_0 - v_i'\beta_1 y_i - \varepsilon_i)(v_i'\beta_0 + v_i'\beta_1 y_i + \varepsilon_i)
$$

$$
- \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)\frac{1}{n} \sum_{i=1}^{n} x_i
$$

(\ast) = \frac{1}{n} \sum_{i=1}^{n} \left( v_i' (\hat{\beta}_0 - \beta_0) \beta_0' v_i + v_i' (\hat{\beta}_0 - \beta_0) \beta_1' v_i y_i + v_i' (\hat{\beta}_0 - \beta_0) \varepsilon_i 
\right.

$$
+ v_i' (\hat{\beta}_1 - \beta_1) \beta_0' v_i y_i + v_i' (\hat{\beta}_1 - \beta_1) \beta_1' v_i y_i^2 + v_i' (\hat{\beta}_1 - \beta_1) y_i \varepsilon_i
$$

$$
- v_i' \beta_0 \varepsilon_i - v_i' \beta_1 y_i \varepsilon_i - \varepsilon_i^2 
\right) - \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)\frac{1}{n} \sum_{i=1}^{n} x_i
$$

We consider each term individually.
Terms 1-6: We can rewrite all six terms of the sum by defining a scalar $m_{i,j}$ for each one:

\[
\begin{align*}
&v_i'(\beta_0 - \beta_0)\beta_0 v_i = v_i'(\beta_0 - \beta_0)m_{i,1} \quad \text{letting } m_{i,1} = \beta_0 v_i, \\
v_i'(\beta_0 - \beta_0)\beta_1 v_i y_i = v_i'(\beta_0 - \beta_0)m_{i,2} \quad \text{letting } m_{i,2} = \beta_1 v_i y_i, \\
v_i'(\beta_0 - \beta_0)\epsilon_i = v_i'(\beta_0 - \beta_0)m_{i,3} \quad \text{letting } m_{i,3} = \epsilon_i, \\
v_i'(\beta_1 - \beta_0)\beta_0 v_i y_i = v_i'(\beta_1 - \beta_1)m_{i,4} \quad \text{letting } m_{i,4} = \beta_0 v_i y_i, \\
v_i'(\beta_1 - \beta_1)\beta_1 v_i y_i^2 = v_i'(\beta_1 - \beta_1)m_{i,5} \quad \text{letting } m_{i,5} = \beta_1 v_i y_i^2, \\
v_i'(\beta_1 - \beta_1)y_i \epsilon_i = v_i'(\beta_1 - \beta_1)m_{i,6} \quad \text{letting } m_{i,6} = y_i \epsilon_i.
\end{align*}
\]

Then we note that for \( j \in \{1, 2, 3, 4, 5, 6\} \) and \( k \in \{0, 1\} \):

\[
\left| \frac{1}{n} \sum_{i=1}^{n} v_i'(\hat{\beta}_k - \beta_k)m_{i,j} \right| \\
\leq \frac{1}{n} \sum_{i=1}^{n} |v_i'(\hat{\beta}_k - \beta_k)| \cdot |m_{i,j}| \quad \text{by the triangle inequality} \\
\leq \left( \frac{1}{n} \sum_{i=1}^{n} \|v_i||m_{i,j}\right) \cdot ||\hat{\beta}_k - \beta_k|| \quad \text{by the Cauchy-Schwartz inequality} \\
\xrightarrow{x} E[\|\textbf{V}\| \cdot |M_j| \cdot \|\hat{\beta}_k - \beta_k\|] \quad \text{by the weak law of large numbers and Slutsky’s}
\]

For \( j = 3, 6, E[\|\textbf{V}\| \cdot |M_j|] = E[\|\textbf{V}\||\hat{\epsilon}|^j = 6\}] E[|\epsilon|] < \infty \) since all \( \textbf{V}, Y \) are bounded and \( \text{Var}(\epsilon) < \infty \) implies \( E[|\epsilon|] < \infty \). For all other \( j, E[\|\textbf{V}\| \cdot |M_j|] < \infty \) since all terms are bounded.

Therefore, for all \( j, E[\|\textbf{V}\| \cdot |M_j| \cdot \|\hat{\beta}_k - \beta_k\|] = 0 \) since \( \hat{\beta}_k \) is consistent. So all six terms converge in probability to 0.

Terms 7-8: We also rewrite these two terms:

\[
\begin{align*}
v_i'\beta_0 \epsilon_i &= m_{i,7} \epsilon_i \quad \text{letting } m_{i,7} = v_i'\beta_0, \\
v_i'\beta_1 y_i \epsilon_i &= m_{i,8} \epsilon_i \quad \text{letting } m_{i,8} = v_i'\beta_1 y_i.
\end{align*}
\]

And we see that for \( j \in \{7, 8\} \):

\[
\frac{1}{n} \sum_{i=1}^{n} m_{i,j} \epsilon_i \\
\xrightarrow{x} E[M_j \epsilon] \quad \text{by the weak law of large numbers} \\
= E[M_j] E[\epsilon] \quad \text{since the error } \epsilon \text{ is independent of the independent variables } \textbf{V}, Y \\
= 0 \quad \text{since } E\epsilon = 0 \text{ and } M_j \text{ are bounded}
\]

So these two terms also converge in probability to 0.
Term 9:
\[
\frac{1}{n} \sum_{i=1}^{n} -\epsilon_i^2 \xrightarrow{P} E[-\epsilon^2] \quad \text{by the weak law of large numbers}
\]
< 0 since \( E[\epsilon^2] = Var(\epsilon) > 0 \)

Term 10:
\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i) \frac{1}{n} \sum_{i=1}^{n} x_i = 0
\]
as \( \sum \hat{x}_i = \sum x_i \) for a linear regression with an intercept.

Combining the results of all 10 terms with Slutsky’s theorem, we see that
\[
(*) \xrightarrow{P} E[-\epsilon^2] < 0
\]

Therefore, the covariance between the bias and the true pollutant concentration is asymptotically negative.

\[ \square \]

S1.2 Proposition 3 Proof with Covariates

We proved proposition 3 in the case where the inverse-regression model does not contain covariates in the appendix. Now, we consider the case with covariates in the model.

**Proof.** In the case of a model with covariates, the inverse model can be written as \( y_i = \alpha_0 z_{a,i} + \alpha_1' z_{a,i} x_i + \epsilon_i \) where \( \alpha_0 = (\beta_0', \beta_2')', \alpha_1 = (\beta_1, \beta_3')', \) and \( z_{a,i}(s, t) = (1, z_i')' \).

Assume that the true pollutant concentrations \( x \) and covariates \( z_k \) are bounded below and above and that \( \epsilon \) is normally distributed with \( Var(\epsilon) < \infty \). We also assume that \( |\alpha_1' z_a| > a \) for some \( a > 0 \), so that division by this quantity does not result in overinflated predictions. Lastly, we assume that \( X'X/n \xrightarrow{P} C \), for a positive definite \( C \), where \( X \) is the matrix of independent variables \( z_{a,i} \) and \( z_{a,i} x_i \).

The final assumption means that \( Var(\alpha) = Var(\epsilon)(X'X)^{-1} = Var(\epsilon)(X'X/n)^{-1}/n \xrightarrow{P} Var(\epsilon)C^{-1}/n \), where the inversion is well defined since \( C \) is positive definite. Therefore for each term, \( Var(\alpha_1) = O(1/n) \) for a finite \( C_d \).

We first state two lemmas:

**Lemma 1:** \( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|\alpha_1' z_{a,i}|^2} \) is bounded in probability.

**Proof.** Let \( \epsilon > 0 \). We will show that there exist some \( N \) such that for \( n \geq N \),
\[ P\left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|z_{a,i}|^2} > \frac{4}{a^2} \right) < \epsilon. \] We have

\[ P\left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|z_{a,i}|^2} > \frac{4}{a^2} \right) \leq P\left( \bigcup \left\{ \frac{1}{|\hat{\alpha}'_{a,i}|^2} > \frac{4}{a^2} \right\} \right) \text{ since at least one term must be larger than the average}\]

\[ \leq \sum P\left( \frac{1}{|\hat{\alpha}'_{a,i}|} > \frac{2}{a} \right) \text{ by sub-additivity}\]

\[ = \sum P\left( \frac{1}{|\hat{\alpha}'_{a,i}|} < \frac{a}{2} \right) \leq \sum P\left( \|\alpha'_{a,i} - \alpha_{a,i}\| < \frac{a}{2} \right) \text{ by the inverse triangle inequality}\]

\[ \leq \sum P\left( \|\alpha'_{a,i} - \alpha_{a,i}\| > \frac{a}{2} \right) \text{ since } |\alpha'_{a,i}| > a\]

\[ \leq \sum P\left( \|\alpha'_{a,i} - \alpha_{a,i}\| > \frac{a}{2M} \right) \text{ by the Cauchy-Schwartz inequality}\]

\[ \leq \sum P\left( \|\alpha'_{a,i} - \alpha_{a,i}\| > \frac{a}{2M} \right) \text{ since } |z_{a,i}| \text{ is bounded by } M\]

\[ = nP\left( \frac{1}{\|\hat{\alpha} - \alpha_{a,i}\|} > \frac{a}{2M} \right) \]

\[ = nP\left( \sum_{d=1}^{D} (\hat{\alpha}_{a,i,d} - \alpha_{a,i,d})^2 > \frac{a^2}{4M^2} \right) \text{ by the definition of the } L_2 \text{ norm}\]

\[ \leq nP\left( \bigcup \left\{ (\hat{\alpha}_{a,i,d} - \alpha_{a,i,d})^2 > \frac{a^2}{4M^2D} \right\} \right) \text{ since at least one term must be larger than the average}\]

\[ \leq n \sum_{d=1}^{D} P\left( (\hat{\alpha}_{a,i,d} - \alpha_{a,i,d})^2 > \frac{a^2}{4M^2D} \right) \text{ by sub-additivity}\]

\[ \leq n \sum_{d=1}^{D} P\left( |\hat{\alpha}_{a,i,d} - \alpha_{a,i,d}| > \frac{a}{2M\sqrt{D}} \right) \]

\[ \leq n \sum_{d=1}^{D} 2 \exp \left\{ -\frac{a^2}{4M^2D} / (2Var(\hat{\alpha}_{a,i,d})) \right\} \text{ by the tail bound of a normal distribution}\]

\[ = \sum_{d=1}^{D} 2n \exp \left\{ -\frac{a^2}{4M^2D} O(n)/2 \right\} \text{ since } Var(\hat{\alpha}_{a,i,d}) = O(1/n) \text{ for all } d\]

\[ < \epsilon \text{ for } n \text{ large enough since } 2n \exp \left\{ -\frac{a^2}{4M^2D} O(n)/2 \right\} \to 0 \text{ as } n \to \infty \text{ for all } d\]

So the quantity is bounded in probability.

\[ \square \]

**Lemma 2:** If \( A_n \xrightarrow{p} 0 \) and \( B_n \) is bounded in probability, then \( A_nB_n \xrightarrow{p} 0 \).

**Proof.** Let \( \epsilon, \delta > 0 \). Then \( \exists M \ni P(|B_n| \geq M) \leq \epsilon/2 \) for \( n \geq n_B \). Since \( A_n \xrightarrow{p} 0 \), there
exists $N_A := N_A(M)$ such that $P(|A_n| \geq \delta / M) < \epsilon / 2$.

\[
P(|A_nB_n| \geq \delta) = P(|A_nB_n| \geq \delta, |B_n| \geq M) + P(|A_nB_n| \geq \delta, |B_n| < M)
\]

\[
\leq P(|B_n| \geq M) + P(|A_n| \geq \delta / M)
\]

So for $n \geq \max\{N_A, N_B\}$,

\[
P(|A_nB_n| \geq \delta) < \epsilon / 2 + \epsilon / 2 = \epsilon
\]

So $|A_nB_n| \xrightarrow{p} 0$.

Next, we see that

\[
\widehat{\text{Cov}}(\mathbf{x} - \mathbf{x}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - x_i) x_i - \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - x_i) \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\left(\alpha'_0 z_{a,i} + \alpha'_1 z_{a,i} x_i + \epsilon_i\right) - \hat{\alpha}'_0 z_{a,i} - x_i}{\alpha'_1 z_{a,i}} \right) x_i
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\left(\alpha'_0 z_{a,i} + \alpha'_1 z_{a,i} x_i + \epsilon_i\right) - \hat{\alpha}'_0 z_{a,i} - x_i}{\alpha'_1 z_{a,i}} \right) \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha'_1 z_{a,i}} \left[ (\alpha'_0 - \hat{\alpha}'_0) z_{a,i} (x_i - \bar{x}) + (\alpha'_1 - \hat{\alpha}'_1) z_{a,i} (x_i^2 - x_i \bar{x}) + \epsilon_i (x_i - \bar{x}) \right]
\]

Terms 1-2: we rewrite the first two terms as follows:

\[
\frac{1}{\alpha'_1 z_{a,i}} (\alpha'_0 - \hat{\alpha}'_0) z_{a,i} (x_i - \bar{x}) = \frac{1}{\alpha'_1 z_{a,i}} (\alpha'_0 - \hat{\alpha}'_0) z_{a,i} m_{i,0} \quad \text{where } m_{i,0} = x_i - \bar{x}
\]

\[
\frac{1}{\alpha'_1 z_{a,i}} (\alpha'_1 - \hat{\alpha}'_1) z_{a,i} (x_i^2 - x_i \bar{x}) = \frac{1}{\alpha'_1 z_{a,i}} (\alpha'_1 - \hat{\alpha}'_1) z_{a,i} m_{i,1} \quad \text{where } m_{i,1} = x_i^2 - x_i \bar{x}
\]

Then for $j \in \{0, 1\}, k \in \{0, 1\}$:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha'_1 z_{a,i}} (\alpha'_k - \hat{\alpha}'_k) z_{a,i} m_{i,j} \right|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha'_1 z_{a,i}} \left| (\alpha'_k - \hat{\alpha}'_k) z_{a,i} \right| \cdot |m_{i,j}| \quad \text{by the triangle inequality}
\]

\[
\leq \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha'_1 z_{a,i}} \cdot |z_{a,i}| \cdot |m_{i,j}| \right) \cdot \|\alpha_k - \hat{\alpha}_k\| \quad \text{by the Cauchy-Schwartz inequality}
\]

\[
\leq \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha'_1 z_{a,i}} \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} |z_{a,i}|^2 \cdot |m_{i,j}|^2 \right)^{1/2} \cdot \|\alpha_k - \hat{\alpha}_k\| \quad \text{by Cauchy-Schwartz}
\]

35
By Lemma 1, the first term is bounded in probability. Also, since the covariates are bounded, there exists a $B > 0$ such that $||z_{a,i}|| \leq B$, and $X$ being bounded means there exists $C > 0$ such that $|m_{i,j}| \leq C$. Then, $\frac{1}{n} \sum_{i=1}^{n} ||z_{a,i}||^2 \leq \frac{1}{n} \sum_{i=1}^{n} B^2 C^2$. Lastly, $||\alpha_k - \hat{\alpha}_k|| \xrightarrow{P} 0$ by the consistency of $\hat{\alpha}_k$.

By Lemma 2, the entire quantity converges to 0 in probability. So the first two terms converge to 0 in probability.

For the third term, 

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i (x_i - \bar{x}) \leq \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\alpha_1' z_{a,i}} \epsilon_i (x_i - \bar{x}) - \frac{1}{\alpha_1' z_{a,i}} \epsilon_i (x_i - \bar{x}) \right) + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i (x_i - \bar{x})$$

by the triangle inequality

Note that the second part of the expression tends to 0 asymptotically since

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i x_i - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i \bar{x}$$

We apply the weak law of large numbers directly to the first term to see $\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i x_i \xrightarrow{P} E \left[ \frac{1}{\alpha_1' z_{a}} \epsilon X \right] = E[\epsilon] E \left[ \frac{1}{\alpha_1' z_{a}} X \right] = 0$ since $\epsilon$ is independent of $X, Z_a$ and $E[\epsilon] = 0$. For the second term we write

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i \bar{x} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha_1' z_{a,i}} \epsilon_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

$\xrightarrow{P} E \left[ \frac{1}{\alpha_1' Z_a} \epsilon \right] E[X]$ by the weak law of large numbers and Slutsky’s theorem

$= E[\epsilon] E \left[ \frac{1}{\alpha_1' Z_a} \right] E[X]$ since $\epsilon$ and $Z_a$ are independent

$= 0$ since $E[\epsilon] = 0$
For the first part of the expression,
\[
\left| \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\alpha_1 Z_{a,i}} \epsilon_i (x_i - \bar{x}) - \frac{1}{\alpha_1' Z_{a,i}} \epsilon_i (x_i - \bar{x}) \right) \right|
\]
\[
= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(\alpha_1 - \alpha_1')Z_{a,i}}{(\alpha_1' Z_{a,i})(\alpha_1' Z_{a,i})} \epsilon_i (x_i - \bar{x}) \right|
\]
\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left| \frac{(\alpha_1 - \alpha_1')Z_{a,i}}{\alpha_1' Z_{a,i}} \right| |\epsilon_i| |x_i - \bar{x}| \quad \text{by the triangle inequality}
\]
\[
\leq ||\alpha_1 - \alpha_1'|| \frac{1}{n} \sum_{i=1}^{n} \frac{||Z_{a,i}||}{\alpha_1' Z_{a,i}} |\epsilon_i| |x_i - \bar{x}| \quad \text{by the Cauchy-Schwartz inequality}
\]
\[
\leq ||\alpha_1 - \alpha_1'|| \left( \frac{1}{n} \sum_{i=1}^{n} \frac{||Z_{a,i}||^2 \epsilon_i^2 (x_i - \bar{x})^2}{(\alpha_1' Z_{a,i})^2} \right)^{\frac{1}{2}} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(\alpha_1' Z_{a,i})^2} \right)^{\frac{1}{2}} \quad \text{by Cauchy-Schwartz}
\]

In this expression, \( ||\alpha_1 - \alpha_1'|| \xrightarrow{P} 0 \) by the consistency of the least squares estimator. 
\[
\left( \frac{1}{n} \sum_{i=1}^{n} ||Z_{a,i}||^2 \epsilon_i^2 (x_i - \bar{x})^2 \right) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{d_2^2 p^2 \epsilon_i^2}{a^2} \xrightarrow{P} \frac{d_2^2 p^2}{a^2} \text{Var}(\epsilon) < \infty \quad \text{since } \bar{z}_{a,i}, x_i - \bar{x} \text{ are bounded (with the bounds denoted as } D, F \text{ respectively), and } \text{Var}(\epsilon) < \infty. \quad \text{So this factor is bounded in probability. Finally, by Lemma 1, the last factor is bounded in probability. By Lemma 2, the entire third term converges to 0 in probability.}
\]

Since all three terms tend to 0 in probability, we can use Slutsky’s theorem to conclude that \( \text{Cov}(\hat{x} - x, x) \xrightarrow{P} 0. \)

\[\square\]

**S2 Predicting pollutant concentrations on a grid**

We provide the details of predicting pollutant concentrations on a grid of locations (Set D) with neither low-cost or reference data (Section 3.4.1). To write the conditional distributions of these random variables, we will drop the \( t \) and use a subscript to notate the set (for example \( x_D = x(S_D, t) \)). Also, let \( D = A \cup B \cup C \). Thus, the conditional distributions are:

\[
x_D | x_D \sim N(\mu_D, \Sigma_D)
\]

where \( \mu_D = \mu_1 + C_{DD}^{-1}C_{DB}(x_D - \mu_1) \)
\( \Sigma_D = C_{DD} - C_{DD}^{-1}C_{DB}C_{BB}C_{DB}^{-1} \)

\[
x_B | \begin{pmatrix} x_A \\ x_C \end{pmatrix}, y_B \sim N(\mathbf{m}_B, \Sigma_B)
\]

where \( \mathbf{m}_B = (\Sigma_B^{-1} \mu_B + \frac{1}{\tau_2} \mathbf{H}_B u_B) \)
\( \Sigma_B = (\Sigma_B^{-1} + \frac{1}{\tau_2} \mathbf{H}_B)^{-1} \)
In these equations, \( \tilde{\mu}_B = \tilde{\mu}_t \), \( \Sigma_B = \Sigma_t \), \( H_B = H(S_B, t) \) and \( u_B = u(S_B, t) \) from Section 3.4. Using these two distributions, we get that

\[
\begin{pmatrix}
  x_D - C_{DD}C_{DD}^{-1}(x_D - \mu 1) \\
x_B
\end{pmatrix} | \begin{pmatrix} x_A \\ x_C \end{pmatrix}, y_B \sim N \left( \begin{pmatrix} \mu 1 \\ S_B m_B \end{pmatrix}, \begin{pmatrix} \Sigma_D & 0 \\ 0 & S_B \end{pmatrix} \right)
\]

where the first element of the left hand side is written in this form so that that its distribution is conditionally independent of the distribution of \( x_B \). This results in a block diagonal covariance matrix. We rewrite as the posterior so that the first element of the left hand side is a function of \( x_D \) and \( x_B \) only:

\[
\begin{pmatrix}
  x_D - A_2 x_B \\
x_B
\end{pmatrix} | \begin{pmatrix} x_A \\ x_C \end{pmatrix}, y_B \\
\sim N \left( \begin{pmatrix} \mu 1 + A_1(x_A - \mu 1) + A_3(x_C - \mu 1) - A_2 \mu 1 \\ S_B m_B \end{pmatrix}, \begin{pmatrix} \Sigma_D & 0 \\ 0 & S_B \end{pmatrix} \right)
\]

(S1)

where \( A_1, A_2, A_3 \) are blocks of columns of \( C_{DD}C_{DD}^{-1} \) corresponding to \( x_A, x_B, x_C \) respectively.

Equation (S1) can be used to predict \( \begin{pmatrix} x_D - A_2 x_B \\ x_B \end{pmatrix} | \begin{pmatrix} x_A \\ x_C \end{pmatrix}, y_B \) by using the posterior mean as the predicted PM\(_{2.5}\) concentration. Since \( \begin{pmatrix} x_D \\ x_B \end{pmatrix} \) is a linear function of \( \begin{pmatrix} x_D - A_2 x_B \\ x_B \end{pmatrix} \), we can take a linear function of the predictions to predict \( \begin{pmatrix} x_D \\ x_B \end{pmatrix} \). This can be written out as

\[
\begin{pmatrix}
  \hat{x}_D \\
  \hat{x}_B
\end{pmatrix} = \begin{pmatrix} I & A_2 \\ 0 & I \end{pmatrix} \begin{pmatrix} x_D - A_2 x_B \\ x_B \end{pmatrix}
\]

\[
= \begin{pmatrix} I & A_2 \\ 0 & I \end{pmatrix} \begin{pmatrix} \mu 1 + A_1(x_A - \mu 1) + A_3(x_C - \mu 1) - A_2 \mu 1 \\ S_B m_B \end{pmatrix}
\]

(S2)
Table 2: Simulation 1a: Runtimes (seconds) for 100 timepoints and 50 sensors. Frequentist implementation and fully parallelized Bayesian implementation are compared.

|                      | Frequentist | Bayesian |
|----------------------|-------------|----------|
| Covariate model      | 3.56        | 322.86   |
| No covariate model   | 2.37        | 296.92   |

Figure S1: Simulation 1a: RMSE by distance of sensor from the co-located reference device, using correctly specified Gaussian process spatial model, across 50 datasets.
Figure S2: Simulation 1b: Average parameter estimates with 95% interval from GP Filter, comparing 50 and 200 non-co-located sites, with 1 co-located site. True parameter values are shown in black. Averaged over 20 datasets.

Figure S3: Simulation 2: RMSEs for correctly specified and misspecified covariate set in the observation model.
Figure S4: Simulation 2: False negative rates for correctly specified and misspecified covariate set in the observation model.

Figure S5: Simulation 3: RMSE when the underlying pollution surface is misspecified.
Figure S6: Simulation 3: FNR when the underlying pollution surface is misspecified.