Parameter recovery using remotely sensed variables

Jonathan Proctor[1] ⓡ, Tamma Carleton[2,3] ⓡSandy Sum[2]

[1]Center for the Environment and Data Science Initiative, Harvard University
[2]Bren School of Environmental Science and Management, University of California, Santa Barbara
[3]National Bureau of Economic Research

March 23, 2023

Abstract

Remotely sensed measurements and other machine learning predictions are increasingly used in place of direct observations in empirical analyses. Errors in such measures may bias parameter estimation, but it remains unclear how large such biases are or how to correct for them. We show empirically that using remotely sensed variables without correction leads to substantial bias in point estimates and standard errors across a diversity of models. We demonstrate that multiple imputation, a standard and easily implementable statistical imputation technique that has yet to be tested in this setting, effectively reduces bias and improves statistical coverage in both cross-sectional and panel data designs.

Author order randomized and archived in the American Economic Association Random Author Order Archive. Proctor: jproctor1@fas.harvard.edu. Carleton: tcarleton@ucsb.edu. Sum: sandysum@ucsb.edu. We thank Patrick Baylis, Avi Feller, Robert Heilmayr, Solomon Hsiang, Peter Huybers, Kelsey Jack, Esther Rolf, and seminar participants at The Workshop for Environmental Economics and Data Science (TWEEDS) and at the University of California Environmental Economics (UC-EE) seminar for helpful comments and suggestions. We thank Eliana Stone for research assistance.

1 Introduction

Since the first images of Earth were captured by satellite in 1960, the availability of satellite imagery has grown substantially, with now over one hundred terabytes of imagery data collected daily. Alongside increasing imagery, a cross-disciplinary research community spurred by advances in computer vision has developed a range of algorithms to transform raw images into predictions of social, economic, and environmental variables. For example, remote sensing algorithms have enabled researchers to track deforestation (Hansen et al., 2013), identify illegal mining activities (Swenson et al., 2011), and monitor economic activity like agricultural land use (Potapov et al., 2022) and wealth (Jean et al., 2016; Chi et al., 2022) at fine resolution and national or even global scales.

These predictions provide a treasure trove of new data to better understand key drivers of human and environmental well-being and their use in empirical research is growing rapidly. For example, measurements of global forest cover and deforestation from Hansen et al. (2013) have been cited over 8,000 times since their release. However, these remotely sensed predictions[1] are indirect measures of the true variables of interest and often exhibit substantial measurement error, which may be correlated with the variable itself or with other variables in the model. As a result, remotely sensed variables may introduce bias into both parameter estimates and associated measures of uncertainty when used in downstream regression analyses. For example, errors in remotely sensed air pollution have been shown to confound estimates of the relationship between air pollution and mortality (Josey et al., 2022). These biases can arise whether remotely sensed variables are used in causal inference settings with clear experimental designs, or in descriptive analyses that are correlational.

While the biases in parameter estimation introduced by measurement error and methods for their correction have been long documented in the statistical literature (e.g., Little

[1] Throughout this paper, we use the phrases "remotely sensed" and "satellite-based" interchangeably. Although our analysis relies exclusively on satellite imagery based predictions, other forms of remote sensing, such as Light Detection and Ranging (LIDAR), similarly exhibit nonrandom errors that can influence downstream regression analyses (Triglav-Čekada, Crosilla and Kosmatin-Fras, 2009).

and Rubin, 2019; Wooldridge, 2010; Wang, McCormick and Leek, 2020), directly transferring error correction methods from other disciplines is complicated by the complex nature of errors in remotely sensed measurements, which can arise from flaws in the imagery, from key features not being visible, or from errors in the translation of the information within the image (e.g., color and texture) into the outcome of interest (e.g., forest cover). Thus, it is still common practice to use satellite-based measures without correction as either the dependent (e.g., BenYishay et al., 2017; Sims and Alix-Garcia, 2017; Marx, Stoker and Suri, 2019; Balboni et al., 2021) or independent (e.g., Chen et al., 2017; Crost et al., 2018; Freeman et al., 2019; Harari, 2020; Kocornik-Mina et al., 2020; Proctor, 2021; Chen, Oliva and Zhang, 2022) variable in regression analysis. The degree of bias introduced by measurement error in remotely sensed variables has yet to be systematically quantified, and a generalizable and easily implementable solution to account for such errors has yet to be proposed.

In this paper, we quantify the extent to which continuous remotely sensed variables introduce parameter bias and lead to incorrect estimates of parameter uncertainty when used in regression analysis as either an independent or dependent variable. We do so both using reanalysis of published research and using a set of real-data simulation experiments that leverage a newly available benchmark dataset providing co-located ground truth ("labels") and remotely sensed predictions for multiple variables across the contiguous United States. While simulations have been extensively used to demonstrate the efficacy of all error correction methods we evaluate here (e.g., De Silva et al., 2017; Freedman et al., 2008; Cole, Chu and Greenland, 2006), such results depend critically on assumptions about the structure of measurement error in the experimental design. Because these assumptions are largely untestable in applied settings, we rely on actual remotely-sensed and ground truth measurements, as opposed to simulated data, to evaluate what types of measurement error are typically present, what types of error lead to bias, and to what degree these biases are amenable to correction.

We show that not accounting for measurement error, as is standard in most applied research, tends to substantially bias parameter point estimates and dramatically decrease coverage[2] across a diversity of empirical settings. For example, we find that 95% confidence intervals estimated using remotely sensed data rarely contain the true parameter of interest, raising concerns about the growing use of remotely sensed data in empirical research. We demonstrate that while mean-reverting measurement error (negative correlations between errors in one variable and itself) is common in remotely sensed variables, differential measurement error (correlations between errors in one variable and levels of another variable) is responsible for the majority of the induced bias.

We then present a method to account for measurement error that is feasible in cases where researchers have even a small quantity of labeled data for calibration. We show that multiple imputation, an "off-the-shelf" data imputation technique used widely in statistics to solve missing data challenges, but so far untested in this setting, improves the accuracy of recovered parameters and prevents exaggerated statistical precision across a broad set of empirical models. In turn, corrected 95% confidence intervals contain the true parameter of interest over 90% of the time. We demonstrate that multiple imputation performs well under common limitations that applied researchers face, such as small samples of ground truth data located far from target areas of interest, and when applied in panel fixed effects settings commonly used for program evaluation. Such settings receive little study in the error-correction literature, but are critical for informing the applicability of error-correction methods in empirical research. Throughout, we compare the performance of multiple imputation to other common error correction methods, showing that it systematically out-performs alternative approaches. Collectively, our findings indicate that multiple imputation is a generalizable and easily implementable method for correcting parameter estimates that rely on remotely sensed variables.[3]

Our findings contribute to a nascent literature that has begun to document measurement error in satellite-based measurements and explore its implications for parameter

[2] Coverage is defined as the likelihood that an estimated confidence interval contains the true parameter of interest.

[3] Multiple imputation can be implemented off-the-shelf in `R` using the `mice` package, in Python using the `scikit-learn` library, and in Stata using the `mi` command.

estimates and uncertainty (Fowlie, Rubin and Walker, 2019; Ratledge et al., 2022). Review articles raise the general issue (Jain, 2020; Gibson et al., 2021), and solutions to the problem have been proposed in the case of binary landcover data (Alix-Garcia and Millimet, 2020; Garcia and Heilmayr, 2022). One paper develops a method to correct for mean-reverting measurement error when making remote sensing predictions of continuous variables by tailoring the loss function during algorithm development (Ratledge et al., 2022). While this method appears to effectively reduce bias from mean-reverting measurement error, it is not designed to address other types of measurement error such as differential measurement error, which we find drives the majority of the bias induced by remotely sensed variables. Further, the method is infeasible to implement for the vast majority of researchers, who use, but do not themselves produce, remotely sensed predictions. Though each unique analysis using remotely sensed data requires individual consideration, this rapidly growing field lacks a comprehensive assessment of the magnitude of bias introduced by errors in remotely sensed measurements. This analysis represents, to our knowledge, the most extensive set of experiments to date aimed at informing the use and correction of remotely sensed measurements in regression analysis. Moreover, while our quantitative insights are most relevant to analyses using remotely sensed variables, the threats to parameter recovery that we identify, as well as the solution that we propose, apply more generally to the use of machine learning (ML) predictions in downstream regression analysis.

The paper proceeds as follows. Section 2 describes the data used throughout the paper. Section 3 details a set of experiments designed to both quantify the biases introduced by remotely sensed variables in regression analysis and to evaluate the performance of multiple imputation and other error correction methods at mitigating these biases. Section 4 presents results from these experiments and Section 5 concludes by highlighting key considerations for researchers conducting regression analysis with remotely sensed variables.

2 Data

Our real-data simulation analysis relies primarily on a multi-label benchmark dataset from Rolf et al. (2021) that includes remotely sensed predictions and corresponding ground truth labels for six variables: forest cover, population density, nighttime luminosity, average household income, elevation, and road length. These predictions were constructed from high-resolution visual imagery using the Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS) framework, a machine learning approach that relies on an unsupervised featurization of imagery called random convolutional features in combination with a ridge regression to train a model to predict an outcome of interest (see Rolf et al. (2021) and Rahimi and Recht (2007) for details). Importantly, MOSAIKS generates predictions with similar error magnitude and structure to other commonly used methods, such as a convolutional neural network (see Supplementary Materials Section C.2, Figure B.2 and Rolf et al. (2021)'s Supplementary Figure 17), making these data generally representative of many modern remotely sensed predictions.

Figure 1 shows these labels and remotely sensed predictions for all six variables at 1km^2 resolution across 100,000 sampled locations in the contiguous United States.[4] Ground truth observations are collected from ∼2016 (see Table A.1 for details). Of these 100,000 grid cells, we randomly sample 40,000 to facilitate computation throughout the analysis. To increase the number and variety of variables considered, we augment this dataset with observations of average temperature and precipitation from PRISM.[5] Throughout our analysis, we focus on continuous remotely sensed variables as these are most commonly used in applied economics studies (e.g., Heft-Neal et al., 2020; Freeman et al., 2019; Ratledge et al., 2022). We standardize all our ground truth and remotely sensed variables by the mean and standard deviation of the ground truth variable to facilitate comparisons of coefficients across variable pairs.

[4] Note that the sampling was population weighted, which is why the missing locations in Figure 1 are not uniformly distributed.

[5] Data available at `https://prism.oregonstate.edu/`. We assign each 1km x 1km grid cell the value of the 0.8km x 0.8km PRISM grid cell that contains its center. Temperature and precipitation are 30 year averages from 1991-2020.

To assess the implications of errors in remotely sensed variables in prior research, we use data from Deschenes, Greenstone and Shapiro (2017)'s study of the effects of the U.S. NO_x budget program on ambient air pollution. These data include county-by-season estimates of $\text{PM}_{2.5}$ across the U.S. over the years 2001-2007, constructed by interpolating ground monitor data across each county's area. We replicate the authors' original analysis and then re-estimate their main specification using a satellite-derived $\text{PM}_{2.5}$ dataset from Van Donkelaar et al. (2021), a widely-used remotely sensed measure of air pollution, in place of the original monitor data. To do so, we average the gridded monthly $\text{PM}_{2.5}$ data from Van Donkelaar et al. (2021) to the county-by-season level.

3 Methods

3.1 Empirical setting

To begin, we consider a straightforward setting in which a researcher is interested in estimating the following simple linear regression:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

We assume that Equation 1 is correctly specified, such that recovered coefficients $\hat{\alpha}$ and $\hat{\beta}$ represent unbiased estimates of the true parameters of interest when Equation 1 is estimated using ground truth data.[6] Similarly, we assume that standard errors on the coefficients recovered from estimating Equation 1 with ground truth data are unbiased estimates of parameter uncertainty. In initial experiments, we consider a cross-sectional research design, such that i indexes location (one of 40,000 1km x 1km grid cells across the U.S.), but we later generalize our analysis to the panel data setting.

We assume the researcher, however, cannot directly estimate Equation 1 because she cannot observe either y or x using traditionally collected ground truth data. Instead, she

[6] We address the implications of measurement error in the ground "truth" data themselves in the discussion.

must rely on a remotely sensed proxy, indicated by $\tilde{y}$ or $\tilde{x}$. For example, if ground truth data for the independent variable x are not available, as illustrated in the left panel of Figure 2, she will estimate what we call the "error-in-X" regression:

$$y_i = \alpha_{\tilde{x}} + \beta_{\tilde{x}}\tilde{x}_i + \epsilon_i, \tag{2}$$

where remotely sensed measures $\tilde{x}$ are used directly as proxies for x. Due to measurement error in $\tilde{x}$, the researcher who estimates Equation 2 will recover different coefficients $\hat{\alpha}_{\tilde{x}}$ and $\hat{\beta}_{\tilde{x}}$ and different estimates of their standard errors than those from Equation 1. Similarly, if ground truth data for y are not available, as illustrated in the middle panel in Figure 2, the researcher will estimate the analogous "error-in-Y" regression, where remotely sensed predictions $\tilde{y}$ are substituted for y in Equation 1. We design a set of experiments to quantify how parameters differ when regressions are estimated using ground truth versus remotely sensed data and to test the effectiveness of various error correction methods at reducing these differences across a diversity of applied settings. We detail these experiments below, and visually illustrate the overall experimental design in Figure B.1.

3.2 Performance metrics

Our analysis quantifies bias introduced by remotely sensed measurements by comparing recovered parameters in Equation 2 to those of Equation 1, in both the error-in-X and error-in-Y cases. We similarly evaluate the performance of corrected models, which leverage multiple imputation or alternative error correction approaches to adjust $\hat{\beta}_{\tilde{x}}$ (or, analogously, $\hat{\beta}_{\tilde{y}}$) and its estimated standard error. The four performance metrics we use to compare uncorrected and corrected regression models to those using ground truth data are: absolute proportional coefficient bias, proportional standard error bias, coverage, and power (see Supplementary Materials Section C.1 for details).

3.3 Experiment one: quantify parameter biases introduced by remotely sensed measurements

In our first experiment, we systematically evaluate the extent to which error in remotely sensed measurements induces bias in parameter estimates across diverse empirical settings. The co-location of the labeled data and predictions from Rolf et al. (2021) enables us to create a set of 42 different ordered pairs of variables[7] (e.g., population density and forest cover; nighttime luminosity and income, etc.). While it is impossible to create a fully representative sample of the many possible empirical settings where remotely sensed data could be used in regression analysis, these pairs provide a large and heterogeneous set of regression models to evaluate the degree of bias introduced by measurement error and the ability of multiple imputation to correct for it.

For each pair of variables, we estimate three regressions, reflecting the three data availability regimes outlined in Figure 2: one where the ground truth labels are used for both variables (as in Equation 1), one where the remotely sensed variable is used for the dependent variable (i.e., error-in-Y), and one where the remotely sensed variable is used for the independent variable (i.e., error-in-X).

We calculate the distribution of uncorrected and ground truth estimates for each pair of variables using a bootstrap procedure illustrated in Figure B.1. Specifically, we create 100 datasets of size 40,000 by randomly sampling with replacement from the original dataset. Coverage is calculated as the fraction of the 100 bootstrap runs in which the estimated confidence interval contains the ground truth point estimate. Power is calculated as the fraction of bootstrap runs where the null of $\beta = 0$ is rejected when this null is also rejected in the ground truth data. Figures show the distribution of bias in the regression coefficient (i.e., Equation S1) and bias in standard errors (i.e., Equation S2)

[7] We have 6 variables from Rolf et al. (2021) and add 2 climate variables from PRISM, leading to 8 total. Each variable can be paired with every other variable twice, once where it is the outcome variable, and once where it is the independent variable. This gives $8 \times 7 = 56$ ordered pair combinations for analysis. We do not explore remotely sensed predictions of climate data, so pairs including temperature or precipitation variables are only estimated with two regressions. This leaves us 42 error-in-X and error-in-Y models with associated ground truth models. All variables are normalized by their standard deviation prior to regression analysis to facilitate comparison across pairs.

over all bootstrap × variable pair combinations, and the distribution of coverage and power over all variable pair combinations.[8] All performance metrics are calculated for only the 40 of the 42 variable pairs that have a significant ($p < 0.05$) relationship in the true data.

3.4 Experiment two: evaluate the efficacy of multiple imputation for correcting parameter biases and inference

In our second experiment, we use the empirical setup described above to evaluate the efficacy of various error correction methods drawn from the statistics literature, including multiple imputation. Similar to the uncorrected model, we compute bias in the regression coefficient, bias in the standard errors, coverage, and power for each pair of variables for a variety of error correction methods over the 100 bootstrapped samples. We focus our main analysis and results on multiple imputation (Rubin and Schenker, 1991) due to its prior performance in other fields (Rubin, 1987; Cole, Chu and Greenland, 2006; Keogh and White, 2014), its ability to account for uncertainty, its more flexible assumptions about the distribution of errors relative to the other calibration methods (Keogh and White, 2014), and its performance in our analysis. All other error correction methods we evaluate are described in Supplementary Materials Section C.4.

Multiple imputation, as well as all other error correction methods we consider, relies on a user-provided calibration dataset where researchers obtain some quantity of ground truth data. For example, the researcher may be estimating the error-in-X regression Equation 2, and have access to a smaller calibration dataset where y, x, and $\tilde{x}$ are all available, as illustrated in Figure 2. Intuitively, the calibration dataset allows the researcher to estimate the structure of the measurement error present in the remotely sensed variables, and then to use that estimated structure to correct parameter estimates and measures of uncertainty when estimating the regression of interest in the main sam-

[8] Averaging coefficients or standard errors over bootstrap runs before calculating bias gives consistent results.

ple. To implement this in our analysis, we partition each bootstrap sample of 40,000 observations into a "main" sample of size 28,000 (70%), where we assume the researcher does not have access to ground truth data for all variables, and a "calibration" sample of size 12,000 (30%), in which additional ground truth data is used for error correction, as described below.

To illustrate how multiple imputation works, consider the error-in-X case. Multiple imputation begins with a first-stage model, such as (but not limited to) linear regression, where the relationship between the ground truth observation x and both the remotely sensed observation $\tilde{x}$, and the ground truth observations of the outcome variable y is estimated in the calibration sample. For example, the researcher might estimate a linear first stage model using data from the calibration sample, as follows:

$$x_i = \delta + \gamma \tilde{x}_i + \psi y_i + e_i \tag{3}$$

Estimated coefficients from Equation 3, which is called the imputation step, describe how the remotely sensed proxy relates to the ground truth measurement, conditional on the outcome variable.[9] For example, when extreme values of x are routinely underestimated, as is common in remotely sensed estimates (e.g., see Figure B.2), $\hat{\gamma}$ is greater than one. Next, the estimated relationship from Equation 3 is used to impute a "corrected" value of $\tilde{x}$ in the main sample, where ground truth measurements x do not exist. Call this prediction $\hat{x}$. In multiple imputation, this imputation step is repeated K times, such that K versions of $\hat{x}$ are generated using predictions from Equation 3. This duplication of the error-correction imputation step can be done in a variety of ways, including through bootstrapping or through estimating Equation 3 with Bayesian linear regression, as we

[9] Multiple imputation can easily be extended to more complex regression settings with additional covariates or controls like fixed effects. This results in additional controls being included in Equations 3 and 4. We detail and implement multiple imputation in a setting with additional controls and fixed effects in Section 4.6. Further, Equation 3 can also be nonlinear; for example, a random forest can be used to flexibly estimate the relationship between x, $\tilde{x}$, and y, as well as any relevant covariates (Van Buuren and Groothuis-Oudshoorn, 2011). Here, we use linear regression, but we show in Figure B.7 that our results are very similar using other imputation models.

do here, where K coefficient draws are taken from a posterior predictive distribution assumed to be Gaussian. Regardless of its specific implementation, multiple imputation uses a form of Equation 3 to generate a set of K corrected values of the remotely sensed variable $\tilde{x}$ in the main sample. Denoting each of these $\hat{x}^k$, these predictions are then used to run the following second stage regression model in the main sample K times:

$$y_i = \alpha^k + \beta^k \hat{x}_i^k + u_i^k, \tag{4}$$

leading to K point estimates, $\hat{\alpha}^k$ and $\hat{\beta}^k$, as well as their standard errors. This set of parameters is then used to derive final multiple imputation coefficients and measures of uncertainty using Rubin's Rule, a group of formulas based on asymptotic theory for combining coefficient estimates and standard errors into final parameter estimates (Rubin and Schenker, 1991).[10] A key benefit of multiple imputation is that final coefficients and standard errors are constructed accounting for uncertainty in the imputation step (Equation 3). This method was originally designed to address systematic non-response in surveys (Rubin and Schenker, 1991), but has been applied widely in biostatistics and other applied statistics fields (Sterne et al., 2009; Liu and De, 2015).[11]

Figure 3 uses an example of one simple linear regression to illustrate how remotely

[10] The Rubin's Rule combined multiple imputation coefficient estimate is the average across the K point estimates from Equation 4: $\hat{\beta}^{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}^k$. Rubin's Rule then combines the within-imputation and the between-imputation variances to generate a single estimate of coefficient variance. The within-imputation variance is calculated as the average of the conventional sampling variance across imputations: $V_w = \frac{1}{K}\sum_{k=1}^{K} V(\hat{\beta}^k)$, where $V(.)$ indicates variance. The between-imputation variance is calculated as the variance of the estimated coefficient of interest across imputations: $V_b = \frac{\sum_{k=1}^{K}(\hat{\beta}^k - \hat{\beta})}{K-1}$. The total variance is the sum of the two: $V(\hat{\beta}) = V_w + V_b + \frac{V_b}{K}$.

[11] Readers may note that multiple imputation somewhat resembles the standard instrumental variables approach to correct for measurement error, in which one error-prone variable is used as an instrument for another, under the assumption that the errors are independent across the two variables (Wooldridge, 2010). A primary difference is that multiple imputation requires a small calibration dataset where ground truth measurements are available, whereas two-stage least squares (2SLS) requires a second, independent, measurement across the entire main sample. Obtaining such a measurement for the entire main sample seems unlikely in the context of remote sensing, which is often leveraged in data-limited settings. Additionally, previous work has shown that two remotely sensed predictions of the same outcome tend to have correlated errors Rolf et al. (2021). Another important difference between multiple imputation and 2SLS is the types of measurement error they can correct – multiple imputation can address non-differential measurement error, as discussed in the results, whereas standard IV approaches do not. Moreover, 2SLS cannot be used to solve biases arising from an error-prone dependent variable (i.e., error-in-Y).

sensed measures can bias parameter recovery and how multiple imputation can correct for such biases. Panel A shows that there is substantial measurement error in remotely sensed estimates of road length; in particular, extremely high levels of road length are systematically underestimated by the machine learning model. Panel B demonstrates that this error introduces substantial bias in the estimation of the relationship between population density and road length; in this "error-in-X" model, the slope coefficient turns out to be biased upward. Multiple imputation uses a calibration sample to learn the relationship between true and predicted road length, controlling for population density, and then adjusts all data points in the main sample to account for this estimated relationship. A final regression with these adjusted data points – repeated multiple times with estimates combined as described above – recovers an unbiased estimate of the true parameter of interest, as shown in Panel C. In panels D and E, this same exercise is shown, but with remotely sensed road length as the outcome variable, making it an "error-in-Y" model. There is substantially less bias for multiple imputation to correct in this example for the error-in-Y than error-in-X case. In our real-data simulation experiment, we repeat the estimation and correction shown in Figure 3 for all 40 pairs of variables in our benchmark dataset.

3.5 Experiment three: examine the performance of multiple imputation in data-limited and panel data research settings

Our final set of experiments evaluates the robustness and generalizability of multiple imputation in more data-limited and complex settings that applied researchers commonly face. We investigate the implications of having a limited number of calibration data observations and of having a calibration set distant from the main sample. We also evaluate the challenges and efficacy of using multiple imputation with control variables and in a triple-difference panel data research design used in a prominent research paper. The methods for each experiment are described accompanying the findings, in the results section.

4 Results

4.1 Errors in remotely sensed measurements bias the distribution of parameter estimates.

The bias introduced by errors in remotely sensed measurements across all regressions between all pairs of variables is shown in Figure 4. Each panel shows the distribution of each performance metric over all bootstrap samples for all 40 of the 42 pairs of variables for which the ground truth data reject the null hypothesis of no empirical relationship at the 0.05 significance level. Performance metrics for regression models using uncorrected remotely sensed predictions are shown in purple. Median estimates of these distributions are indicated with a black dot, while means are shown with a red dot. Error-in-X models are shown in column A and error-in-Y models are shown in column B.

These results reveal that remotely sensed variables tend to introduce substantial bias into linear regression coefficients. The median coefficient bias of the uncorrected estimates is 23% across all regression models and bootstrap samples in the error-in-X case, and 10% in the error-in-Y case. Note the long tail in these distributions: for some pairs of variables, substituting ground truth for remotely sensed observations leads to biases of over 100% (e.g., night lights regressed on income with a mean bias of 700%, and elevation regressed on income with a mean bias of 110%). These long tails lead the mean bias to exceed the median bias, with a mean bias of 69% in the error-in-X case and 37% in the error-in-Y case. Figure B.3 shows that coefficients tend to be *exaggerated* in the error-in-X model but *attenuated* in error-in-Y models, although biases in both directions are common in both cases.

Errors in remotely sensed measurements also lead to biased estimates of parameter uncertainty (Figure 4, second row). Standard errors are biased large in the error-in-X

case, with a median bias of 12%, but a long right tail. Somewhat more concerning is that in the error-in-Y case, the uncorrected standard errors are biased downward relative to true values, with a median bias of -12%, indicating that replacing dependent variable ground truth observations with remotely sensed values tends to overstate the precision of point estimates.

Importantly, large coefficient bias in the uncorrected approach leads to exceptionally poor coverage in both the error-in-X and error-in-Y cases (Figure 4, third row). Whether a remotely sensed variable is used as the independent variable or dependent variable, mean coverage is below 25%, indicating that recovered 95% confidence intervals rarely contain the true parameter of interest. This lack of coverage is concerning given the increasing use of remotely sensed data for parameter recovery and causal inference. Power is not a cause for concern in the uncorrected models, as mean power is higher than 95% in both the error-in-X and error-in-Y cases (Figure 4, bottom row).

4.2 Diagnosing the origins of bias caused by remotely sensed measurements

Many different forms of measurement error could be responsible for the biases observed in Figure 4. The *linear measurement error model* (Keogh et al., 2020) is a fairly general model of measurement error structure that helps elucidate the patterns we recover. In this model, the remotely sensed variable is expressed as an affine function of the true variable:

$$\begin{aligned} \tilde{y} &= \theta + \lambda y + u \qquad \text{(error-in-}Y\text{)} \\ \tilde{x} &= \theta + \lambda x + u \qquad \text{(error-in-}X\text{)}, \end{aligned} \tag{5}$$

where we assume u to be mean zero, $cov(y, u) = 0$ for error-in-Y, and $cov(x, u) = 0$ for error-in-X. For example, classical measurement error, perhaps the most commonly assumed error structure in economics, follows Equation 5 with the additional assumptions that $\theta = 0$, $\lambda = 1$, and $cov(y, u) = cov(x, u) = 0$ in both error-in-Y and error-in-X cases. Another special case of the linear measurement error model is Berkson error, in which a mismeasured variable is generated from a prediction or calibration equation, leading to Equation 5 with $\lambda < 1$, a formal characterization of mean-reverting measurement error (Keogh et al., 2020; Ratledge et al., 2022).

Under the most general form of the linear measurement error model, uncorrected error-in-X and error-in-Y regression models recover the following slope coefficients (see Supplementary Materials Section C.2 for details) in expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{\tilde{y}}] &= \lambda\beta + \frac{\sigma_{xu}}{\sigma_x^2} && \text{(error-in-}Y\text{)} \\
\mathbb{E}[\hat{\beta}_{\tilde{x}}] &= \beta\frac{\lambda\sigma_x^2}{\lambda^2\sigma_x^2 + \sigma_u^2} + \frac{\sigma_{yu}}{\lambda^2\sigma_x^2 + \sigma_u^2} && \text{(error-in-}X\text{)},
\end{aligned} \tag{6}$$

where σ_x^2 is the variance in the true variable x and σ_u^2 is the variance in the residuals from the linear error model, u. Covariances between errors in one variable and values of the other are indicated by σ_{xu} and σ_{yu}. When these covariances are non-zero, the error is "differential" (Carroll et al., 2006). Equation 6 recovers the standard prediction that classical measurement error (in which $\lambda = 1$ and all covariance terms are zero) causes no bias in error-in-Y models but attenuates coefficients in error-in-X models by a magnitude determined by the "reliability ratio" $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$. In practice, however, biases can additionally be influenced by λ and by differential measurement error. Under these more general conditions, biases can be present in both error-in-X and error-in-Y cases, and can lead to either attenuation or exaggeration of estimated coefficients.[12] It is clear from the substantial bias in the error-in-Y case shown in Figure 4, and from the exaggeration of coefficients in the error-in-X case shown in Figure B.3, that the assumptions of classical measurement error do not hold in this setting.

[12] For example, with $\lambda < 0$ and no differential measurement error, uncorrected error-in-Y models will exhibit attenuated coefficients $\beta_{\tilde{y}} = \lambda\beta < \beta$, while uncorrected error-in-X models can be biased in either direction, depending on the relative magnitudes of λ and the reliability ratio. The special case of non-differential Berkson measurement error involves a specific λ that exactly balances with the reliability ratio to lead to no bias in the error-in-X case (Keogh et al., 2020). With differential measurement error as in the general form in Equation 6, biases can arise in either direction for both model types.

Instead, our results suggest that remotely sensed measures are best described by a differential linear measurement error model with $\lambda < 1$.[13] To see this, first note that Figure B.4 shows substantial mean reverting measurement error, in which extreme values are systematically underestimated in remotely sensed predictions (Bound and Krueger, 1991). This behavior generates values of $\lambda < 1$,[14] which is expected when remotely sensed measures are generated from a prediction model (Keogh et al., 2020) and is consistent with evidence from remotely sensed income in Ratledge et al. (2022). Second, Figure B.5 shows non-zero covariance between errors in one variable (shown on the y-axes of each panel) and levels of other variables (shown on the x-axes of each panel). This implies that the second terms in both expressions in Equation 6 are non-zero and contribute to bias. While mean reversion, differential error, and classical measurement errors all likely play a role in coefficient bias, we show in Figure B.6 that differential measurement error is responsible for most of the biases we uncover. This figure shows that a measurement error model allowing for classical and differential measurement errors, but no mean-reverting measurement error, explains 61% of the coefficient bias in the error-in-X case and 90% of the bias in the error-in-Y case, across models (Figure B.6, bottom row). In contrast, a measurement error model that allows only for classical and mean-reverting measurement errors, but no differential measurement error, explains none of the variation in either case (Figure B.6, middle row). When we allow for all three forms of error, the linear measurement error model explains virtually all of the variation in observed biases (Figure B.6, top row). Thus, empirically it appears that differential measurement error, as opposed to mean reversion, is the most important contributor to coefficient bias introduced by remotely sensed measurements. This motivates our evaluation of multiple imputation, which has been shown in simulation to effectively correct for differential measurement error (Shaw et al., 2020), and suggests that recently proposed methods to correct for mean-reverting measurement error (Ratledge et al., 2022) may not account

[13] The Berkson error model also exhibits $\lambda < 1$, but Berkson error is non-differential and involves a specific value of λ that leads to no bias in the error-in-X case (Carroll et al., 2006).

[14] We estimate values of λ ranging from 0.47 for income to 0.9 for forest cover across our six remotely sensed variables.

for the main source of bias induced by errors in remotely sensed estimates.

The biases in standard errors shown in Figure 4 are also consistent with differential, mean-reverting measurement error. In general, reduced variance in the outcome variable in an error-in-Y model due to $\lambda < 1$ will lower the sum of squared errors, reducing estimated uncertainty in $\beta_{\tilde{y}}$, consistent with the results shown in the second row of Figure 4. In contrast, reduced variance in the independent variable will lower the estimated variance in $\tilde{x}$ and inflate standard errors, also consistent with results in Figure 4. However, the presence of differential measurement error in addition to mean reversion complicates this intuition and makes the direction of bias in uncertainty parameters theoretically ambiguous (Carroll et al., 2006).

Together, these findings highlight the importance of accounting for non-classical measurement error when estimating regressions using remotely sensed variables. Importantly, multiple imputation presents an error correction method that applies to the most general form of Equation 5, and has been shown to be effective at correcting bias in cases with mean reversion and differential error (Shaw et al., 2020; Josey et al., 2022).

4.3 Multiple imputation successfully addresses bias in parameter estimates across a diversity of regression models.

While uncorrected regression models estimated using remotely sensed measurements exhibit substantial biases, we find that multiple imputation is highly effective at correcting them. Figure 4 shows in blue the distributions of all performance metrics across all regression models after using multiple imputation. Median coefficient bias is reduced from 23% down to 2% for the error-in-X case and from 10% down to 2% in the error-in-Y case. Only 6% of the estimates were biased by more than 25% after multiple imputation was applied in the error-in-X and error-in-Y cases, compared to 49% and 29% before correction.

The standard errors estimated by multiple imputation are on average 7% and 9%

larger than those estimated using ground truth data in the error-in-X and error-in-Y cases, respectively, due to uncertainty from the first stage error correction model being incorporated into the final parameter estimates (as shown in Equation 4). Thus, multiple imputation mitigates the problem of overly precise standard errors uncovered in the uncorrected error-in-Y case, while also reducing the upward bias in standard errors uncovered in the uncorrected error-in-X case.

With low bias and standard errors that account for both sample and imputation uncertainty, models estimated using multiple imputation tend to have excellent coverage: 95% confidence intervals estimated from multiple imputation include the point estimate from the ground truth labels >90% of the time for both error-in-X and error-in-Y cases (Figure 4, third row). The additional uncertainty from imputation, however, leads multiple imputation to have marginally lower power than the uncorrected approach (mean power falls from 98% to 97% for both the error-in-X and error-in-Y cases). Thus, for simple linear regression models using remotely sensed variables, multiple imputation appears to reduce bias in parameter estimates and improve coverage at the cost of modest reductions in statistical power. Figure B.7 shows that this conclusion is consistent whether multiple imputation is implemented using a Bayesian linear regression approach, as we have used throughout our main analysis, or alternative methods such as linear regression bootstrapping or predictive mean matching.

4.4 Multiple imputation performs well with calibration sets that are small and that are distant from the main sample.

A primary cost to the researcher of implementing a correction method like multiple imputation is that a calibration set of ground truth labels co-located with both the remotely sensed measurements and the remaining variables in the analysis must be obtained. The experiments shown above use a randomly selected calibration sample of size 12,000 and impute values for a main sample size of 28,000. However, in many empirical applications it may be difficult or impossible to collect such a large and spatially well-distributed calibration dataset. Here, we assess the performance of multiple imputation under a range of calibration set sizes and under different spatial distributions of calibration data to evaluate the method's generalizability to a wider range of real-world applications.

Sample size: Figure 5 shows how the performance of multiple imputation varies with the size of the calibration set. The uncorrected model performance is depicted by the horizontal purple line, which doesn't vary with calibration set size because no calibration data are used to correct the regression estimates. Grey lines indicate bias, coverage, and power for each of the 40 regression models, while black lines and blue dots indicate median (bias) or mean (coverage and power) values over models.[15] Intuitively, coefficient bias from the models corrected using multiple imputation increases as the calibration set size falls. The rate of decrease in performance is similar in the error-in-X and error-in-Y cases, with median bias increasing from 2% with 12,000 calibration set observations to 13% with 180 observations. However, because uncorrected regressions have much higher coefficient biases in the error-in-X case, multiple imputation is on average more beneficial in this setting. Specifically, in the error-in-X case, corrected median coefficient bias remains nearly half the size of that of the uncorrected models, even with only 180 calibration observations. In the error-in-Y case, multiple imputation has lower median bias than the uncorrected model as long as the calibration set is above roughly 500 observations. With fewer observations, however, the imputation procedure is poorly constrained and bias is increased above the uncorrected model.

The second row of Figure 5 shows that estimated standard errors become increasingly inflated as calibration set size declines. This is due to increased uncertainty in the calibration model leading to increased uncertainty in the final estimates. Further, as a result of increased coefficient bias, model coverage decreases slightly with reduced calibration

[15] In Figures 5 and 6 we report median coefficient and standard error bias, but mean coverage and power. We do so because of the long tails of the distribution of bias, and because coverage and power are binary for each bootstrap run of each model. Qualitative results are similar using means and medians.

sample size, though less than proportionally with the increase in coefficient bias due to increases in the estimated standard errors (Figure 5, third row). Finally, large standard errors from small calibration samples lead to meaningful declines in statistical power: mean power drops to 74% for error-in-X and 73% for error-in-Y when the calibration set size is reduced to 180 observations.

Together, these results suggest that multiple imputation can reduce coefficient bias even in cases where only a few hundred ground truth labels can be obtained, although performance improves with the size of the calibration set. It also cautions that when only minimal calibration data can be obtained, error correction techniques may exacerbate, rather than mitigate, bias relative to using the remotely sensed measures directly, particularly in error-in-Y cases.

Spatial proximity: Another consideration in applied settings is the spatial proximity of calibration data to the main sample area of interest. In some contexts, remotely sensed data may be widely available, but it may be feasible to collect ground truth data only in a limited geographic region. For example, while remotely sensed pollution measures are ubiquitous, researchers may have access to air pollution monitors in only a handful of sparsely sampled locations. This spatial separation between calibration and main samples raises the possibility that multiple imputation will become less effective, as the structure of the measurement error estimated in the calibration data may be less applicable to main sample observations located far away. To evaluate this possibility, we design an experiment (detailed in Supplementary Materials Section C.3) in which we systematically increase the physical distance between the observations in the main and calibration datasets and record the performance of multiple imputation at each separation distance.

Figure 6 shows the four performance metrics for multiple imputation regression models (vertical axes) plotted against the distance between observations in the calibration and main datasets (horizontal axes) for both the error-in-X and error-in-Y cases. As in

Figure 5, the uncorrected model performance is depicted by the horizontal purple line. Grey lines indicate bias, coverage, and power for each of the 40 regression models, while black lines and blue dots indicate median (bias) or mean (coverage and power) values over models. The figure shows that increasing the distance between the main and calibration samples increases bias and decreases coverage, but does not substantially change parameter uncertainty or power.

In the error-in-X case, multiple imputation outperforms the uncorrected model for coefficient bias up until the maximum evaluated distance between the main and calibration dataset of 1776km (16 degrees), though median bias increases from 2% in the baseline experiment to 24%. Correspondingly, coverage gradually declines from a mean of 92% at no spatial extrapolation to close to the uncorrected level of 19% at 1776km. For the error-in-Y case, multiple imputation outperforms the uncorrected model only up to separation distances of roughly two hundred kilometers, on average. As in Figure 5, the loss of performance is similar in both cases, but the better baseline performance of the uncorrected model in the error-in-Y case leads to more limited gains from multiple imputation. Overall, the results of these spatial extrapolation calibration experiments are broadly encouraging, but also caution against relying on multiple imputation when calibration data are located very far from the main sample of interest, especially for error-in-Y settings.

Importantly, in these extreme cases, removing the calibration data from the estimating sample (i.e., using "standard" in place of "efficient" multiple imputation[16]) is a simple solution that can substantially improve bias and coverage, at the cost of reduced precision. The performance of standard multiple imputation is shown by the dotted lines in Figure 6. Across the range of spatial separation between the main and calibration samples, the bias of standard multiple imputation is roughly one half to two-thirds that of efficient

[16] Throughout the paper, we emphasize results from what is called the "efficient" version of multiple imputation, in which the calibration set is appended to the main sample when estimating the regression shown in Equation 4. Efficient multiple imputation is generally preferable as it makes the best use of all available data. Using a "standard" version of multiple imputation, where only the main sample is used in estimation, tends to provide less precise estimates, but, as we show here, can reduce bias and improve coverage when spatially extrapolating between the calibration and main samples.

multiple imputation and the coverage is roughly double in both the error-in-X and error-in-Y cases.

Together, these experiments document the returns to higher quantity and quality of calibration data when implementing multiple imputation. They also demonstrate the overall robustness of multiple imputation as an error correction method in data-limited settings. Our results suggest that multiple imputation can reduce parameter bias even with a relatively small or distant calibration set, but that in settings with extremely limited calibration data, biases can be amplified and power reduced. Importantly, across all settings analyzed, the coverage of multiple imputation models exceed that of the uncorrected models, so long as standard multiple imputation is used when spatially extrapolating.

4.5 Multiple imputation performs well relative to other correction methods.

Above, we compared bias, coverage, and power metrics between uncorrected models and models corrected with multiple imputation. Here, we additionally evaluate the performance of multiple imputation as compared to other common error correction methods, each of which is described in detail in Supplementary Materials Section C.4.

Figure B.8 shows that with a randomly distributed calibration dataset multiple imputation has lower bias and higher coverage in the error-in-X case than all other approaches we consider. The closest contender to multiple imputation is "complete case analysis," the approach of directly applying the coefficient estimated in the calibration sample to the main sample, which has almost the same performance as multiple imputation. However, complete case analysis has slightly lower coverage and its performance drops rapidly relative to multiple imputation methods at lower sample sizes (shown in Figure B.11) and with more distant calibration datasets (shown in Figure B.13).

For the error-in-Y case, the choice of error correction method is much less consequential, as all correction approaches perform quite well at lowering coefficient and standard error bias and raising coverage (see Figure B.9). All correction methods exhibit a similar decline in performance with decreased sample size and with increased spatial distance between main and calibration datasets (see Figures B.12 and B.14).[17] The only exception to this general finding is that complete case analysis leads to slightly more bias than other correction methods (see Figure B.9). As in the error-in-X case, this approach is also less robust to small and distant calibration samples (see Figures B.12 and B.13).

In sum, these results are consistent with prior literature on statistical error correction, demonstrating that multiple imputation approaches outperform other error correction methods for most metrics. However, multiple imputation is rarely, if ever, empirically evaluated against these alternative error correction methods in experiments that examine the influence of sample size or spatial separation between calibration and main samples (McNeish, 2017). Overall, we find robust evidence that multiple imputation is as good as or better than common alternative error correction techniques, even when calibration data are limited or spatially distant.

4.6 Multiple imputation performs well in fixed effects experimental designs.

The analyses above use remotely sensed variables in cross-sectional simple linear regression frameworks. In practice, however, many researchers use remotely sensed variables in more complex research designs. We conduct two experiments to evaluate the potential bias introduced by remotely sensed variables, as well as the efficacy of multiple imputation, in such contexts.

First, we replicate the analysis shown in Figure 4 using a cross-sectional fixed effects research design, rather than simple linear regression. By including state fixed effects in the estimating equation, we identify the relationship between the outcome and independent

[17] External calibration methods generally perform poorly, but exhibit little decay in performance with changes in calibration data quantity and quality. We address this approach in the Discussion.

variable using only within-state variation across 1km×1km grid cells. To implement this, for each pair of variables and each bootstrap sample, we first residualize all variables with respect to the fixed effects in both the main sample and the calibration sample. We then perform all analysis steps as outlined in previous sections.[18] Figure B.15 shows that the challenges to parameter estimation from using remotely sensed variables in the simple regression framework are replicated when using spatial fixed effects, as is the ability of multiple imputation to address them. This suggests that the threats to estimation posed by errors in remotely sensed variables may not be easily remedied by flexible spatial controls, and that multiple imputation is effective at reducing bias and improving coverage in research designs relying on spatial fixed effects.

Second, we consider using remotely sensed predictions in a panel data setting with fixed effects and a set of control variables. Specifically, we replicate the main specification from Deschenes, Greenstone and Shapiro (2017)'s study of the effects of the U.S. NO_x budget program on ambient air pollution. This regression model uses panel data and a suite of spatial and temporal fixed effects to construct a "triple-difference" research design in which $\text{PM}_{2.5}$ is compared across states, years, and seasons that are (versus are not) covered by the NO_x budget program.[19] The estimating equation is:

$$\text{PM2.5}_{cst} = \beta \mathbb{1}\{NBP\,Operating\}_{cst} + \boldsymbol{W}_{cst}\rho + \mu_{ct} + \upsilon_{cs} + \nu_{st} + \epsilon_{cst} \tag{7}$$

where c is county, s is season (either "summer" or "winter", each of which is six months), and t is year. PM2.5_{cst} is the ambient concentration of $\text{PM}_{2.5}$ measured in $\mu g/m^3$ and $\mathbb{1}\{NBPOperating\}_{cst}$ is an indicator function that is equal to one when the NO_x budget

[18] This approach is motivated by the Frisch-Waugh-Lovell Theorem (Lovell, 2008). As fixed effects are control variables like any other, they can be projected out of the outcome and treatment variables, rather than being directly controlled for, without any implication for recovered coefficients of interest. Note that this approach is more computationally efficient than including the fixed effects directly, and it removes issues in settings where a fixed effect needed for prediction in the main sample is not identifiable in the calibration sample, such as with spatial fixed effects and spatially disjoint main and calibration sets. This is discussed further in the replication of Deschenes, Greenstone and Shapiro (2017) below.

[19] Note that the NO_x budget program was only operational during summer months. See Deschenes, Greenstone and Shapiro (2017) for details on the empirical design.

program is operational for a given county, season, and year. β is the coefficient of interest and indicates the influence of the NO_x budget program on ambient $PM_{2.5}$. $\boldsymbol{W}_{cst}$ is a matrix of weather controls, including precipitation, temperature and dew point temperature,[20] μ_{ct} is a set of county-by-year fixed effects, υ_{cs} is a set of county-by-season fixed effects, and ν_{st} is a set of season-by-year fixed effects.

The authors used daily $PM_{2.5}$ concentrations from ground monitors for their analysis, interpolated to the county-season-year level. Here, we show that directly substituting these data for satellite-derived $PM_{2.5}$ leads to substantial bias in the coefficient of interest, β. To do so, we first replicate the finding from Deschenes, Greenstone and Shapiro (2017) that the NO_x budget program reduced average county $PM_{2.5}$ by 1.03 $\mu g/m^3$. Our result, in Table 1, column 1 is identical to the authors' estimate in their Table 2, column 5. Second, we replace the authors' station data with $PM_{2.5}$ observations from the widely-used remotely sensed Van Donkelaar et al. (2021) dataset and re-estimate Equation 7. We find that using remotely sensed air pollution data attenuates the estimated effect of the budget program on air pollution by ∼50% relative to the original paper and reduces standard errors by ∼35% (Table 1, column 2). In turn, the 95% confidence interval estimated in column 2 does not contain the original point estimate from Deschenes, Greenstone and Shapiro (2017). This suggests that the previous findings of substantial bias, overly precise standard errors, and low coverage in the error-in-Y case shown above using data from Rolf et al. (2021) generalize to the context of satellite-based air pollution data and a fixed effects panel data setting with high-dimensional controls.

We then evaluate whether multiple imputation can effectively correct for the estimated bias. To mimic the likely situation that the calibration set would be available in certain geographical units but not others, we randomly assign full county time series from the replication dataset into either a main sample (70% of the counties) or calibration sample (30% of the counties). This reflects a setting where, for example, air pollution monitor data are available in a handful of counties, but not in all of them. To ensure there are

[20] Temperature and dew point temperature are represented using the share of days in each county-season-year that fall within a set of 20 bins defined by quantiles of the daily distribution.

adequate calibration samples for both the participating ("treated") and non-participating ("control") counties, we sample the main and calibration sets proportionally from the treated and control counties identified in Deschenes, Greenstone and Shapiro (2017).[21]

A rich set of fixed effects are used in Deschenes, Greenstone and Shapiro (2017) to create a triple-difference research design. While this lends credibility to the causal interpretation of the estimated β coefficient, these fixed effects introduce some complexity into the application of multiple imputation. In particular, treating these fixed effects as standard controls by including them directly in both stages of multiple imputation is infeasible. This is because the county-specific fixed effects μ_{ct} and υ_{cs} can be estimated in the first (i.e., imputation) step of multiple imputation only for the counties falling into the calibration set. Making predictions from this model in the main sample is impossible because the fixed effects for those main sample counties are unknown.

We address this challenge by residualizing all regression variables by μ_{ct} and υ_{cs} prior to the multiple imputation analysis, effectively controlling for these county-year and county-season fixed effects throughout the entire multi-step procedure.[22] Note that because the season-year fixed effects ν_{st} are estimated using data from all counties in the sample, they can be treated identically to other controls in the regression. After residualization, we conduct multiple imputation in a standard manner, using the calibration sample to estimate how residualized ground truth pollution relates to residualized remotely sensed pollution and the other residualized model controls, including the season-year fixed effects (analagously to Equation 3). Denoting residualized variables with a double dot superscript, as in $\mathrm{P\ddot{M}2.5}$, we estimate:

$$\mathrm{P\ddot{M}2.5}_{cst} = \gamma \mathrm{P\tilde{\ddot{M}}2.5}_{cst} + \psi \ddot{\mathbb{1}\{NBP\,Operating\}}_{cst} + \ddot{\boldsymbol{W}}_{cst}\varrho + \ddot{\nu}_{st} + e_{cst}. \tag{8}$$

Then in the main dataset (where we assume ground truth data for $\text{PM}_{2.5}$ are not avail-

[21] "Treated" counties are in states that have an operational NO_x budget program at some point in the sample, while "control" counties are in all other states investigated by Deschenes, Greenstone and Shapiro (2017).

[22] This approach is, again, motivated by the Frisch-Waugh-Lovell Theorem (Lovell, 2008).

able), we use the estimated calibration model from Equation 8 to impute residualized values of true $\mathrm{PM}_{2.5}$, K times. We can then perform K estimations of the following regression, using each of the imputed datasets k (analogously to Equation 4):

$$\mathrm{P}\hat{\ddot{\mathrm{M}}}2.5^{k}_{cst} = \beta^{k}\mathbb{1}\{NBP\ddot{O}perating\}_{cst} + \ddot{\boldsymbol{W}}_{cst}\rho^{k} + \ddot{\nu}^{k}_{st} + \epsilon^{k}_{cst} \tag{9}$$

In both Equations 8 and 9, standard errors are clustered by state-season, following Deschenes, Greenstone and Shapiro (2017). Coefficients and standard errors are recovered from across K multiple imputations using Rubin's Rule. As in all the experiments above, we use efficient multiple imputation, which includes the calibration dataset along with the main dataset when estimating Equation 9. We repeat the random splitting of the data into main and calibration sets and the multiple imputation analysis 200 times to recover a distribution of bias and of the effectiveness of multiple imputation.[23] Reported point estimates and standard errors from multiple imputation are calculated as means over the 200 runs; median estimates are nearly identical.

Column 3 in Table 1 shows that multiple imputation removes ∼60% of the bias in the coefficient introduced by the remotely sensed pollution data and ∼40% of the bias in the standard errors.[24] Moreover, the corrected 95% confidence interval now contains the original point estimate. These findings demonstrate the ability of multiple imputation to generalize to a regression framework exploiting panel data with a large set of semi-parametric control variables.

It is important to note that in a fixed effects research design, the exact implementation of multiple imputation will depend on the chosen fixed effects and the spatial and temporal structure of the calibration and main samples. While we have shown one adaptation of multiple imputation designed to meet the needs of a particular dataset and research

[23] Note that in practice, researchers would only have access to one main dataset and one calibration dataset. Here, we bootstrap the entire procedure in order to capture sampling variability, as we have access to both ground truth and remotely sensed pollution in all counties.

[24] Note that the $\mathrm{PM}_{2.5}$ data from Van Donkelaar et al. (2021) rely on remotely sensed variables as well as other inputs, such as station data and a chemical transport model. This highlights the applicability of multiple imputation to error-prone predicted variables beyond those that are purely remotely sensed.

design, this procedure can easily be adjusted to match the nature of the calibration data available and the structure of the fixed effects estimated in other contexts. Residualizing fixed effects when they cannot be included as standard controls due to the structure of the calibration set, while including all other fixed effects and control variables in both stages of multiple imputation, generalizes easily to most settings and appears to be effective.

Here, we have considered remotely sensed variables in a panel data setting, but prior work has similarly documented the effectiveness of multiple imputation in panel data settings where measurement error arises for other reasons, such as survey non-response, entry errors, or inability to comprehensively track individuals over time in a longitudinal study (e.g., De Silva et al., 2019, 2017; Spratt et al., 2010; Nevalainen, Kenward and Virtanen, 2009).[25] However, to our knowledge, this prior literature has not evaluated a spatially structured calibration set, as we have done here by removing entire county time series from the main sample, nor explored how to handle fixed effects that cannot be treated as standard controls in multiple imputation. This spatial structure, and its implications for fixed effects estimation, is very likely to be a key feature of available calibration data in many applied economics settings, and it is encouraging to see that even in this complex empirical setting, multiple imputation performs well at mitigating parameter biases.

5 Discussion

As the uses and benefits of remotely sensed data continue to expand across many disciplines, it is increasingly important that the challenges these data raise be examined and that corresponding solutions be identified, tested, and improved. In this paper, we make progress toward these goals by evaluating the risks that measurement errors in remotely sensed data pose for parameter recovery in regression analyses. First, we quantify the

[25] There is a large literature investigating the effectiveness of multiple imputation in longitudinal study settings, but to our knowledge none consider the estimation of panel data models with spatial and/or temporal fixed effects. In some longitudinal simulation studies, multiple imputation has been found to exhibit variable performance, depending on assumptions about error structure (Twisk et al., 2013).

biases introduced by remotely sensed variables when used in downstream regression analyses. We uncover substantial bias in regression coefficients and associated standard errors when using a large set of remotely sensed environmental and economic variables in simple linear regression models. Second, we demonstrate that a standard statistical technique for imputation of missing data, multiple imputation, performs well at mitigating these biases across a diversity of contexts, as long as researchers have access to some amount of ground truth data. These results apply most directly to studies leveraging remotely sensed variables in regression analysis, but are relevant more broadly to analyses relying on machine learning predictions in downstream regressions (Wang, McCormick and Leek, 2020).

These results call into question the findings of previous papers that directly use remotely sensed measures in regression analyses without correction. However, there are a few important features and limitations to keep in mind. First, the empirical returns to using multiple imputation as a correction method are higher for remotely sensed measurements with lower predictive power. For example, Figure B.16 shows that within the Rolf et al. (2021) benchmark dataset used throughout this paper, variables with higher R^2 in the underlying remote sensing model exhibit lower bias when used in regression analysis without correction, and therefore have a lower value of applying multiple imputation. As remotely sensed measurements improve, biases introduced in downstream regression analyses are likely to become smaller. However, the growing use of remotely sensed socioeconomic indicators, which tend to be more difficult to sense than directly visible natural phenomena like forest cover, indicates that measurement error and its correction will remain important considerations.

Second, we have assumed throughout the analysis that ground-based measurements are fully accurate. Of course, measurement error is also present in traditional data collection methods and threatens parameter recovery in traditional analyses as well (Little and Rubin, 2019). In some cases, particularly in contexts without established data management systems, ground-based measurements may actually exhibit larger measurement error than satellite-based predictions (Lobell et al., 2020). However, remotely sensed measurement error is often substantially larger than errors we consider in traditional settings; for example, satellite-based estimates of income or wealth generally can explain just a half to two-thirds of the variation in ground truth data (Chi et al., 2022; Jean et al., 2016; Yeh et al., 2020; Rolf et al., 2021; Ratledge et al., 2022). Thus, while ground truth data are rarely perfectly measured, there is a much larger scope for parameter bias arising from such substantial remotely sensed prediction errors.

Third, multiple imputation is under-studied within the context of the empirical models most often employed by applied economists. While we have shown that multiple imputation performs well in a triple-difference panel fixed effects research design used in prior work, more theoretical and simulation-based analysis is necessary to fully characterize the benefits and limitations of multiple imputation in panel data settings.

Finally, in most of the results emphasized here, we have assumed researchers have access to a calibration dataset in which ground truth measurements are available for *both* the dependent and independent variables (see data availability regimes in Figure 2). This is called "internal calibration" in the statistics literature. However, in some cases researchers may only have access to an "external calibration" dataset, in which ground truth data are available only for the remotely sensed measure (whether it is the dependent or independent variable). We show in Figures B.8 and B.9 that multiple imputation becomes much less effective in this setting.

In sum, we show across a variety of settings that multiple imputation is highly effective at reducing parameter biases introduced into regression analysis due to remotely sensed measurements. While there are important limitations to its effectiveness that should be considered, this method is simple, easy to implement via existing packages in software platforms such as `R`, `Stata`, and `Python`,[26] and it generalizes well across a wide range of empirical contexts, including when calibration data are limited and when regression

[26] There are different packages available for multiple imputation in most platforms. For example, `mice` is available in `R`, `IterativeImputer` is available within `scikit-learn` in `Python`, and a variety of `mi` commands are available in `Stata`.

models leverage panel data and standard fixed effects research designs.

References

Alix-Garcia, Jennifer and Daniel L Millimet. 2020. "Remotely incorrect." *August* 24(2020):5.

Balboni, Clare, Robin Burgess, Anton Heil, Jonathan Old and Benjamin A Olken. 2021. Cycles of fire? Politics and forest burning in Indonesia. In *AEA Papers and Proceedings*. Vol. 111 pp. 415–19.

BenYishay, Ariel, Silke Heuser, Daniel Runfola and Rachel Trichler. 2017. "Indigenous land rights and deforestation: Evidence from the Brazilian Amazon." *Journal of Environmental Economics and Management* 86:29–47.

Bound, John and Alan B Krueger. 1991. "The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?" *Journal of labor economics* 9(1):1–24.

Boutwell, James L and John V Westra. 2013. "Benefit transfer: A review of methodologies and challenges." *Resources* 2(4):517–527.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski and Ciprian M Crainiceanu. 2006. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chen, Joyce J, Valerie Mueller, Yuanyuan Jia and Steven Kuo-Hsin Tseng. 2017. "Validating migration responses to flooding using satellite and vital registration data." *American Economic Review* 107(5):441–45.

Chen, Shuai, Paulina Oliva and Peng Zhang. 2022. "The effect of air pollution on migration: evidence from China." *Journal of Development Economics* 156:102833.

Chi, Guanghua, Han Fang, Sourav Chatterjee and Joshua E Blumenstock. 2022. "Microestimates of wealth for all low-and middle-income countries." *Proceedings of the National Academy of Sciences* 119(3).

Cole, Stephen R, Haitao Chu and Sander Greenland. 2006. "Multiple-imputation for measurement-error correction." *International journal of epidemiology* 35(4):1074–1081.

Crost, Benjamin, Claire Duquennois, Joseph H Felter and Daniel I Rees. 2018. "Climate change, agricultural production and civil conflict: Evidence from the Philippines." *Journal of Environmental Economics and Management* 88:379–395.

De Silva, Anurika Priyanjali, Margarita Moreno-Betancur, Alysha Madhu De Livera, Katherine Jane Lee and Julie Anne Simpson. 2017. "A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study." *BMC medical research methodology* 17(1):1–11.

De Silva, Anurika Priyanjali, Margarita Moreno-Betancur, Alysha Madhu De Livera, Katherine Jane Lee and Julie Anne Simpson. 2019. "Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study." *BMC medical research methodology* 19(1):1–14.

Deschenes, Olivier, Michael Greenstone and Joseph S Shapiro. 2017. "Defensive investments and the demand for air quality: Evidence from the NOx budget program." *American Economic Review* 107(10):2958–89.

Fowlie, Meredith, Edward Rubin and Reed Walker. 2019. Bringing satellite-based air quality estimates down to earth. In *AEA Papers and Proceedings*. Vol. 109 pp. 283–88.

Freedman, Laurence S, Douglas Midthune, Raymond J Carroll and Victor Kipnis. 2008. "A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression." *Stat. Med.* 27(25):5195–5216.

Freeman, Richard, Wenquan Liang, Ran Song and Christopher Timmins. 2019. "Willingness to pay for clean air in China." *Journal of Environmental Economics and Management* 94:188–216.

Fuller, Wayne A. 1995. "Estimation in the presence of measurement error." *International Statistical Review/Revue Internationale de Statistique* pp. 121–141.

Garcia, Alberto and Robert Heilmayr. 2022. "Conservation impact evaluation using remotely sensed data." *Available at SSRN 4179782* .

Gibson, John, Susan Olivia, Geua Boe-Gibson and Chao Li. 2021. "Which night lights data should we use in economics, and where?" *Journal of Development Economics* 149:102602.

Hansen, Matthew C, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland et al. 2013. "High-resolution global maps of 21st-century forest cover change." *Science* 342(6160):850–853.

Harari, Mariaflavia. 2020. "Cities in bad shape: Urban geometry in India." *American Economic Review* 110(8):2377–2421.

Heft-Neal, Sam, Jennifer Burney, Eran Bendavid, Kara K Voss and Marshall Burke. 2020. "Dust pollution from the Sahara and African infant mortality." *Nature Sustainability* 3(10):863–871.

Jain, Meha. 2020. "The benefits and pitfalls of using satellite data for causal inference." *Review of Environmental Economics and Policy* .

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." *Science* 353(6301):790–794.

Josey, Kevin P, Priyanka deSouza, Xiao Wu, Danielle Braun and Rachel Nethery. 2022. "Estimating a Causal Exposure Response Function with a Continuous Error-Prone Exposure: A Study of Fine Particulate Matter and All-Cause Mortality." *Journal of Agricultural, Biological and Environmental Statistics* pp. 1–22.

Keogh, Ruth H and Ian R White. 2014. "A toolkit for measurement error correction, with a focus on nutritional epidemiology." *Stat. Med.* 33(12):2137–2155.

Keogh, Ruth H, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis et al. 2020. "STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment." *Statistics in medicine* 39(16):2197–2231.

Kocornik-Mina, Adriana, Thomas KJ McDermott, Guy Michaels and Ferdinand Rauch. 2020. "Flooded cities." *American Economic Journal: Applied Economics* 12(2):35–66.

Little, Roderick JA and Donald B Rubin. 2019. *Statistical analysis with missing data.* Vol. 793 John Wiley & Sons.

Liu, Yang and Anindya De. 2015. "Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study." *International journal of statistics in medical research* 4(3):287.

Lobell, David B, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic and Siobhan Murray. 2020. "Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis." *American Journal of Agricultural Economics* 102(1):202–219.

Lovell, Michael C. 2008. "A simple proof of the FWL theorem." *The Journal of Economic Education* 39(1):88–91.

Marx, Benjamin, Thomas M Stoker and Tavneet Suri. 2019. "There is no free house: Ethnic patronage in a Kenyan slum." *American Economic Journal: Applied Economics* 11(4):36–70.

McNeish, Daniel. 2017. "Missing data methods for arbitrary missingness with small samples." *Journal of Applied Statistics* 44(1):24–39.

Nevalainen, Jaakko, Michael G Kenward and Suvi M Virtanen. 2009. "Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification." *Statistics in medicine* 28(29):3657–3669.

Potapov, Peter, Svetlana Turubanova, Matthew C Hansen, Alexandra Tyukavina, Viviana Zalles, Ahmad Khan, Xiao-Peng Song, Amy Pickens, Quan Shen and Jocelyn Cortez. 2022. "Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century." *Nature Food* 3(1):19–28.

Proctor, Jonathan. 2021. "Atmospheric opacity has a nonlinear effect on global crop yields." *Nature Food* 2(3):166–173.

Rahimi, Ali and Benjamin Recht. 2007. "Random features for large-scale kernel machines." *Advances in neural information processing systems* 20.

Ratledge, Nathan, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler and Marshall Burke. 2022. "Using machine learning to assess the livelihood impact of electricity access." *Nature* 611(7936):491–495.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht and Solomon Hsiang. 2021. "A generalizable and accessible approach to machine learning with global satellite imagery." *Nature communications* 12(1):1–11.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys.* Vol. 81 John Wiley & Sons.

Rubin, Donald B and Nathaniel Schenker. 1991. "Multiple imputation in health-care databases: An overview and some applications." *Statistics in medicine* 10(4):585–598.

Shaw, Pamela A, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Ruth H Keogh, Victor Kipnis, Janet A Tooze, Michael P Wallace, Helmut Küchenhoff et al. 2020. “STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics.” *Statistics in medicine* 39(16):2232–2263.

Sims, Katharine RE and Jennifer M Alix-Garcia. 2017. “Parks versus PES: Evaluating direct and incentive-based land conservation in Mexico.” *Journal of Environmental Economics and Management* 86:8–28.

Spratt, Michael, James Carpenter, Jonathan AC Sterne, John B Carlin, Jon Heron, John Henderson and Kate Tilling. 2010. “Strategies for multiple imputation in longitudinal studies.” *American journal of epidemiology* 172(4):478–487.

Sterne, Jonathan AC, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood and James R Carpenter. 2009. “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.” *Bmj* 338.

Swenson, Jennifer J, Catherine E Carter, Jean-Christophe Domec and Cesar I Delgado. 2011. “Gold mining in the Peruvian Amazon: global prices, deforestation, and mercury imports.” *PloS one* 6(4):e18875.

Triglav-Čekada, Mihaela, Fabio Crosilla and Mojca Kosmatin-Fras. 2009. “A simplified analytical model for a-priori LiDAR point-positioning error estimation and a review of LiDAR error sources.” *Photogrammetric Engineering & Remote Sensing* 75(12):1425–1439.

Twisk, Jos, Michiel de Boer, Wieke de Vente and Martijn Heymans. 2013. “Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis.” *Journal of clinical epidemiology* 66(9):1022–1028.

van Buuren, Stef. 2012. *Flexible Imputation of Missing Data.* Chapman & Hall/CRC Interdisciplinary Statistics Philadelphia, PA: Chapman & Hall/CRC.

Van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate imputation by chained equations in R.” *Journal of statistical software* 45:1–67.

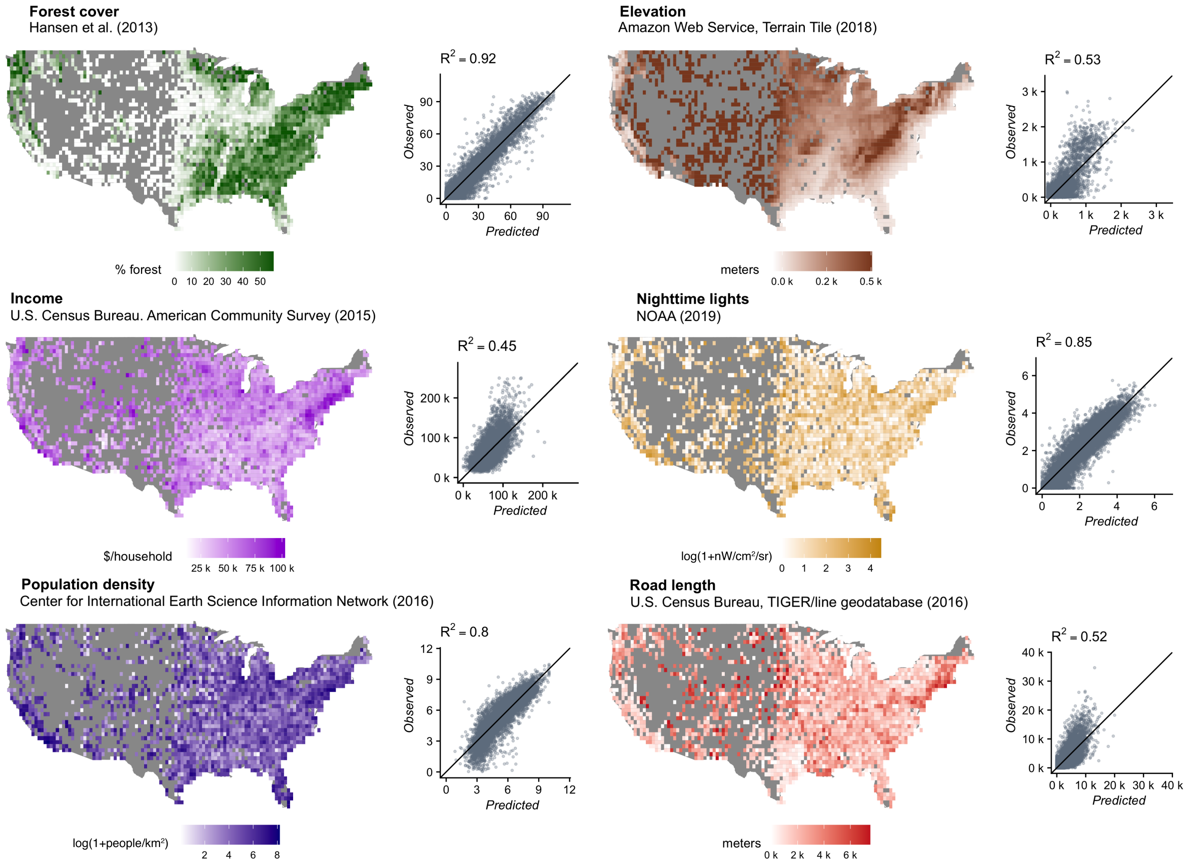
Van Donkelaar, Aaron, Melanie S Hammer, Liam Bindle, Michael Brauer, Jeffery R Brook, Michael J Garay, N Christina Hsu, Olga V Kalashnikova, Ralph A Kahn, Colin Lee et al. 2021. “Monthly global estimates of fine particulate matter and their uncertainty.” *Environmental Science & Technology* 55(22):15287–15300.

Wang, Siruo, Tyler H McCormick and Jeffrey T Leek. 2020. “Methods for correcting inference based on outcomes predicted by machine learning.” *Proceedings of the National Academy of Sciences* 117(48):30266–30275.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon and Marshall Burke. 2020. “Using publicly available satellite imagery and deep learning to understand economic well-being in Africa.” *Nature communications* 11(1):1–11.

Figures



Forest cover
Hansen et al. (2013)

Elevation
Amazon Web Service, Terrain Tile (2018)

Income
U.S. Census Bureau. American Community Survey (2015)

Nighttime lights
NOAA (2019)

Population density
Center for International Earth Science Information Network (2016)

Road length
U.S. Census Bureau, TIGER/line geodatabase (2016)

Figure 1: Ground-truth labels and remotely sensed predictions of forest cover, elevation, income, nighttime lights, population density, and road length from Rolf et al. (2021). Maps show ground-truth labels for 80,000 1km x 1km grid cells which were sampled with a population-weighted uniform-at-random sampling scheme from across the continental United States and are aggregated to 20km x 20km for visualization. Scatters show the relationship between ground truth (y-axis) and remotely sensed predictions (x-axis) for each variable. Text under each variable name gives the original ground-truth data source.

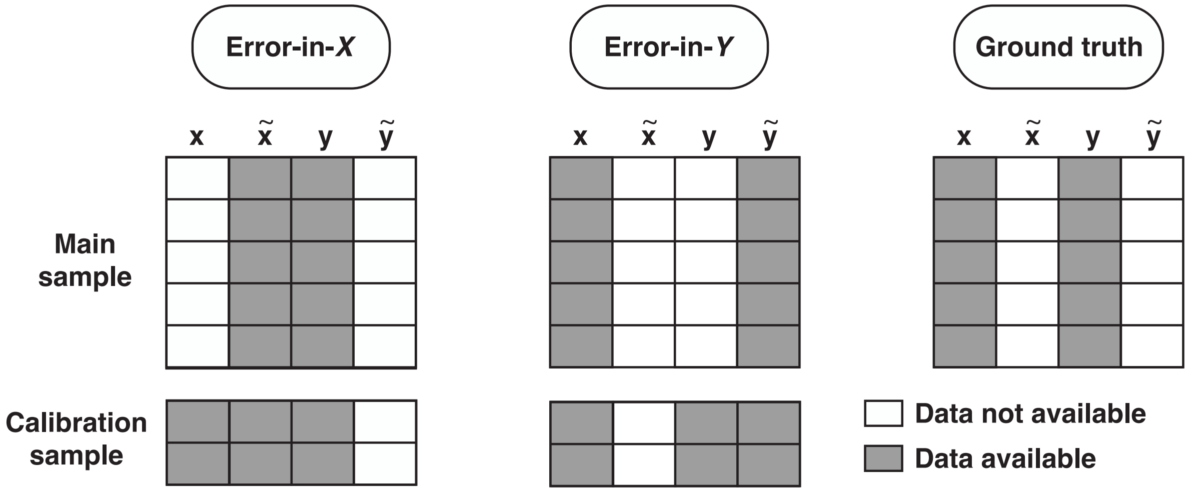


Figure 2: Three data availability regimes with different implications for parameter recovery and bias correction. Figure shows three possible scenarios for data availability, each of which is evaluated in this analysis. First, in the error-in-X case shown on the left, the main analysis sample includes ground truth data for the dependent variable y, but only remotely sensed measurements for the independent variable x (denoted $\tilde{x}$). The calibration sample, which is generally smaller than the main sample, includes these observations *plus* ground truth observations for the independent variable x. In contrast, the error-in-Y case shown in the middle column includes ground truth x and remotely sensed $\tilde{y}$ in the main sample, and additional ground truth y in the calibration sample. In the ground truth case, no bias is present and no calibration is necessary, as ground truth observations are available in the entire main sample.

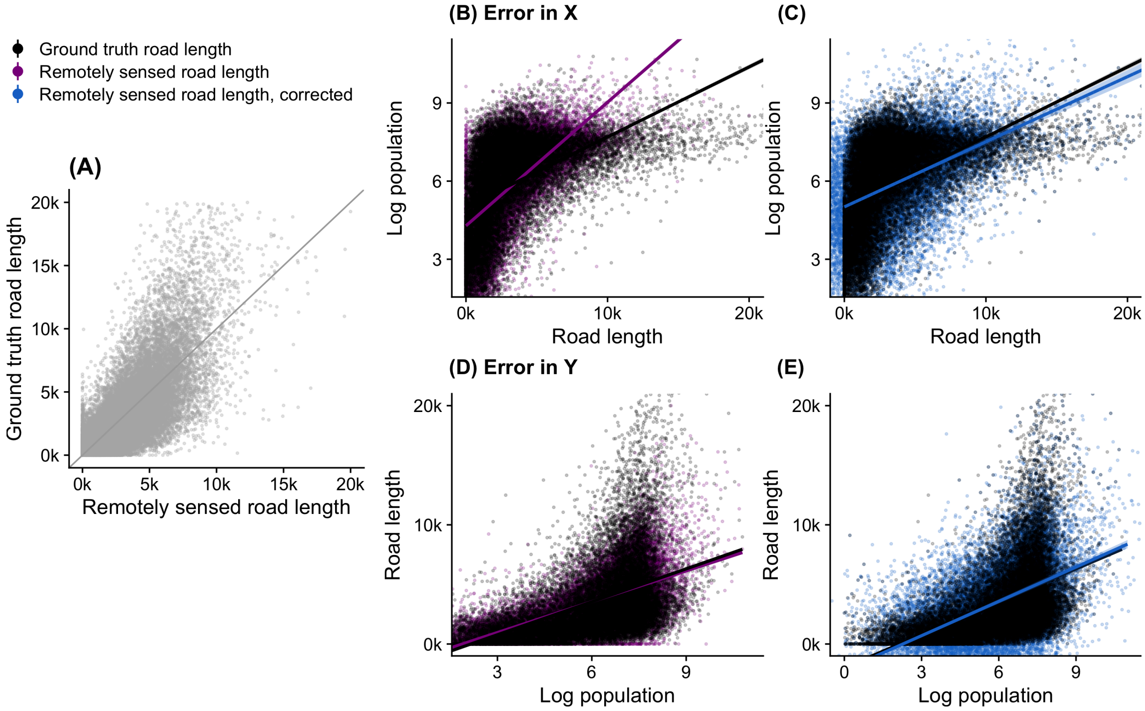


Figure 3: Remotely sensed predictions introduce bias in parameters recovered in downstream regression analyses: example from predictions of road length. Figure uses ground truth population data and remotely sensed predictions of road length to illustrate how measurement error in remotely sensed predictions can bias downstream regression coefficients and estimated standard errors. Panel **(A)** shows ground truth observations of total road length within 1km × 1km grid cells on the y-axis, plotted against satellite-based predictions of total road length on the x-axis. Panel **(B)** shows that the measurement error evidenced in panel **(A)** leads to a biased estimate of the relationship between population density and road length. Panel **(C)** shows data points and a regression model both adjusted using multiple imputation to correct the bias introduced by remotely sensed road length. Panels **(D)** and **(E)** repeat these analyses in an "error-in-Y" model in which remotely sensed road length is the outcome variable and population density is the independent variable. In this case, bias from measurement error is minimal, but multiple imputation corrects for overly precise standard error estimates.

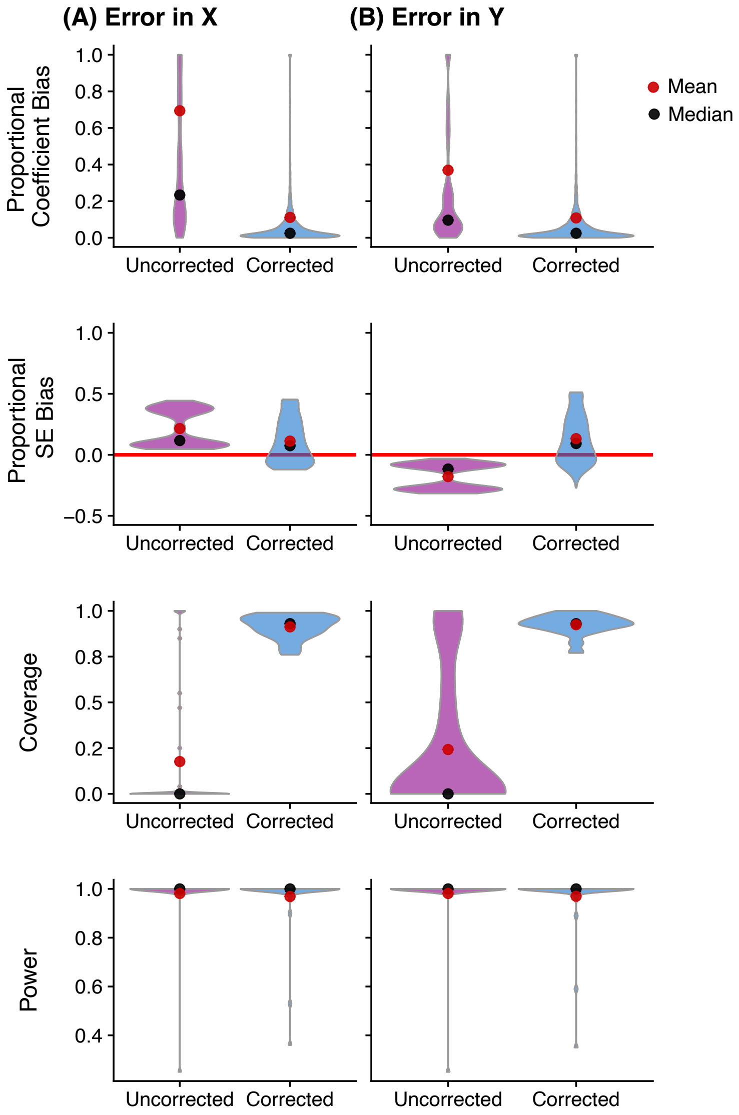


Figure 4: Bias, coverage, and power for regression models using remotely sensed variables both with and without correction via multiple imputation. Figure shows the distribution of bias, coverage, and power over a set of 100 bootstrapped estimates of 40 regression models, each of which estimates the relationship between two socioeconomic and/or environmental variables (e.g., income and temperature; road length and forest cover). Purple distributions indicate regression models in which remotely sensed variables are used without correction as either a dependent (panel **(A)**, "error-in-X") or independent (panel **(B)**, "error-in-Y") variable, while blue distributions indicate regression models in which multiple imputation was used with a corresponding calibration set to correct bias in recovered parameter estimates. The top two rows show the proportional bias in regression coefficients and standard errors (where 0.25 indicates a 25% bias), while the bottom two rows show coverage and power. Data for violin plots has been winsorized for visual display purposes only.

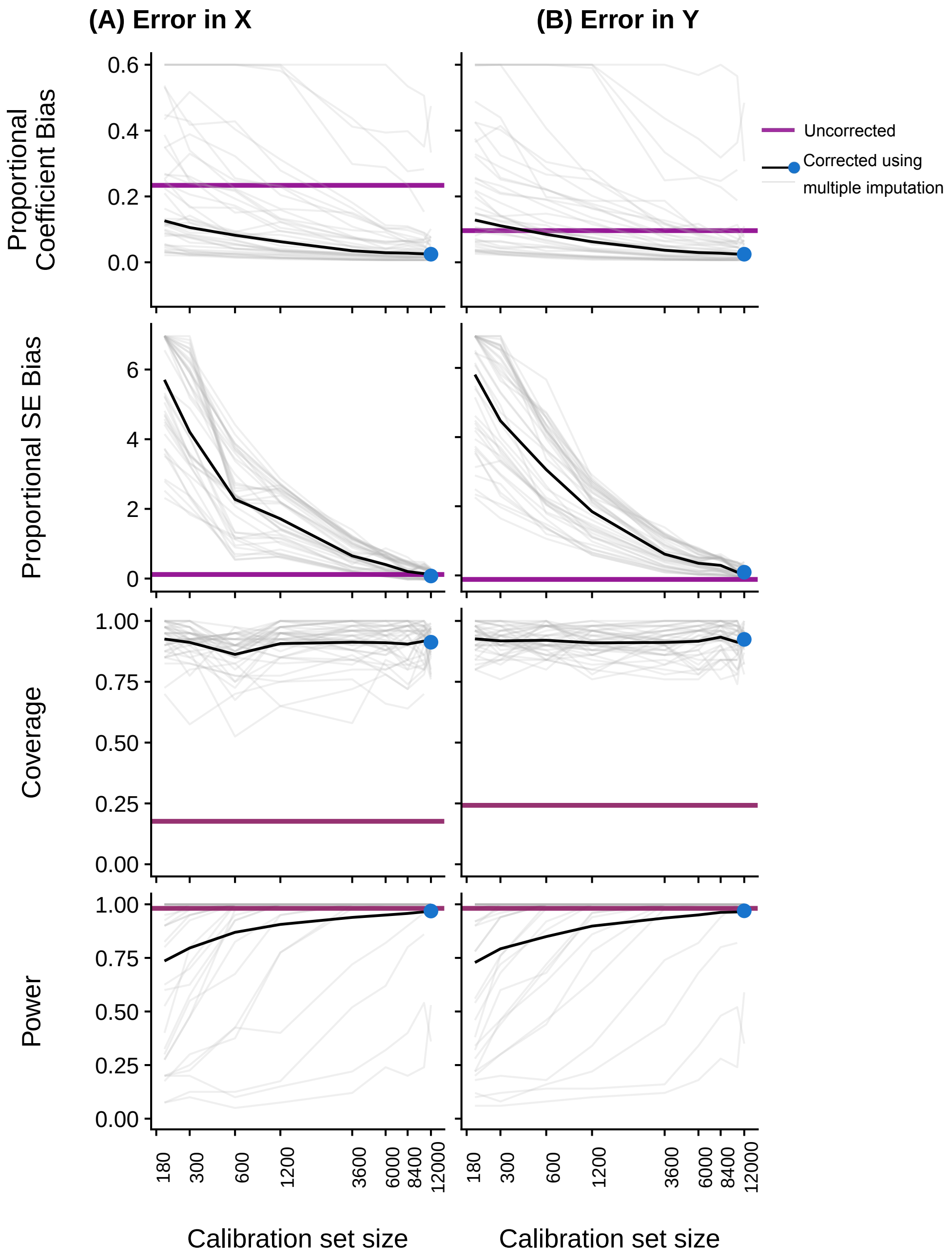


Figure 5: The effect of calibration set size on the ability of multiple imputation to correct biases introduced by remotely sensed variables. Figures show median bias, mean coverage, and mean power as a function of the size of the dataset available for calibration in the multiple imputation procedure. Horizontal purple lines show values for the uncorrected model, which does not rely on a calibration set. Solid black lines show median bias and mean coverage and power values across all regression models. Light grey lines show these measures for each of 40 regression models estimating the relationship between two socioeconomic and/or environmental variables (winsorized for display). Blue dots indicate values for a calibration set size of 12,000, as is used throughout the rest of the analysis (e.g., in Figure 4). As in Figure 4, panel **(A)** shows results for regressions in which remotely sensed variables are used as independent variables (i.e., "error-in-X"), while panel **(B)** shows results for regressions in which remotely sensed variables are used as dependent variables (i.e., "error-in-Y").

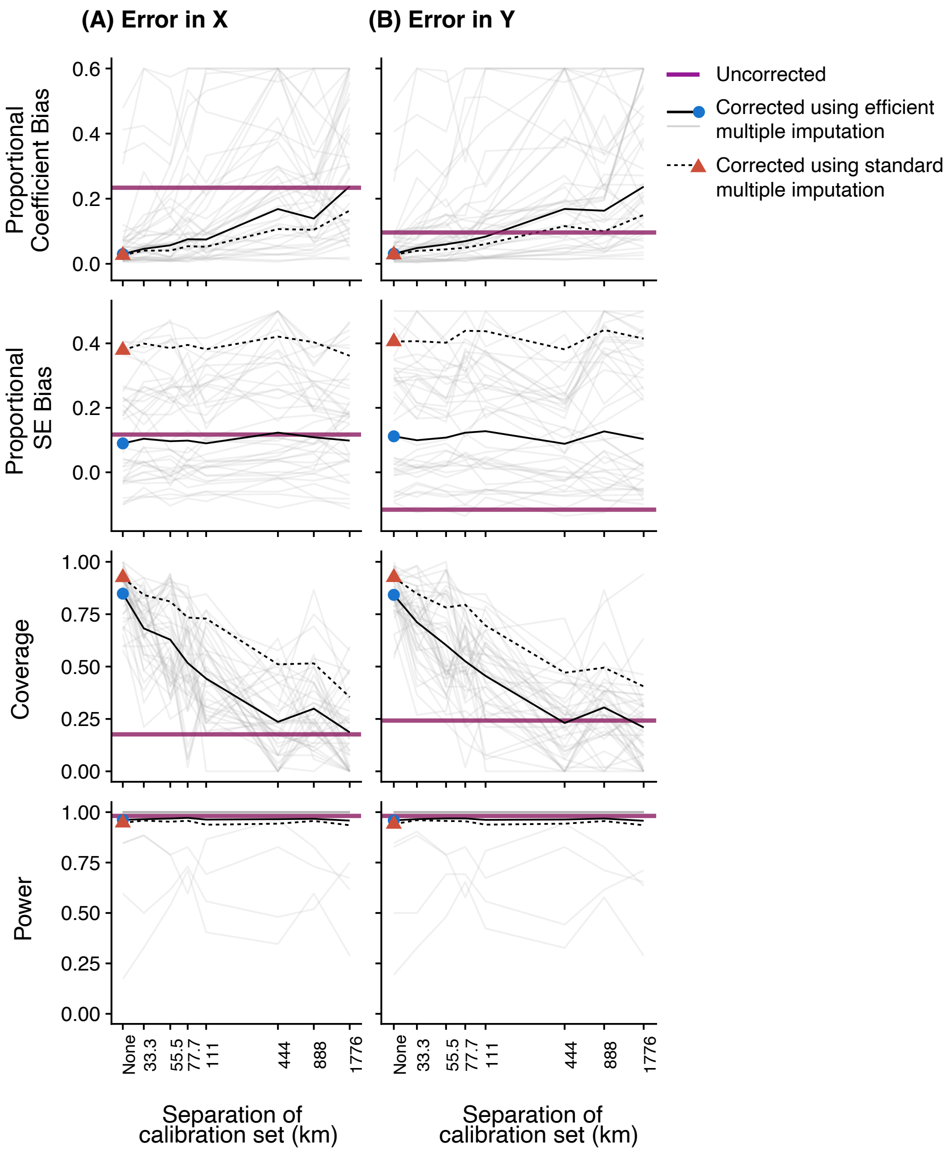


Figure 6: The effect of distance between calibration and main datasets on the ability of multiple imputation to correct biases introduced by remotely sensed variables. Figures show median bias, mean coverage, and mean power as a function of the distance between the calibration set and the main regression sample. Horizontal purple lines indicate values for the uncorrected model, which does not rely on a calibration set. Solid black lines show median bias and mean coverage and power values across all regression models. Light grey lines show these measures for each of 40 regression models estimating the relationship between two socioeconomic and/or environmental variables (winsorized for display). Blue dots indicate average values for a random sampling of the calibration set (i.e., no spatial separation between calibration and main samples imposed). Dotted lines show values for a "standard" version of multiple imputation, where, unlike the "efficient" version of multiple imputation used throughout the text, the calibration set is not appended to the main set when estimating the parameter of interest. As in Figure 4, panel **(A)** shows results for regressions in which remotely sensed variables are used as independent variables (i.e., "error-in-X"), while panel **(B)** shows results for regressions in which remotely sensed variables are used as dependent variables (i.e., "error-in-Y").

	Dependent variable:		
	$PM_{2.5}$ Ground monitor	$PM_{2.5}$ Satellite Uncorrected	$PM_{2.5}$ Satellite Multiple Imputation
	(1)	(2)	(3)
NO_x budget program	−1.03 (0.27)	−0.52 (0.18)	−0.82 (0.22)
t-statistic	−3.80	−2.95	−3.73
Main sample (N)	4,172	4,172	2,912
Calibration sample (N)			1,260

Table 1: Replication of Deschenes, Greenstone and Shapiro (2017) using remotely sensed air pollution, with and without correction via multiple imputation. Column (1) shows the paper's original estimate and standard errors of the effect of the NO_x budget program on ambient $PM_{2.5}$. Standard errors and t-statistic are calculated from the full original dataset, where standard errors are clustered at the state-season level. Column (2) shows the same estimate using uncorrected satellite $PM_{2.5}$ data from Van Donkelaar et al. (2021) in place of ground monitor data. Standard errors and t-statistic are computed identically to column (1). Column (3) shows point estimates and standard errors corrected using multiple imputation, where 30% of the data was used as a calibration dataset. The 70/30 split of the data sample is done with 200 bootstrap samples and parameters shown (including standard errors and the t-statistic) reflect means across this distribution of bootstrap samples (median estimates are nearly identical).

Parameter recovery using remotely sensed variables

Supplementary Materials

Jonathan Proctor ⓡ Tamma Carleton ⓡ Sandy Sum

Proctor: Harvard University, jproctor1@fas.harvard.edu. Carleton: University of California, Santa Barbara, tcarleton@ucsb.edu. Sum: University of California, Santa Barbara, sandysum@ucsb.edu.

A Supplementary Tables

Variable	Units	Native resolution
Forest cover	% forest cover	$\sim$ 30 m × 30 m
Elevation	meters	$\sim$ 611.5 m × 611.5 m
Population density	log(people per sq. km)	$\sim$ 1 km × 1 km
Nighttime lights	log(nanoWatts /cm^2/sr)	$\sim$ 500 m × 500 m
Income	USD per household	census block group
Road length	meters	polyline
Housing price	USD per sq. ft.	geocoded point data

Table A.1: Description of variables obtained from Rolf et al. (2021). Table lists the variables that form the benchmark dataset leveraged throughout this analysis. Each variable is measured both using remote sensing and using ground truth observations. Most of the ground truth data are based on measurements from 2010-2015. Ground-truth values (labels) for these tasks are assembled from publicly available data, and all remotely sensed predictions are made available by the authors at `https://codeocean.com/capsule/6456296/tree/v2`. More details on the original data are described in Rolf et al. (2021).

B Supplementary Figures

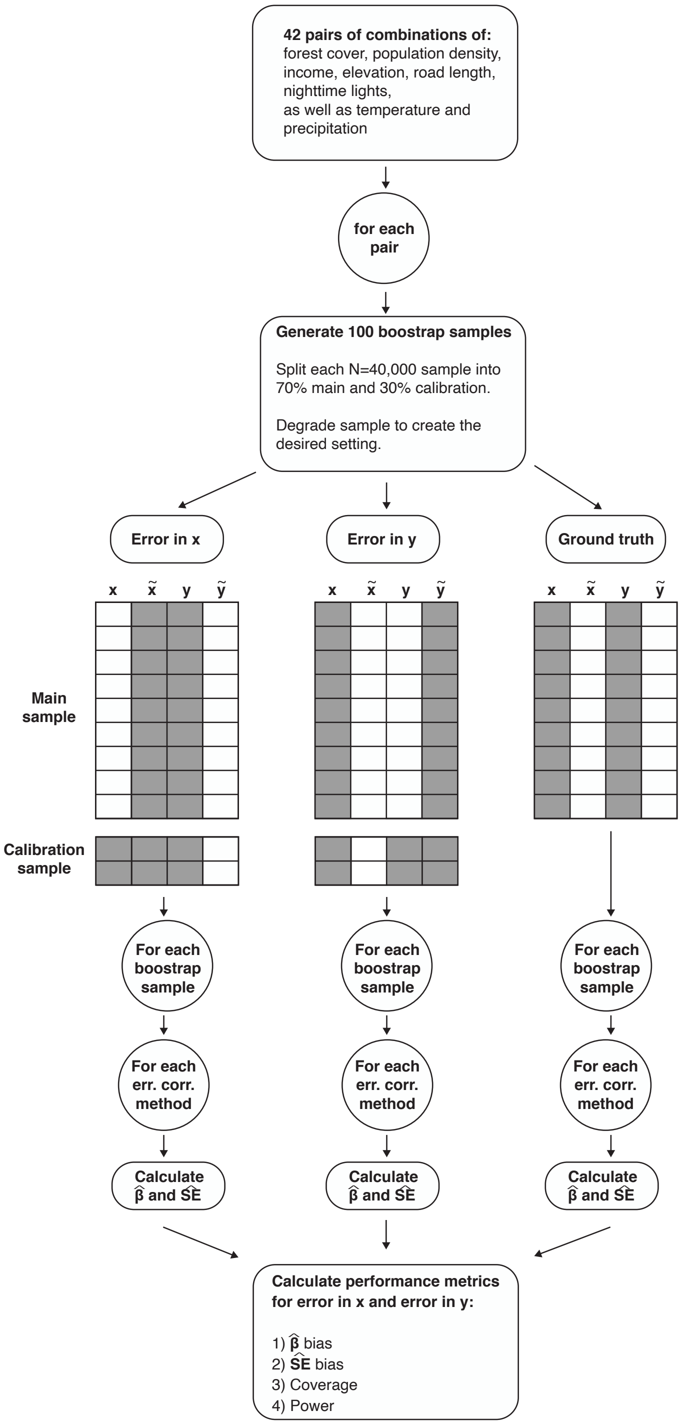

42 pairs of combinations of:
forest cover, population density, income, elevation, road length, nighttime lights,
as well as temperature and precipitation
for each pair
Generate 100 boostrap samples
Split each N=40,000 sample into 70% main and 30% calibration.
Degrade sample to create the desired setting.
Error in x
Error in y
Ground truth
x
x̃
y
ỹ
Main sample
Calibration sample
For each boostrap sample
For each err. corr. method
Calculate $\hat{\beta}$ and $\widehat{SE}$
Calculate performance metrics for error in x and error in y:
1) $\hat{\beta}$ bias
2) $\widehat{SE}$ bias
3) Coverage
4) Power

Figure B.1: Experimental design used to evaluate the impact of remotely sensed variables on downstream regression analyses. Schematic outlines experiments used to evaluate the bias, coverage, and power implications of measurement error introduced into regression analysis by remotely sensed variables.

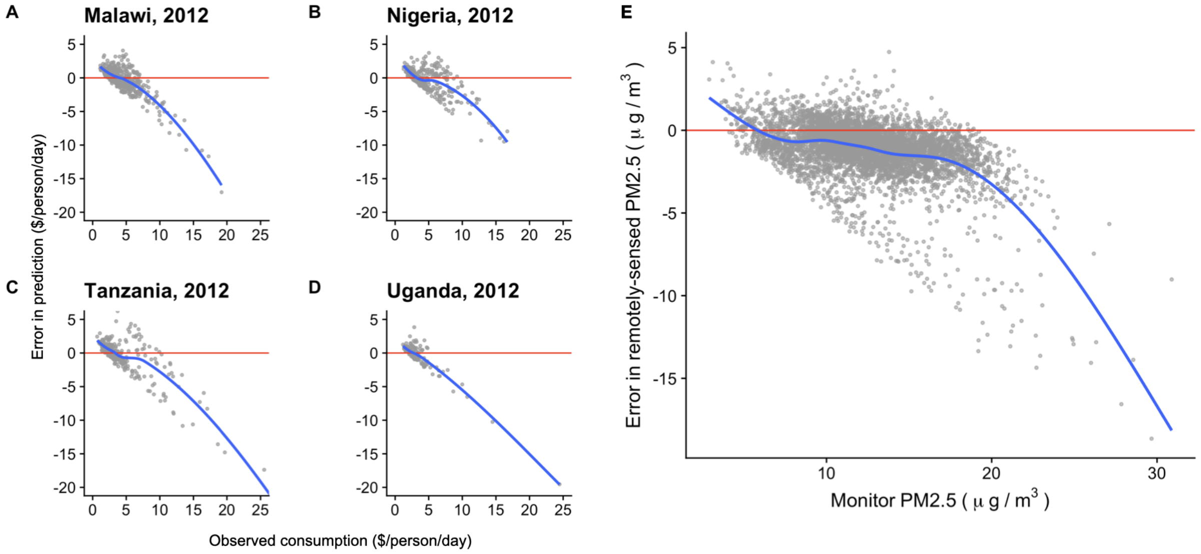


Figure B.2: Remotely sensed consumption from Jean et al. (2016) and remotely sensed PM$_{2.5}$ from Van Donkelaar et al. (2021) display mean-reverting measurement error. Panels A-D show errors in consumption predicted from satellite images (measured in \$/person/day) by Jean et al. (2016) plotted against corresponding ground truth values of consumption for four different countries in Africa. Panel E shows errors in remotely sensed county-level PM$_{2.5}$ (measured in $\mu g/m^3$) from Van Donkelaar et al. (2021) plotted against U.S. EPA monitor-based PM$_{2.5}$ over the years 2001-2007. Downward slopes in all panels indicate mean-reverting measurement error (Bound and Krueger, 1991), consistent with $\lambda < 1$ in Equation 5.

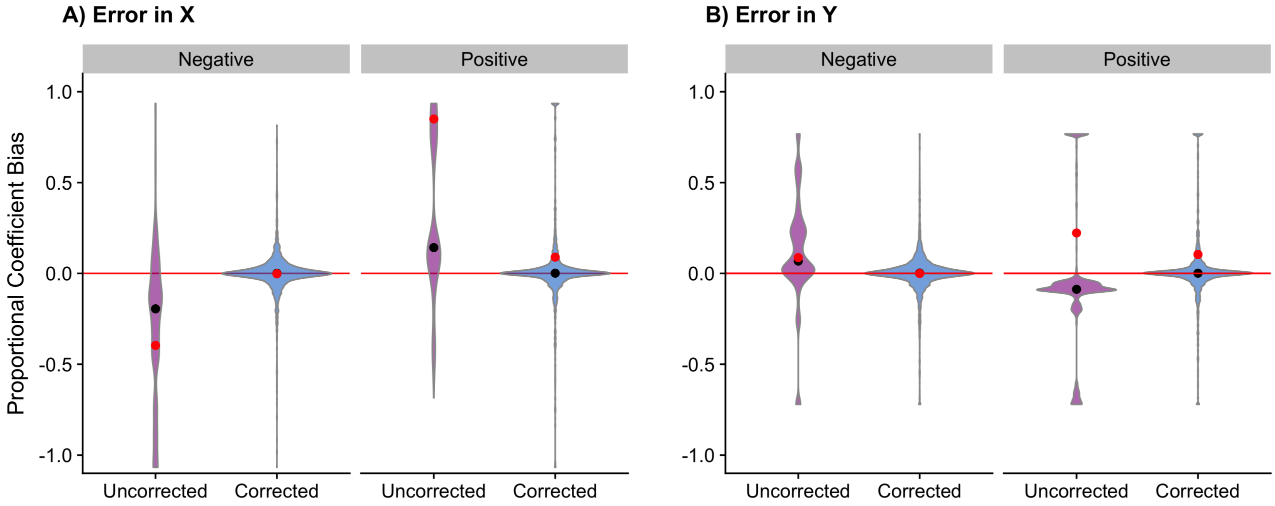


Figure B.3: Remotely sensed variables tend to inflate coefficients in error-in-X models and attenuate coefficients in error-in-Y models. Figure decomposes results from the top row of Figure 4 into coefficient biases in regression models with true slope coefficients less than zero (labeled "Negative") versus those with true slope coefficients greater than zero (labeled "Positive"). Here, proportional bias is signed and is computed as $\frac{\hat{\tilde{\beta}}-\hat{\beta}}{\hat{\beta}}$, following notation from Section 3. Using this definition, attenuation bias would appear as positive bias for true negative coefficients (first column of each panel) and as negative bias for true positive coefficients (second column of each panel). While there is large heterogeneity displayed, uncorrected error-in-X models tend to exhibit coefficient inflation, while uncorrected error-in-Y models tend to exhibit attenuation (shown in purple). In contrast, coefficients corrected using multiple imputation (shown in blue) display minimal bias and exhibit no systematic inflation or attenuation of coefficients. Data for violin plots has been winsorized at the 2.5% and 97.5% levels to cap outliers for visual display purposes only. The unwinsorized mean is indicated by the red circle while the median is indicated by the black circle.

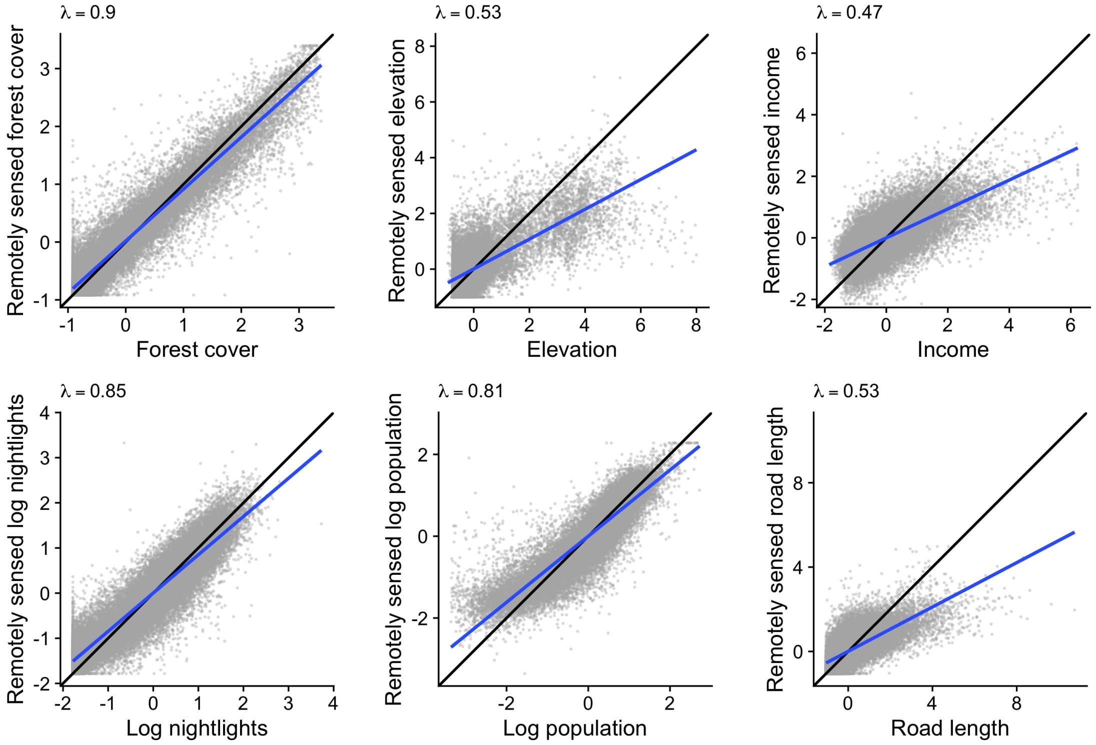


Figure B.4: Remotely sensed predictions of six diverse variables exhibit mean-reverting measurement error. Figure shows the relationship between remotely sensed values ($y-$axis) and ground truth values (x-axis) for all six variables used throughout the analysis and obtained from Rolf et al. (2021). Slope coefficients λ for each subplot correspond to the Equation 5 and are indicated visually by the blue line. The 45 degree line is indicated in black. All estimated values of λ are less than unity, demonstrating that these remotely sensed predictions exhibit mean-reverting measurement error, which contributes to bias in both error-in-X and error-in-Y regression models (see Equation 6).

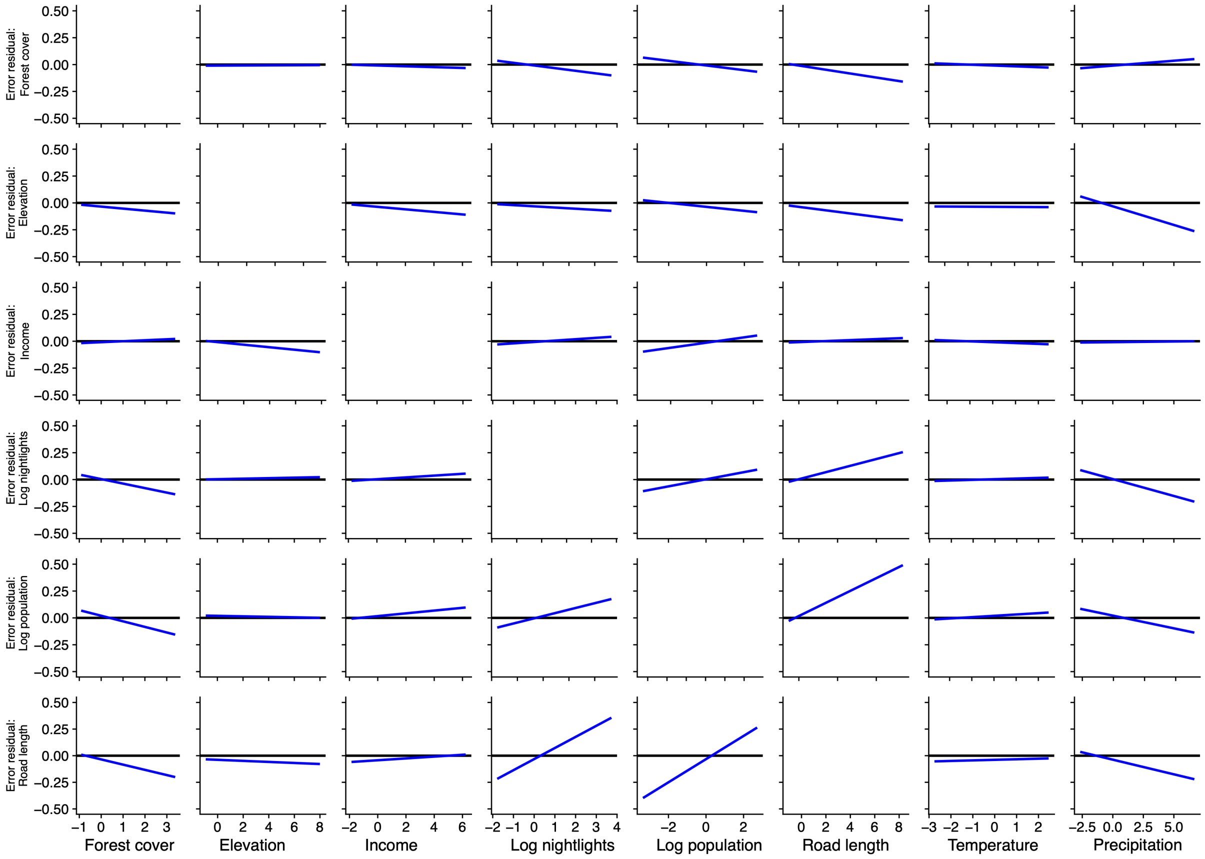


Figure B.5: Relationships between error residuals in one variable and levels of another variable show differential measurement error for many variable pair combinations. Residuals from the linear measurement error model in Equation 5 are plotted on the y-axis against ground truth values of all other variables on the x-axis. Off-diagonal boxes show the covariance between a variable's remotely sensed residual errors and the true value of *another* variable. Any nonzero correlations indicate the presence of differential measurement error in remotely sensed variables (i.e., non-zero values of σ_{yu} or σ_{xu} in Equation 5). Diagonals are omitted as they provide information about λ (shown in Figure B.4), not information about differential error.

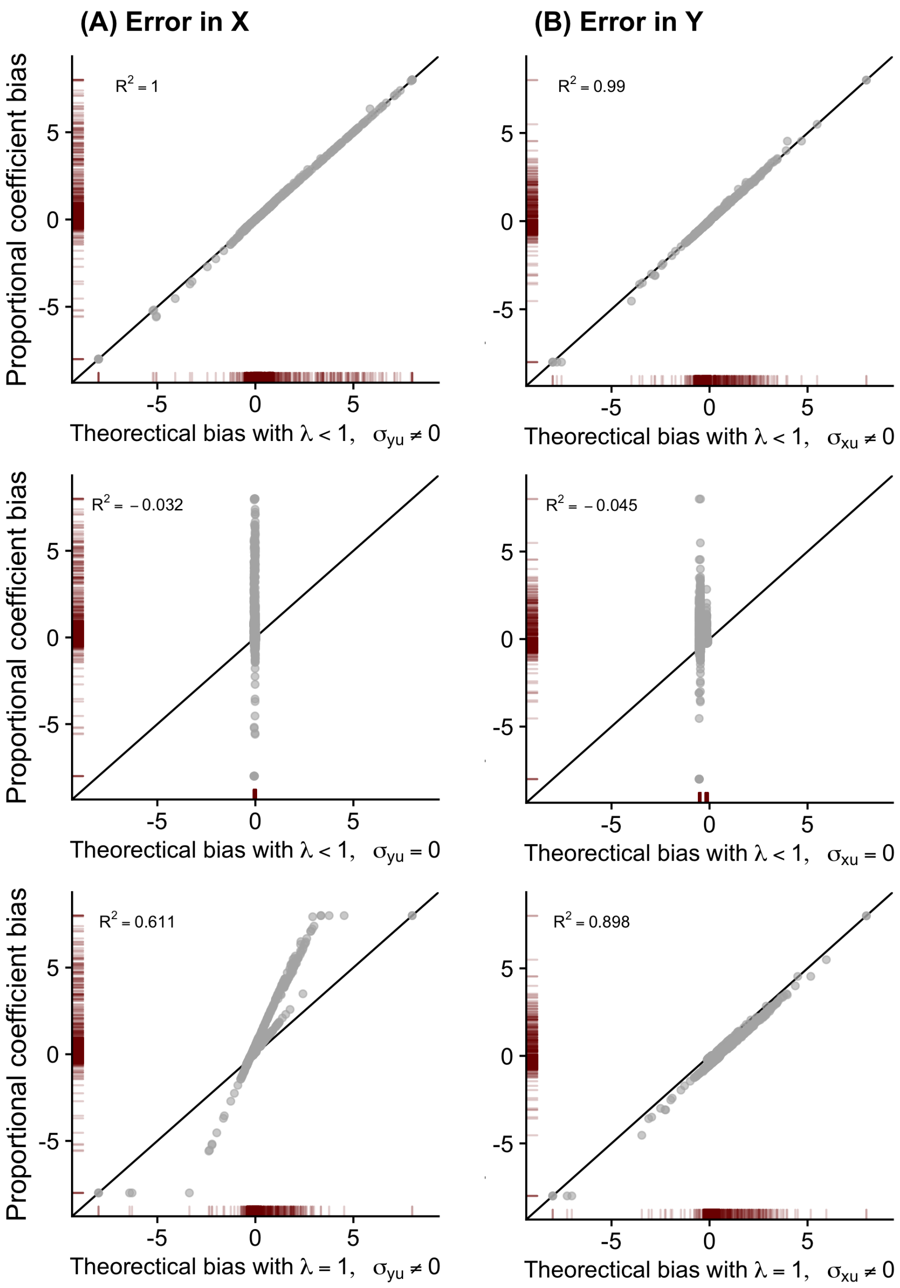


Figure B.6: Decomposing the source of bias introduced by remotely sensed measurements in downstream regression analyses. As detailed in Section 4.2, both mean reverting measurement error and differential measurement error can contribute to bias in error-in-X and error-in-Y regression models. This figure decomposes overall bias to show that differential measurement error is the most important factor in explaining the coefficient biases we recover. Each row plots observed coefficient bias (y-axis) against the theoretically predicted bias that would arise under alternative assumptions about measurement error structure (x-axis). The bottom row assumes no mean reversion, but allows for the presence of differential measurement error. The middle row assumes non-differential measurement error, but allows for mean reversion. The top row is the most general, and allows for both mean reversion and differential measurement error. This top row shows that together, these two forces explain all of the observed coefficient biases we recover across our diverse regression models.

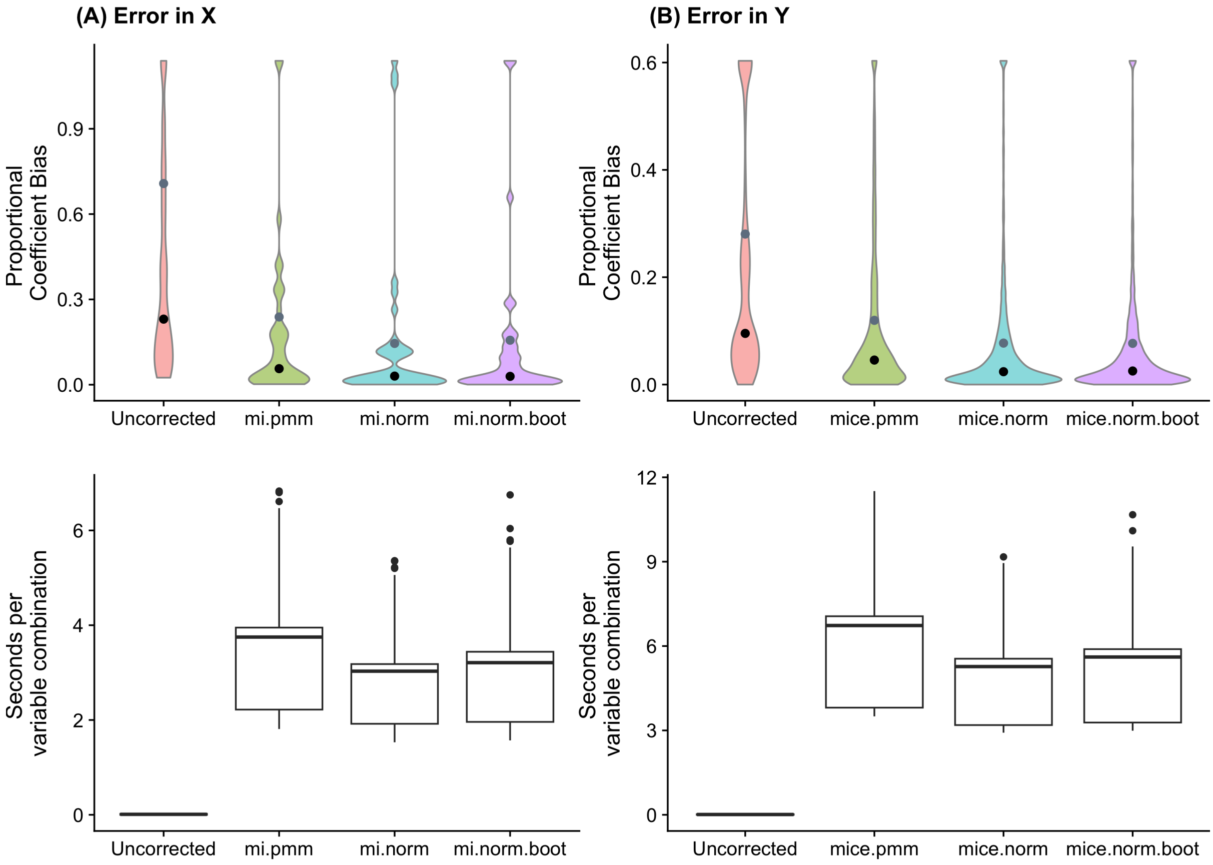


Figure B.7: Proportional coefficient bias and mean computation time for commonly used multiple imputation methods. Figure shows performance and compute time for the three most commonly used multiple imputation methods implemented in the `mice` package in `R`. The top row shows proportional coefficient bias while the bottom row shows the relative time taken to perform the bias correction for a single variable ($N = 40,000$). Results indicate that the Bayesian linear regression specification for multiple imputation (labeled "mi.norm"), which we use throughout our analysis, out-performs predictive mean matching (labeled "mi.pmm") and bootstrap linear regression (labeled "mi.norm.boot"), while being computationally the least demanding. Results shown throughout the main text are from error correction method "mi.norm".

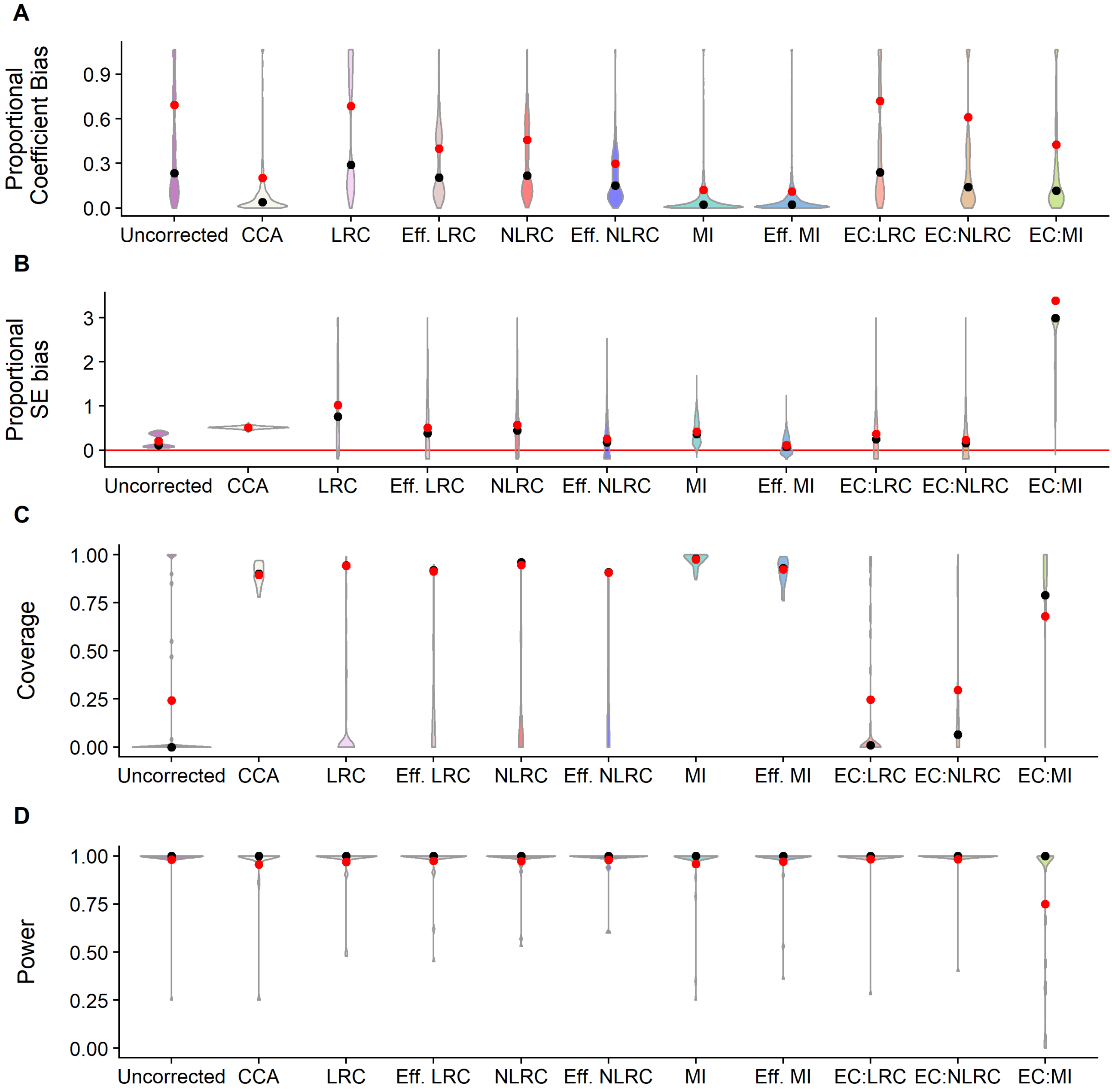


Figure B.8: Performance of alternative error correction models in the error-in-X case. Figure shows proportional coefficient bias, proportional standard error bias, coverage, and power across all regression models tested for each of nine error correction models, one data subsampling approach, and the uncorrected regression approach. Data for violin plots has been winsorized at the 2.5% and 97.5% to cap outliers for visual display purposes only. The unwinsorized mean is indicated by the red circles and the corresponding median is in black. Error correction approaches are detailed in Section C.4.

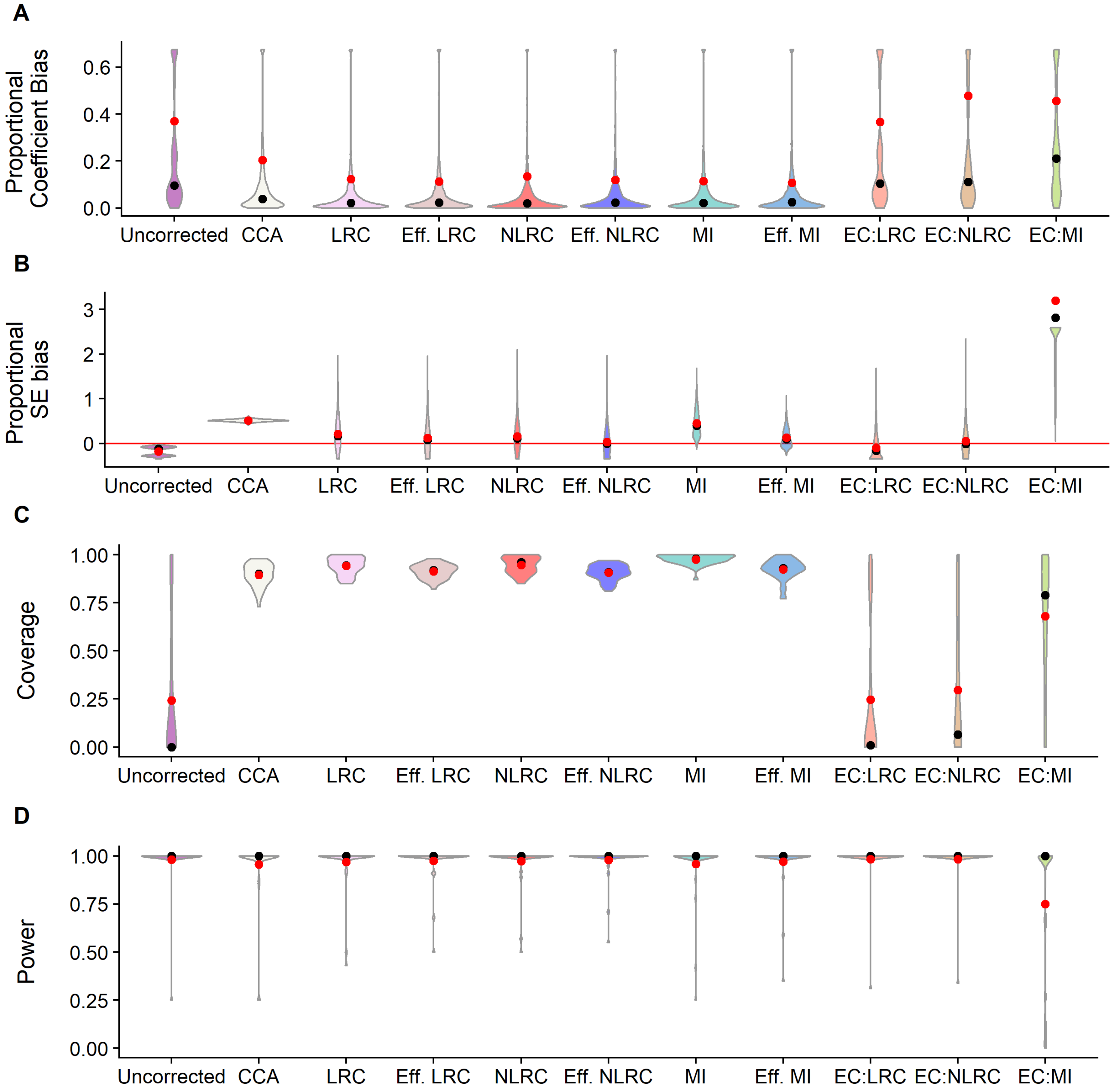


Figure B.9: Performance of alternative error correction models in the error-in-Y case. Figure shows proportional coefficient bias, proportional standard error bias, coverage, and power across all regression models tested for each of nine error correction models, one data subsampling approach, and the uncorrected regression approach. Data for violin plots has been winsorized at the 2.5% and 97.5% to cap outliers for visual display purposes only. The unwinsorized mean is indicated by the red circles and the corresponding median is in black. Error correction approaches are detailed in Section C.4.

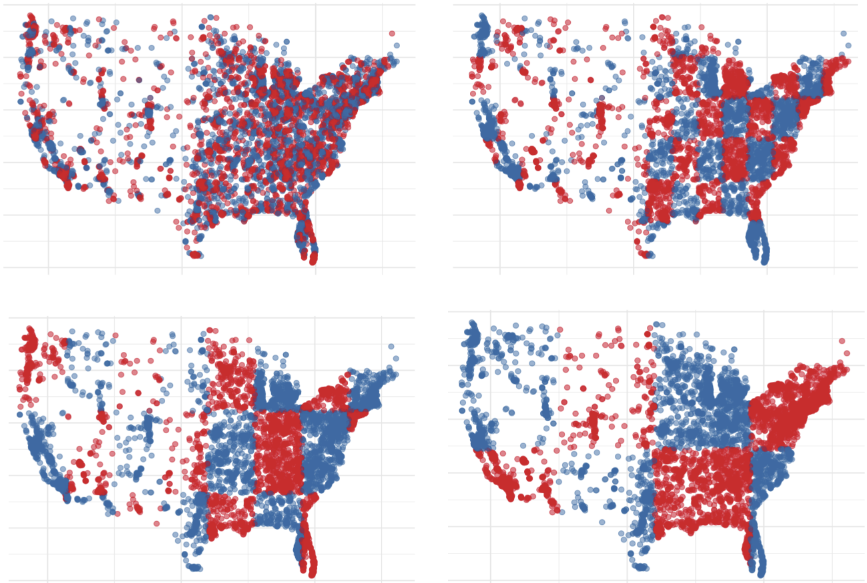


Figure B.10: Illustration of the data availability regime used to test multiple imputation when the calibration sample is spatially separated from the main sample. Figure visually displays how the calibration and main datasets are separated in the spatial experiment outlined in Section 4.4. From top left, clockwise: when separation distance is 1, 4, 8, and 16 degrees. To implement this experiment, we draw the main sample from the red boxes and the calibration sample from the blue boxes. As the width of the grid becomes larger, bias correction using multiple imputation becomes more difficult, as any point in the main sample is, in expectation, farther in physical distance from a point in the calibration sample. For each spatial separation distance, we "jitter" the grid up, down, and left for different bootstrap runs to ensure that results are representative across geographic divisions.

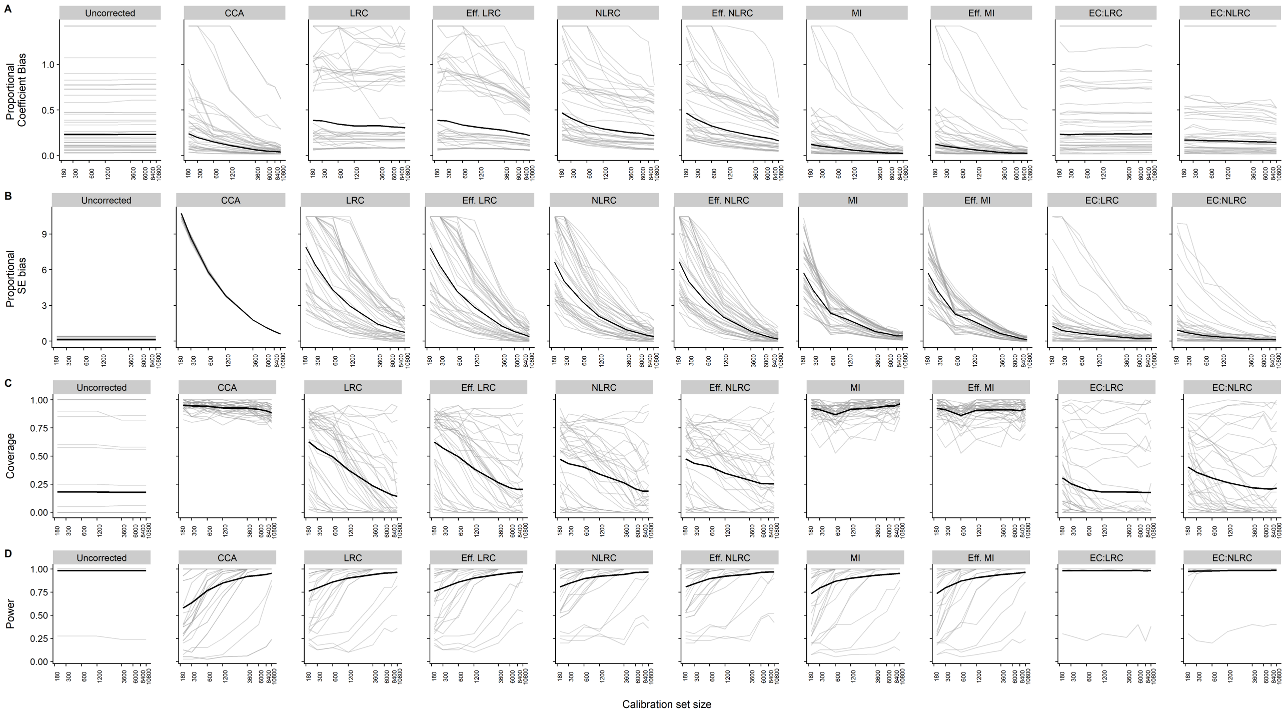


Figure B.11: Effects of calibration sample size on bias, coverage, and power across all alternative error correction techniques (error-in-X case). Proportional coefficient bias, proportional standard error bias, coverage, and power across all models plotted against calibration sample size for each error correction technique. Results can be interpreted as in Figure 5, but include additional error correction approaches beyond multiple imputation. Error correction approaches are detailed in Section C.4.

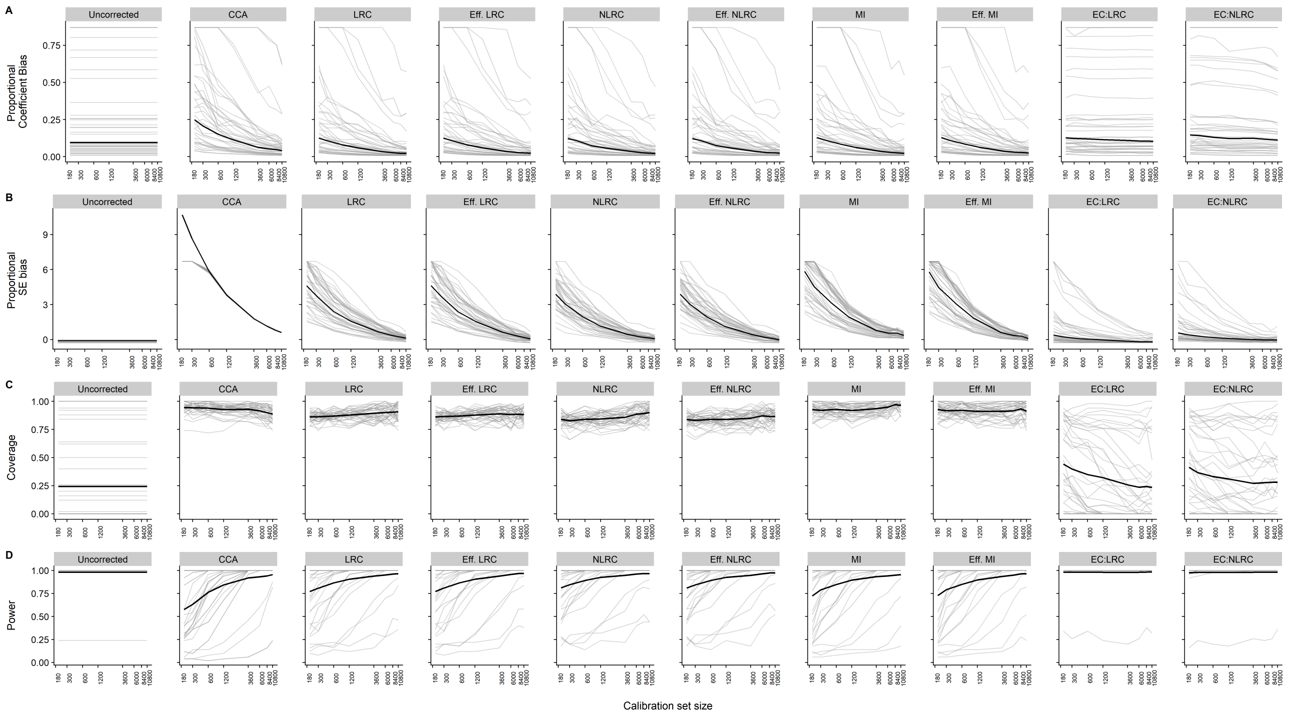


Figure B.12: Effects of calibration sample size on bias, coverage, and power across all alternative error correction techniques (error-in-Y case). Proportional coefficient bias, proportional standard error bias, coverage, and power across all models plotted against calibration sample size for each error correction technique. Results can be interpreted as in Figure 5, but include additional error correction approaches beyond multiple imputation. Error correction approaches are detailed in Section C.4.

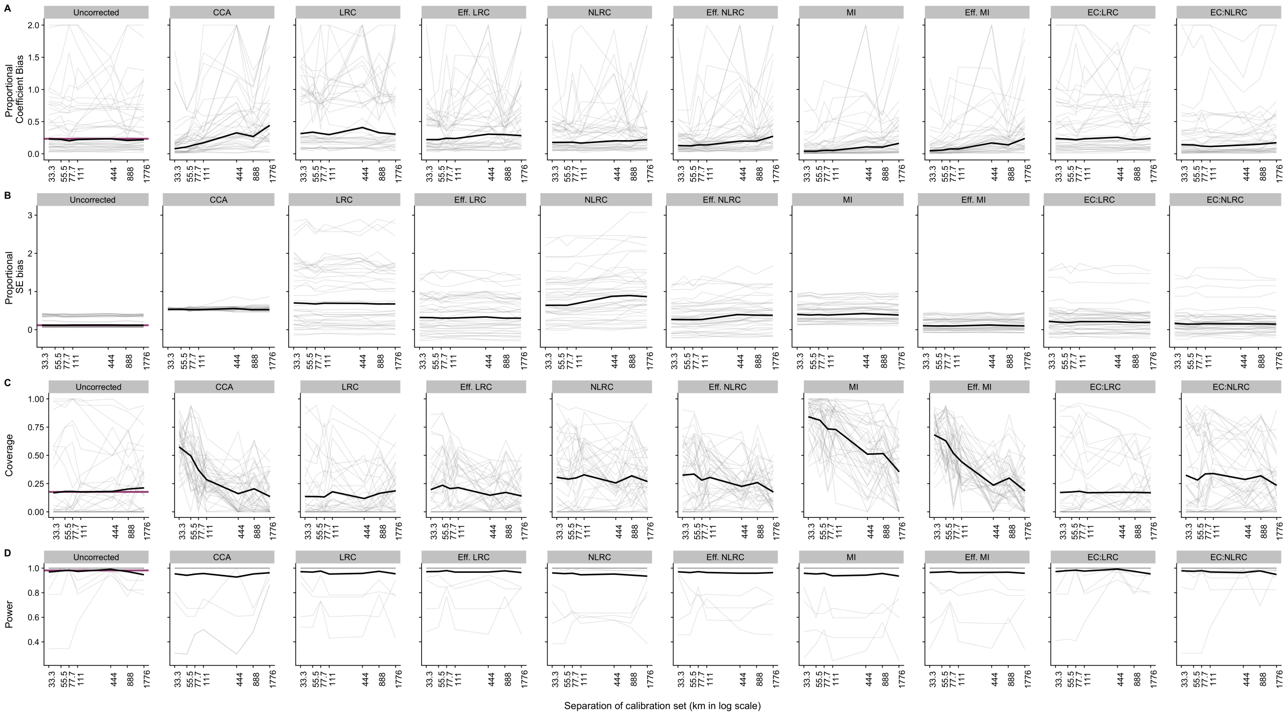


Figure B.13: The effect of distance between calibration and main datasets on the ability of all alternative error correction techniques to correct biases introduced by remotely sensed variables (error-in-X case). Proportional coefficient bias, proportional standard error bias, coverage, and power across all models plotted against distances between calibration sample and main sample for each error correction technique. Results can be interpreted as in Figure 6, but include additional error correction approaches beyond multiple imputation. Error correction approaches are detailed in Section C.4. As in the main text, horizontal purple lines show results for a random sampling of the calibration set (i.e., no spatial separation between calibration and main samples imposed) for the uncorrected model. For completeness we also show here average uncorrected values for each grid size. As expected, these values change very little as the grid size changes. Thus, for simplicity, we show the purple line only in the main text figures.

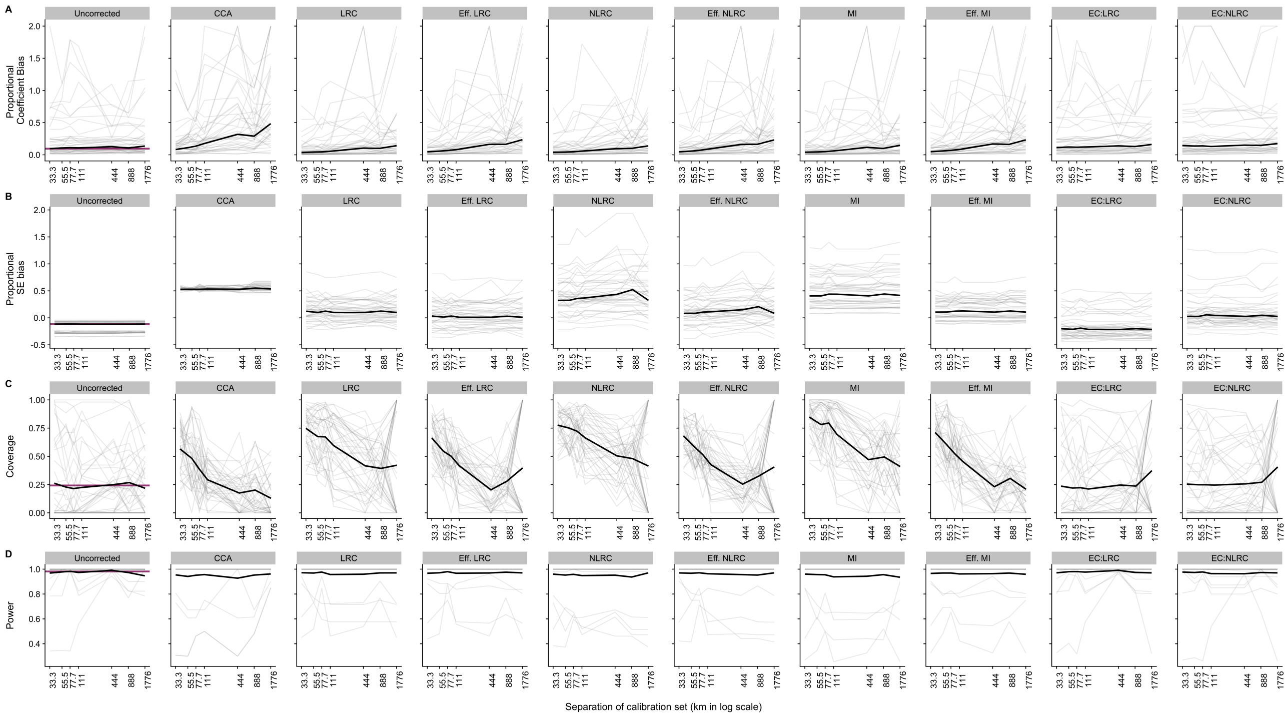


Figure B.14: The effect of distance between calibration and main datasets on the ability of all alternative error correction techniques to correct biases introduced by remotely sensed variables (error-in-Y case). Proportional coefficient bias, proportional standard error bias, coverage, and power across all models plotted against distances between calibration sample and main sample. Results can be interpreted as in Figure 6, but include additional error correction approaches beyond multiple imputation. Error correction approaches are detailed in Section C.4. As in the main text, horizontal purple lines show results for a random sampling of the calibration set (i.e., no spatial separation between calibration and main samples imposed) for the uncorrected model. For completeness we also show here average uncorrected values for each grid size. As expected, these values change very little as the grid size changes. Thus, for simplicity, we show the purple line only in the main text figures.

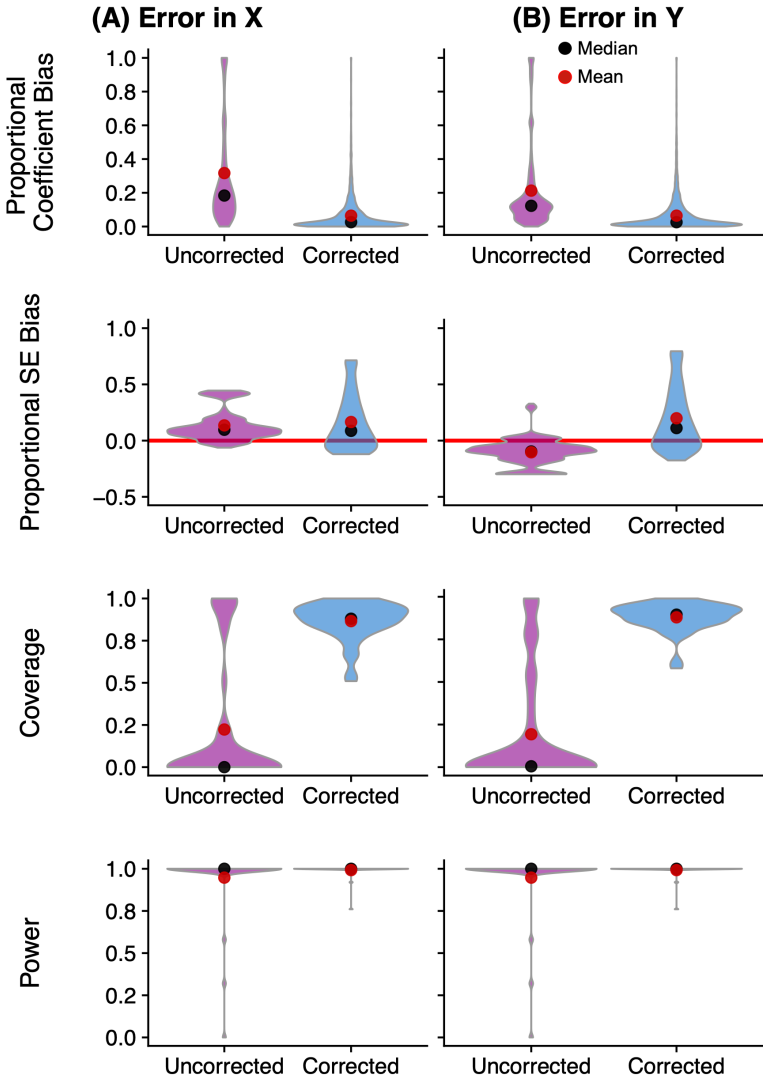


Figure B.15: Bias, coverage, and power for regression models using remotely sensed variables when the estimating equation includes spatial fixed effects. Figure shows analogous results to Figure 4, but here all regression models include state-level fixed effects (see text in Section 4.6 for details). Data for violin plots has been winsorized at the 2.5% and 97.5% to cap outliers for visual display purposes. The unwinsorized mean is indicated by the red circles and the corresponding median is indicated by the black circles.

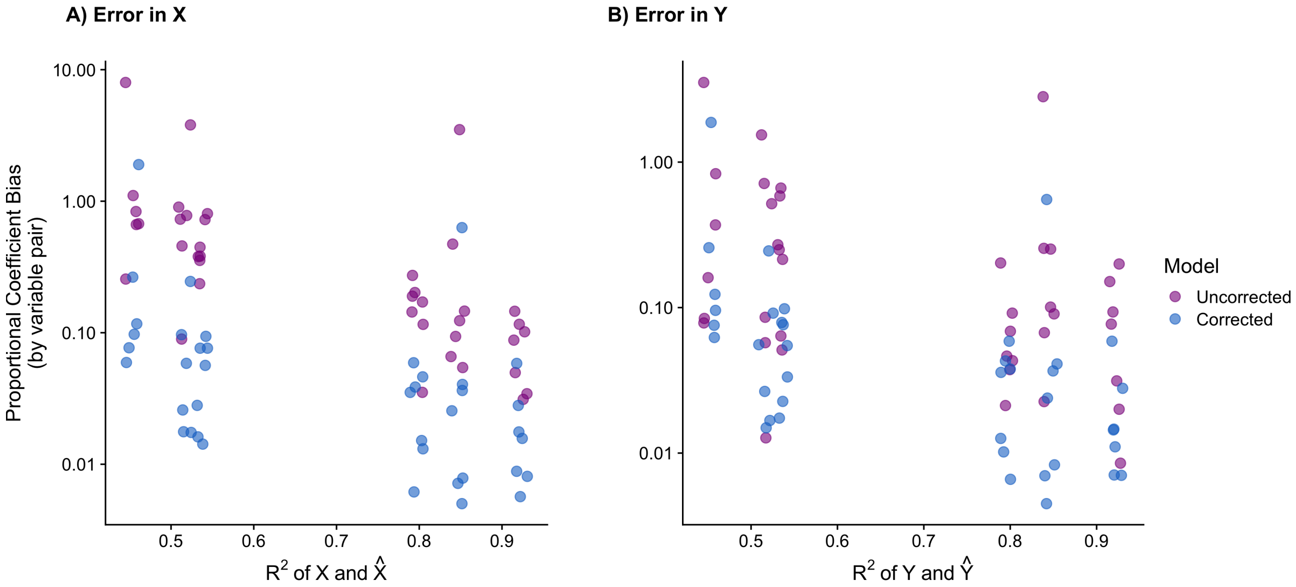


Figure B.16: Remotely sensed variables with higher accuracy exhibit lower coefficient bias in downstream regressions. Figure shows the proportional coefficient bias in uncorrected (purple) and corrected via multiple imputation (blue) downstream regression models plotted against the R^2 of the original remote sensing prediction. Multiple imputation consistently lowers coefficient bias for all regression models, but as R^2 decreases, the ability to correct for bias falls. Variable pairs with less than 0.5 R^2 suffer from large ($\geq$10%) bias in uncorrected models, and in some cases large biases remain even after correction. Points are jittered horizontally for visualization due to sets of variable pairs sharing the same R^2.

C Supplementary Discussions

C.1 Performance metrics

We use four performance metrics to compare uncorrected and corrected regression models to those using ground truth data. First, we compute bias in the regression coefficient:

$$\textit{bias in the regression coefficient} = \left| \frac{(\hat{\beta} - \hat{\beta}_{\tilde{z}})}{\hat{\beta}} \right|, \tag{S1}$$

where $\hat{\beta}$ is the regression coefficient estimated from Equation 1 using ground truth data, and $\hat{\beta}_{\tilde{z}}$ is the error-in-X or error-in-Y coefficient estimated *either* directly from Equation 2 (or the analogous error-in-Y regression) or from a version of Equation 2 that uses an error correction technique, such as multiple imputation, to correct for bias. We consider proportional bias to account for the different strengths of relationships between pairs of variables, and we consider absolute bias because the sign of β varies across variable pairs.

Second, we compute bias in standard error estimates:

$$\textit{bias in standard errors} = \frac{SE(\hat{\beta}_{\tilde{z}}) - SE(\hat{\beta})}{SE(\hat{\beta})}, \tag{S2}$$

where variables are defined as in Equation S1. Note that Equation S2 can be either positive or negative, reflecting overly conservative or overly precise standard errors in a model using remotely sensed data. Finally, we compute two statistics that combine the estimated coefficients and their uncertainty: *coverage* is calculated as the likelihood that a regression model using remotely sensed data recovers a 95% confidence interval containing the ground truth point estimate; and *power* is calculated as the likelihood that a regression model using remotely sensed data rejects a null hypothesis of no relationship between two variables when the ground truth regression also rejects this null (i.e., p<0.05).

C.2 Derivation of biases in the linear measurement error model

Here we derive the biases introduced by measurement error under the linear measurement error model, a general error model that encompasses both the familiar classical measurement error model as well as the Berkson error model as special cases (Keogh et al., 2020). We demonstrate how various restrictions imposed on this general measurement error model change the nature of bias, both in the error-in-X and error-in-Y cases.

As in the main text, we consider a simple linear regression framework in which the coefficient of interest is the slope parameter β (we suppress subscripts throughout this section for parsimony):

$$y = \alpha + \beta x + \varepsilon$$

The "true" value of β is that which would be recovered from a regression with no measurement error in either x or y.

C.2.1 Errors in independent variables (error-in-X)

Under the linear measurement error model, the error-prone variable (in our case, remotely sensed predictions) is written as an affine function of the accurately measured variable (in our case, ground truth observations), as follows:

$$\tilde{x} = \theta + \lambda x + u,$$

where $\tilde{x}$ represents the imperfect measurements of x and u represents residual, mean zero measurement errors. With this definition, we can write the expectation of the slope parameter estimated using remotely sensed $\tilde{x}$ in place of x as:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{\tilde{x}}] &= \frac{\sigma_{\tilde{x}y}}{\sigma_{\tilde{x}}} = \frac{\sigma_{(\theta+\lambda x+u,y)}}{\sigma_{\theta+\lambda x+u}} \\
&= \frac{\lambda\sigma_{xy} + \sigma_{yu}}{\sigma_{\theta+\lambda x+u}} \\
&= \frac{\lambda\sigma_{xy} + \sigma_{yu}}{\lambda^2\sigma_x + \sigma_u} && \text{plug in } \sigma_{xy} = \beta\sigma_x \\
&= \frac{\lambda\beta\sigma_x + \sigma_{yu}}{\lambda^2\sigma_x + \sigma_u} \\
&= \beta \underbrace{\frac{\lambda\sigma_x}{\lambda^2\sigma_x + \sigma_u}}_{\text{random and mean-reverting error}} + \underbrace{\frac{\sigma_{yu}}{\lambda^2\sigma_x + \sigma_u}}_{\text{differential error}} .
\end{aligned} \tag{S3}$$

Equation S3 shows that under the *general* linear measurement error model, there are three components contributing to bias for the error-in-X case. First, random error u causes attenuation of β through the "reliability ratio" $\frac{\sigma_x}{\sigma_x+\sigma_u}$ within the first term. Second, differential measurement error causes bias through a nonzero covariance σ_{yu} in the second term. Finally, the relationship between $\tilde{x}$ and x itself, captured through λ and displaying mean reversion when $\lambda < 1$, will introduce bias through both the first and second terms. The direction of bias under this error model is ambiguous and will depend on the relative magnitudes of λ, σ_u, and σ_{yu}.

In Figure B.6, we decompose Equation S3 into its component parts to assess which features drive observed biases. To do so, we derive bias under two less general forms of Equation S3. First, under the *non-differential* linear measurement error model the covariance σ_{yu} is assumed to be zero and $\mathbb{E}[\hat{\beta}_{\tilde{x}}]$ simplifies to:

$$\mathbb{E}[\hat{\beta}_{\tilde{x}}] = \beta \frac{\lambda \sigma_x}{\lambda^2 \sigma_x + \sigma_u}. \tag{S4}$$

Note that mean reversion with $\lambda < 1$ biases β upward, while the reliability ratio causes familiar attenuation bias. Whether $\hat{\beta}_{\tilde{x}}$ is inflated or attenuated under this error model depends on which of these two forces dominates. If measurement error is Berkson, λ and the reliability ratio balance one another such that no bias arises in the error-in-X case (Carroll et al., 2006).

Second, under a *differential classical* measurement error model, $\lambda = 1$ and $\mathbb{E}[\hat{\beta}_{\tilde{x}}]$ becomes:

$$\mathbb{E}[\hat{\beta}_{\tilde{x}}] = \beta \frac{\sigma_x}{\sigma_x + \sigma_u} + \frac{\sigma_{yu}}{\sigma_x + \sigma_u} \tag{S5}$$

While the reliability ratio attenuates β, σ_{yu} can be greater or less than zero, leading to theoretically ambiguous bias.

In Figure B.6, we plot observed proportional bias on the y-axis against theoretical bias following Equations S3 (top row), Equation S4 (middle row), and Equation S5 (bottom row). These findings show that the full flexibility of Equation S3 is needed to fully explain the patterns of bias we observe in remotely sensed variables.

C.2.2 Errors in dependent variables (error-in-Y)

Under the linear measurement error model for the error-in-Y case, we write:

$$\tilde{y} = \theta + \lambda y + v$$

where all terms are as defined above. With this definition, the expectation of the slope parameter estimated using $\tilde{y}$ instead of y is:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{\tilde{y}}] &= \frac{\sigma_{x\tilde{y}}}{\sigma_x} = \frac{\sigma_{(x,\theta+\lambda y+v)}}{\sigma_x} \\
&= \frac{\lambda \sigma_{xy} + \sigma_{xu}}{\sigma_x} \\
&= \lambda\beta + \frac{\sigma_{xu}}{\sigma_x} && \text{plug in } \sigma_{xy} = \beta\sigma_x
\end{aligned} \tag{S6}$$

Here, there are two components contributing to bias for the error-in-Y version of the general linear measurement error model. First, differential measurement error causes bias through a nonzero covariance σ_{xu} in the second term in Equation S6. Second, the relationship between $\tilde{y}$ and y itself, captured through λ, will introduce attenuation bias through the first term when $\lambda < 1$ under mean reversion. The direction of bias under this error model is ambiguous and will depend on the relative magnitudes of λ and σ_{xu}.

As above for the error-in-X case, we consider bias under two more restrictive error models in order to decompose overall bias into its component parts. First, under the *non-differential* linear measurement error model, the covariance σ_{xu} is assumed to be zero and $\mathbb{E}[\hat{\beta}_{\tilde{y}}]$ simplifies to:

$$\mathbb{E}[\hat{\beta}_{\tilde{y}}] = \lambda\beta, \tag{S7}$$

where it is clear that mean reverting measurement error $\lambda < 1$ will lead to attenuation of slope coefficients for the error-in-Y case.

Second, under a *differential classical* measurement error model, we assume $\lambda = 1$ and $\mathbb{E}[\hat{\beta}_{\tilde{y}}]$ becomes:

$$\mathbb{E}[\hat{\beta}_{\tilde{y}}] = \beta + \frac{\sigma_{xu}}{\sigma_x}, \tag{S8}$$

where bias is determined by the covariance σ_{xu}.

As discussed above, Figure B.6 plots observed proportional bias on the y-axis against theoretical bias, following Equations S6 (top row), Equation S7 (middle row), and Equation S8 (bottom row) for the error-in-Y case. These findings show that for error-in-Y, just like error-in-X, the flexibility of Equation S3 is needed to fully explain the patterns of bias we observe in remotely sensed variables.

C.3 Calibration sample separation experiments

To evaluate the ability of multiple imputation and other bias correction methods to improve parameter recovery when the calibration sample is spatially distant from the main sample, we design an experiment based on an experimental design in (Rolf et al., 2021) where we evaluate models using main and calibration samples that are increasingly far away from each other in space. Specifically, we create grids over the continental U.S. with side lengths of $\delta \in [0.2,\ 0.3,\ 0.4,\ 0.7,\ 1,\ 4,\ 8,\ 16]$ degrees latitude and longitude. We then use this grid to divide the main and calibration sample into spatially separate sets by randomly sampling the main and calibration sets from boxes that are not adjacent (width-wise and height-wise) within the grid, creating a checkerboard pattern as shown in Figure B.10. As δ increases, the calibration set becomes on average further away from the main sample. This separation makes it more difficult for each bias correction method to correct for bias, as observations in the calibration set are now less likely to be similar to those in the main sample. To minimize the noise from any specific placement of the grid, we move or "jitter" the grid by shifting it down, up, or left by half the width of the grid, before running the analysis for each bootstrap run. Average results are taken over these differently jittered bootstrap runs similarly to the rest of the analysis.

C.4 Alternative bias correction methods

In this section, we briefly describe the various bias correction methods that we evaluate in our setting, and their respective assumptions. For reference, comprehensive reviews of statistical methods of regression bias correction include Fuller (1995); Carroll et al. (2006); Freedman et al. (2008) and Keogh and White (2014). Throughout, we denote observations in the calibration and main dataset set with c and m subscripts. We describe all error correction methods for the error-in-X case; error-in-Y methods are analogous.

1. **Complete Case Analysis (CCA)**
   Complete case analysis estimates regression parameters directly in the calibration set, ignoring the main dataset and therefore any remotely sensed data entirely. This approach is analogous to "benefits transfer" in economics (Boutwell and Westra, 2013), as it simply applies a regression parameter estimated in one sample to a new context without any adjustment. To implement CCA, the regression coefficient of interest is simply estimated using only the ground truth values available in the calibration dataset. All data in the main sample, and therefore all remotely sensed predictions, are dropped. For CCA to provide an unbiased estimate of the true parameter of interest in the main sample, the calibration dataset must have the same relationship between y and x as the main sample.

2. **Single imputation linear regression calibration (LRC)**
   Regression calibration has been one of the most commonly-used methods in the measurement error model literature (Freedman et al., 2008) because of its simplicity. To implement linear regression calibration, the relationship between a ground truth variable, its remotely sensed counterpart, and the ground truth regressor (error-in-Y case) or regressand (error-in-X case) is estimated in the calibration sample and then used to make predictions of the true variable in the main sample. In a second stage, the *predicted* ground truth values are then used to estimate the regression model in the main sample. This is the simplest form of imputation, but has been shown to perform poorly when measurement error is non-classical (Cole, Chu and Greenland, 2006). It also does not carry uncertainty from the calibration step into the final estimates, which can lead to overly-precise parameter estimates.

   The method is implemented as follows:

   (a) Estimate a linear model of $x_c = \delta_{RC} + \gamma_{RC}\tilde{x}_c + \phi_{RC}y_c + e_{RC}$ in the calibration sample.

   (b) Predict $\widehat{x_{RC}}$ using $\widehat{\delta_{RC}}$, $\widehat{\gamma_{RC}}$, $\widehat{\phi_{RC}}$ as well as observations of $\tilde{x}_m$ and y_m in the main sample.

   (c) Estimate the linear regression in the main sample treating $\widehat{x_{RC}}$ as you would if it were not measured with error: $y_m = \alpha + \beta_{RC}\widehat{x_{RC}} + \epsilon$

   (d) Obtain $\widehat{\beta_{RC}}$ as the linear regression calibration corrected parameter.

3. **Single imputation efficient linear regression calibration (Eff. LRC)**
   This method is implemented identically to linear regression calibration above, but in step (c), the calibration set is appended to the main sample before estimation. This yields an estimate of $\widehat{\beta_{ERC}}$, which is an inverse-variance-weighted average of the estimate of the coefficient in the calibration set and the estimate of $\widehat{\beta_{RC}}$ using linear regression calibration (Freedman et al., 2008; Keogh and White, 2014).

4. **Single imputation nonlinear regression calibration (NLRC)**
   Nonlinear regression calibration is an extension of single imputation linear regression calibration wherein the first stage model is a nonlinear and flexible model. We use a random forest in our implementation, but other nonlinear methods are possible. The method is implemented as follows:

   (a) Estimate a nonlinear model of $x_c = \boldsymbol{\delta_{RF}}\left(\tilde{x}_c, y_c\right)$ in the calibration sample.

   (b) Predict $\widehat{x_{RF}}$ using $\boldsymbol{\delta_{RF}}$ as well as observations of $\tilde{x}_m$ and y_m in the main sample.

   (c) Estimate the linear regression treating $\widehat{x_{RF}}$ as you would if it were not measured with error: $y_m = \alpha + \beta_{RF}\widehat{x_{RF}} + \epsilon$

   (d) Obtain $\widehat{\beta_{RF}}$ as the nonlinear regression calibration corrected parameter.

5. **Single imputation efficient nonlinear regression calibration (Eff. NLRC)**
   This method is the same as above for single imputation nonlinear regression calibration, but in step (c), the calibration set is appended to the main sample before estimation.

6. **Single imputation external linear regression calibration (EC: LRC)**
   In external calibration, the dependent variable is missing in the calibration set (for the error-in-X case). For the error-in-Y case, the independent variable is missing in the calibration set. The method is implemented similarly to linear regression calibration, but data from the dependent variable is omitted in step (a):

   (a) Estimate a linear model $x_c = \delta_{ExRC} + \gamma_{ECRC}\tilde{x}_c + e_{ECRC}$ in the calibration sample, noting that y_m is omitted here as compared to method 2 (LRC).

   (b) Predict $\widehat{x_{ECRC}}$ using $\widehat{\delta_{ECRC}}$, $\widehat{\gamma_{ECRC}}$ as well as observations of $\tilde{x}_m$ in the main sample.

(c) Estimate the linear regression treating $\widehat{x_{ECRC}}$ as you would if it were not measured with error: $y_m = \alpha + \beta_{ECRC}\widehat{x_{ECRC}} + \epsilon$

(d) Obtain $\widehat{\beta_{ECRC}}$ as the external linear regression calibration corrected parameter.

7. **Single imputation external nonlinear regression calibration (EC: NLRC)** This approach is the same as above, but uses a nonlinear function (random forest) to model the first stage (a).

8. **Multiple imputation: Bayesian linear model (MI)** Multiple imputation is described in detail in the main text. In general, this method replaces each "missing" or, in our case, mis-measured, value $\tilde{x}$ with a vector of $K > 1$ possible imputed values. There are many methods to do this imputation, and here we lay out the Bayesian linear model approach, which allows for parameter uncertainty in the first stage. We show in Figure B.7 that our main results do not depend on this particular approach to imputation. Under the Bayesian linear model imputation, the procedure (adapted by van Buuren (2012) from Rubin (1987)) is as follows:

   (a) Fit a linear regression model of the ground-truth x to the remotely-sensed variable $\tilde{x}$ and to y using the calibration sample. Obtain the estimated parameters $\hat{\delta}$, $\hat{\gamma}$, $\hat{\phi}$.

   (b) Calculate the covariance matrix $S = \widehat{\sum}_{\tilde{x},y}$.

   (c) Using the estimated parameters, $\hat{\delta}$, $\hat{\gamma}$, $\hat{\phi}$, and the covariance matrix, $\widehat{\sum}_{\tilde{x},y}$, draw an estimate of the set of three coefficients for each imputation $k \in \{1, 2, ..., K\}$ from the standard (three dimensional) multivariate Gaussian distribution.

   (d) Calculate $\hat{x}^k = \hat{\delta}^k + \hat{\gamma}^k\tilde{x}_m + \hat{\phi}^k y_m$, which is the imputed value for the k^{th} imputation.

   (e) Perform the linear regression analysis K times using the K values of $\hat{x}^k$. Obtain K estimates of $\hat{\beta}$. Pool these together using Rubin's Rule (see Rubin (1987)) to obtain one final multiple imputation estimate of the parameter, $\widehat{\beta_{MI}}$.

9. **Efficient multiple imputation: Bayesian linear model (Eff. MI)** This method is the same as above for multiple imputation, but in step (e), the calibration set is appended to the main sample before estimation.