Parameter recovery using remotely sensed variables

Jonathan Proctor[1] ⓡ, Tamma Carleton[2,3] ⓡSandy Sum[4]

[1]University of British Columbia
[2]University of California, Berkeley
[3]National Bureau of Economic Research
[4]University of California, Davis

Abstract

Remotely sensed measurements are increasingly used in empirical analyses. Errors in such measures may bias parameter estimation, but it remains unclear how large biases are or how to correct them. We use a large dataset of diverse remotely sensed variables and reanalyze four prior papers to establish stylized facts regarding parameter recovery with remotely sensed observations. We show that non-classical errors in these variables cause substantial coefficient bias and low coverage, particularly when used as regressors. We demonstrate that multiple imputation, a standard and easily implementable imputation technique developed for missing data problems, effectively reduces bias and improves coverage in cross-sectional and panel data research designs.

Dataverse data: https://doi.org/10.7910/DVN/XW0AUT
Github repository (code-only): https://github.com/sandysum/satellite-mi

Author order randomized and archived in the American Economic Association Random Author Order Archive. Proctor: jon.proctor@ubc.ca. Carleton: tcarleton@berkeley.edu. Sum: sandysum@ucdavis.edu. We thank Anastasios Angelopoulos, Patrick Baylis, Marshall Burke, Avi Feller, Robert Heilmayr, Solomon Hsiang, Peter Huybers, Kelsey Jack, Frederik Noack, Esther Rolf, Tijana Zrnic, and seminar participants at: the National Bureau of Economic Research Environment and Energy Economics Program Meeting; the UC Davis Economics and Agricultural and Resource Economics seminar; the Young Scholars Symposium on Natural Resource Governance at Indiana University; the Workshop for Environmental Economics and Data Science (TWEEDS); and the University of California Environmental Economics (UC-EE) online seminar for helpful comments and suggestions. We are grateful to Kathryn Baragwanath, Nilesh Shinde, Matthew Brooks, and Faraz Usmani for generously sharing their replication materials as well as to Nathan Ratledge, Gabriel Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke for helpful correspondence. We thank Youg Sanghvi and 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. Spurred by advances in computer vision, a cross-disciplinary research community has developed a range of algorithms that transform these 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; Christensen et al., 2025), monitor agricultural land use (Potapov et al., 2022) and measure income, wealth, and human development (Jean et al., 2016; Chi et al., 2022; Sherman et al., 2026) at fine resolution and national or even global scales.

These predictions provide a treasure trove of new data 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 are indirect measures of the true variables of interest and often exhibit substantial measurement error, which may introduce bias into both parameter estimates and associated measures of uncertainty when used in downstream regression analyses. 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 measurement error induced biases and methods for their correction have been long studied in the statistical and econometrics literatures (e.g., Little and Rubin, 2019; Bound et al., 2001), directly transferring error correction methods to the remote sensing setting is complicated by the complex nature of the errors, 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; Marx et al., 2019; Balboni et al., 2021) or independent (e.g., Kocornik-Mina et al., 2020; Proctor, 2021; Chen et al., 2022) variable in regression analysis. Despite substantial advances in recent work studying the use of remotely sensed variables in regression analysis (e.g., Alix-Garcia and Millimet, 2023; Ratledge et al., 2022; Sanford et al., 2025; Pelletier et al., 2025; Lu et al., 2025), 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 first quantify the extent to which remotely sensed variables intro-

duce 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 four empirical research papers and using a set of real data simulation experiments that leverage a benchmark dataset providing co-located ground truth data (also called "labels") and continuous remotely sensed predictions for multiple variables across the contiguous United States.[1] While simulations have been extensively used to demonstrate the efficacy of statistical error correction methods (e.g., Cole et al., 2006; Freedman et al., 2008; De Silva et al., 2017), 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 to evaluate what types of measurement error are typically present, which lead to biases, and to what degree these biases are amenable to correction.

We find that not accounting for measurement error, as is standard in most applied research, tends to substantially bias parameter point estimates and dramatically decrease coverage across diverse empirical settings. For example, in our simulation using real data, we find that 95% confidence intervals estimated using remotely sensed data rarely contain the target parameter of interest estimated using ground truth data. In our replication of existing research, coefficients estimated using remotely sensed values range from 50% too low to 80% too high, when compared to the parameters recovered using ground truth. Errors in remotely sensed variables tend to be non-classical: we demonstrate that while mean-reverting measurement error (negative correlations between errors in one variable and itself) is common, differential measurement error (correlations between errors in one variable and levels of another variable) is also widespread and can either offset or exacerbate the downward bias caused by mean-reverting measurement error in both simulations and in replications of published work.

Second, we present a method to correct for this bias that is feasible in cases where researchers have a small quantity of labeled data for calibration. Specifically, we show that multiple imputation, an "off-the-shelf" data imputation technique widely used in statistics to solve missing data challenges, but so far untested in this setting, lowers the bias in recovered parameter estimates and prevents exaggerated statistical precision across a broad set of empirical models. In our simulation, we show that 95% confidence intervals

[1]Throughout our analysis, we focus on continuous remotely sensed variables, as opposed to discrete classifications, due to their common use. Errors in discrete variables are inherently different from errors in continuous variables, as any measurement error will be non-random and negatively correlated with the true value. As a result, even small instances of misclassification in a discrete variable can lead to large biases in parameter estimates (Millimet, 2011). A recent paper extending our approach to errors in categorical data, allowing for nonparametric modeling of misclassification, demonstrates that it is also highly effective at reducing bias in the discrete setting (Wardle, 2025).

corrected using multiple imputation contain the target parameter of interest estimated using ground truth data over 90% of the time. In replications of prior empirical papers, we find that correction with multiple imputation dramatically affects the quantitative and qualitative findings of all four empirical studies, changing recovered coefficients by -49% to +78% across diverse datasets and econometric models. 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 data settings commonly used for program evaluation; such settings are largely unstudied in the error correction literature. Throughout, we compare the performance of multiple imputation to other common error correction methods that similarly leverage a calibration dataset, showing that it nearly always outperforms 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.

Our findings contribute to a nascent literature documenting measurement error in satellite-based datasets and exploring its implications for regression analysis (Jain, 2020; Fowlie et al., 2019). Some solutions to the problem have been proposed in the case of binary or categorical landcover data. Specifically, Alix-Garcia and Millimet (2023) incorporate institutional knowledge on specific sources of error in binary deforestation datasets into a modified maximum likelihood estimator. They show that this method is effective at mitigating bias in a program evaluation with a binary response variable in Mexico, with implications for other deforestation settings. Garcia and Heilmayr (2024) show that the binary and non-repeatable nature of remotely sensed deforestation data (pixels can get deforested only once) introduces bias into downstream econometric analyses, even in the absence of mismeasurement, but that simple solutions such as spatial aggregation can be highly effective. Torchiana et al. (2023) point out that errors in remotely sensed time series of image classifications (e.g., landcover classes) can overestimate transitions between classes and propose a hidden Markov model solution that corrects transition probabilities under a set of assumptions about the structure of measurement error over time. Relative to these papers, our analysis is more general, as our correction method can be applied over time and space, to any continuous remotely sensed outcome variable, when error occurs in the dependent or independent variable, and without strong assumptions or knowledge of the measurement error structure. However, our solution has more stringent data requirements – multiple imputation is feasible only when researchers have access to some ground truth data for calibration – and requires modification when applied to discrete classification problems (Wardle, 2025).

Other solutions have emerged that rely on adjustments to the upstream remote sensing

(e.g., Rambachan et al., 2024; Ratledge et al., 2022; Sanford et al., 2025). Ratledge et al. (2022) correct for mean-reverting measurement error in satellite-based wealth predictions by tailoring the loss function during algorithm development. While this method effectively reduces bias from mean-reverting measurement error, it does not address other error types, such as differential measurement error, which drives the majority of the bias we find in our experiments. Conversely, Sanford et al. (2025) explicitly focus on differential measurement error, developing a debiasing technique to adjust for it when training a predictive remote sensing model of forest cover. Both approaches, as well as the related Rambachan et al. (2024) methodology, address measurement error only in the dependent variable. Neither directly addresses bias in standard errors. Importantly, these methods are infeasible to implement for most researchers, who use, but do not themselves produce, remotely sensed predictions. In contrast, our analysis focuses on methods that account for errors in remotely sensed data after it has been produced.

Finally, our work complements prior studies addressing downstream biases caused by errors in machine learning predictions more broadly.[2] Wang et al. (2020) leverage statistical methods similar to multiple imputation to adjust regressions with machine learning generated outcome variables. However, their proposed method applies only to error-prone dependent variables and requires causal relationships of interest to include variables used in the upstream prediction algorithm. Angelopoulos et al. (2023) develop a "prediction-powered inference" approach that applies beyond regression analysis, but apply it only to dependent variables in linear regression. Neither of these analyses focus on remote sensing, address error-prone independent variables, or assess bias correction efficacy under the practical constraints – such as limited or spatially clustered ground truth data – that most applied researchers face. PPI has recently been adapted to regressions with remotely sensed variables (Pelletier et al., 2025; Lu et al., 2025), but its performance across diverse settings, with limited ground truth data, and in a variety of panel data regression frameworks has not yet been established.

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 paper represents an extensive set of experiments aimed at informing the use and correction of remotely sensed measurements in regression analysis. Moreover, our proposed correction method is general, simple, easy to implement in multiple software programs,[3] and can be used by

[2] We note that our setting is quite distinct from the double/debiased machine learning methods presented in Chernozhukov et al. (2018), where the focus is on using machine learning directly to estimate the causal relationships of interest (as opposed to generating the data inputs for a standard statistical regression framework).

[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.

researchers who do not produce satellite-based data themselves. 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 predictions in downstream regressions.

2 Conceptual framework

There are several mechanisms through which error-prone remotely sensed measurements can bias downstream parameter recovery. Here, we use a standard measurement error model in a simple linear regression framework to elucidate these mechanisms and provide general expressions for recovered biases. We then present multiple imputation as an approach for mitigating such biases. In our empirical analysis, we quantify bias and apply multiple imputation in more complex settings. Supplementary Materials Section A.1 provides derivations of the expressions presented in this section.

2.1 Bias in recovered parameters

We begin by considering a straightforward setting in which the true population relationship between variables y and x is:

$$y = \alpha + \beta x + \varepsilon. \tag{1}$$

A researcher, however, has access only to remotely sensed $\tilde{y}$ (which we call the error-in-Y estimation) or remotely sensed $\tilde{x}$ (which we call the error-in-X estimation), leading her to estimate one of the following regressions:

$$\begin{aligned} \tilde{y} &= \alpha_{\tilde{y}} + \beta_{\tilde{y}} x + \epsilon_{\tilde{y}} \quad \text{(error-in-}Y\text{)} \\ y &= \alpha_{\tilde{x}} + \beta_{\tilde{x}} \tilde{x} + \epsilon_{\tilde{x}} \quad \text{(error-in-}X\text{).} \end{aligned} \tag{2}$$

Errors in $\tilde{y}$ and $\tilde{x}$ have the potential to bias recovered parameters of interest $\hat{\beta}_{\tilde{y}}$ and $\hat{\beta}_{\tilde{x}}$ away from the true population parameter β. To characterize the resulting biases, we assume here that errors in remotely sensed variables follow the general linear measurement error model, which is used widely in both theoretical (Keogh et al., 2020) and empirical (Ratledge et al., 2022) work to assess measurement error bias. We specify this error model in order to derive analytic expressions for parameter biases. However, we note that this assumed structure is not necessary for our proposed solution, multiple imputation, to successfully address parameter biases. Multiple imputation can address a broader class of error structures, and thus its applicability is not limited to this setting.

Under this model, the error-prone remotely sensed variable is assumed to be an affine function of the true variable:

$$\begin{aligned} \tilde{y} &= \theta_y + \lambda_y y + u_y \quad \text{(error-in-}Y\text{)} \\ \tilde{x} &= \theta_x + \lambda_x x + u_x \quad \text{(error-in-}X\text{)}, \end{aligned} \tag{3}$$

where u_x and u_y are assumed to be mean zero and uncorrelated with x and y, respectively (that is, we assume that $cov(y, u_y) = cov(x, u_x) = 0$). θ allows for a level shift between the remotely sensed and true value, λ allows the remotely sensed value to be a scaling of the true value, and u allows for random error.[4]

This error model is quite general. For example, classical measurement error follows Equation 3 with the additional assumptions that $\theta_z = 0$ and $\lambda_z = 1$ for $z \in \{x, y\}$, and that $cov(y, u_x) = cov(x, u_y) = 0$. Another common case of the linear measurement error model emerges when a mismeasured variable is generated from a prediction or calibration equation, which shrinks variance in error-prone variables relative to the truth and leads to $0 < \lambda_z < 1$ (this is also called mean-reverting measurement error).

Under the most general form of the linear measurement error model, error-in-Y and error-in-X regression models recover the following slope coefficients in expectation:

$$\begin{aligned} \mathbb{E}[\hat{\beta}_{\tilde{y}}] &= \lambda_y \beta + \frac{\sigma_{xu_y}}{\sigma_x^2} \quad \text{(error-in-}Y\text{)} \\ \mathbb{E}[\hat{\beta}_{\tilde{x}}] &= \beta \frac{\lambda_x \sigma_x^2}{\lambda_x^2 \sigma_x^2 + \sigma_{u_x}^2} + \frac{\sigma_{yu_x}}{\lambda_x^2 \sigma_x^2 + \sigma_{u_x}^2} \quad \text{(error-in-}X\text{)} \end{aligned} \tag{4}$$

where σ_x^2 is the variance in the true variable x and $\sigma_{u_x}^2$ is the variance in the residuals from the error-in-X error model in Equation 3. Covariances between errors in one variable and values of the other are indicated by σ_{xu_y} and σ_{yu_x}; when these are non-zero, the error is called "differential" (Carroll et al., 2006).

Equation 4 recovers the standard prediction that classical measurement error ($\lambda_z = 1$ and zero covariance terms) causes no bias in error-in-Y models but attenuates coefficients in error-in-X models by a magnitude determined by the "reliability ratio" $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_x}^2}$. In practice, however, biases are additionally influenced by λ_z and differential measurement error, indicated by the covariance terms in both rows of Equation 4.

Mean-reverting and differential measurement errors are not uncommon. For example, mean-reverting measurement error occurs when predictions have lower variance than true

[4] Of course, in practice, the relationship between the remotely sensed and true value may take other forms, though we find in Section 4 that the linear measurement error model is a good approximation in the data sample used for our real data simulation experiment. We note that with unrestricted values of θ and λ, mean zero error terms in Equation 3 are trivial. Moreover, as long as Equation 3 is not misspecified, error terms are uncorrelated with regressors and $cov(y, u_y) = cov(x, u_x) = 0$.

observations and are used as the outcome variable in a program evaluation regression, as illustrated in Ratledge et al. (2022) for predicted wealth and Pelletier et al. (2025) for predicted crop yields. Similarly, differential measurement error occurs when errors in a remotely sensed outcome variable are correlated with a treatment variable. To see this, consider a regression of remotely sensed forest cover ($\tilde{y}$) on road infrastructure (x). Suppose the training sample had few examples of rural roads in densely forested areas and therefore predicted low forest cover near all roadways, inducing a negative correlation between road infrastructure and forest cover prediction errors (i.e., $\sigma_{xu_y} < 0$). This differential measurement error would cause downward bias in the estimated coefficient. Another mechanism through which differential measurement error could occur is if the remote sensing algorithm for y uses x as a proxy. For example, consider a randomized control trial (RCT) quantifying the impact of a large asset (e.g., livestock) on earnings. Suppose that earnings are remotely sensed by a model that has learned that areas with more livestock grazing tend to have higher earnings, either because livestock contribute to income generation (the effect the RCT seeks to estimate) or because people with higher earnings tend to purchase more livestock (a confounder). In the latter case, earnings would be over-predicted for people randomly treated with a livestock asset, inducing a positive correlation between errors in predicted income and treatment ($\sigma_{xu_y} > 0$) that will bias the estimated treatment effect upward.

Under the general linear measurement error model, 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.[5] In Supplementary Materials Section A.1, we show how different assumptions regarding λ_x, λ_y, σ_{xu_y} and σ_{yu_x} influence the expected direction and magnitude of recovered biases. While this exposition relies on a simple linear regression model, the same biases can emerge in more common applied research designs with rich controls and/or fixed effects. In such settings, the same expressions in Equation 4 apply once y and x have been residualized with respect to the controls (Lovell, 2008). In fact, such controls can, in some cases, exacerbate biases, as small errors in non-residualized variables can become relatively large in residualized data (Wooldridge, 2010). We examine such settings empirically in Section 4.5.

Of course, measurement error can influence recovered standard errors as well as point estimates. Estimation of the regression models in Equation 2 will recover the following

[5]For example, with $0 < \lambda_y < 1$ and no differential measurement error, error-in-Y models will exhibit attenuated coefficients $\beta_{\tilde{y}} = \lambda_y \beta < \beta$, while error-in-$X$ models can be biased in either direction, depending on the relative magnitudes of λ_x and the reliability ratio. With differential measurement error as in the general form in Equation 4, biases can arise in either direction for both model types.

variance estimates:[6]

$$\text{var}(\hat{\beta}_{\tilde{y}}) = \frac{\lambda_y^2 \sigma_\varepsilon^2}{N\sigma_x^2} + \frac{1}{N\sigma_x^2}\left[\sigma_{u_y}^2 - \frac{\sigma_{xu_y}^2}{\sigma_x^2}\right] \qquad \text{(error-in-}Y\text{)}$$

$$\text{var}(\hat{\beta}_{\tilde{x}}) = \frac{\sigma_\varepsilon^2}{N(\lambda_x^2\sigma_x^2 + \sigma_{u_x}^2)} + \frac{\beta^2\sigma_{u_x}^2\sigma_x^2 - 2\beta\lambda_x\sigma_x^2\sigma_{yu_x} - \sigma_{yu_x}^2}{N(\lambda_x^2\sigma_x^2 + \sigma_{u_x}^2)^2} \qquad \text{(error-in-}X\text{)} \qquad (5)$$

where the variance of the estimated slope coefficient under the standard assumptions of the Classical Linear Regression Model (CLRM), $\frac{\sigma_\varepsilon^2}{N\sigma_x^2}$, is indicated in brown. Equation 5 shows that for the error-in-Y case, mean-reverting measurement error, where $0 < \lambda_y < 1$, shrinks estimates of parameter uncertainty due to failures of the CLRM assumptions. However, in combination, random error σ_{u_y} and differential measurement error $\sigma_{xu_y}^2$ can either exaggerate or attenuate this effect, depending on their magnitudes. If there is only classical measurement error ($\lambda_y = 1$ and $\sigma_{xu_y} = 0$), the estimated variance collapses to $\frac{\sigma_\varepsilon^2 + \sigma_{u_y}^2}{N\sigma_x^2}$ and we recover the canonical result that classical error in Y inflates variance. In the error-in-X case, biases can go in either direction. When there is no differential measurement error, a value of λ_x between 0 and 1 will inflate both terms, all else equal, increasing recovered standard errors. However, in the presence of classical and/or differential measurement error, the bias cannot be signed. Even in the classical measurement error case ($\lambda_x = 1$ and $\sigma_{yu_x} = 0$), estimated variance simplifies to $\frac{\sigma_\varepsilon^2}{N(\sigma_x^2+\sigma_{u_x}^2)} + \frac{\beta^2\sigma_{u_x}^2\sigma_x^2}{N(\sigma_x^2+\sigma_{u_x}^2)^2}$, which can either exceed or fall below variance recovered when assumptions of CLRM are upheld.

2.2 Multiple imputation

To address biases introduced by remotely sensed variables, we follow prior statistical literature by recasting the problem of measurement error as a problem of missing data (Keogh and Bartlett, 2021), for which simple and general solutions have been previously developed. Specifically, we treat remote sensing applications as settings in which ground truth data are available in some locations, but missing in the full sample desired for analysis, where only an error-prone proxy is available. A well-established statistics literature tackles such missing data problems, originating with the study of systematic survey nonresponse (Rubin, 1987; Little and Rubin, 2019). Specifically, "multiple imputation" – a flexible and common statistical error correction method – has been used across a diversity of measurement error settings to adjust regression analysis to mitigate downstream biases (e.g., Freedman et al. (2008); Liu and De (2015); De Silva et al. (2017); Keogh et al.

[6]Note that this derivation assumes homoscedasticity and no cross-sectional dependence of the errors from Equations 1 and 2 (e.g., $\text{var}(\varepsilon) = \sigma_\varepsilon^2 I_n$). Analogous expressions can be derived under more general error structures.

(2020)). In the following sections, we introduce the typical implementation of multiple imputation, and discuss prior theory and evidence indicating the conditions under which multiple imputation is likely to perform well.

2.2.1 Implementation

Multiple imputation can be used in a remote sensing context when a researcher's main estimation sample contains remotely sensed variables, but she additionally has access to some smaller quantity of ground truth data, potentially from another location or time period, that she can match to remotely sensed observations. Figure 1 illustrates the data requirements for this secondary dataset, which is called the "calibration" sample, for both error-in-X and error-in-Y settings. Multiple imputation allows the researcher to estimate the relationship between the ground truth and the remotely sensed variables using the calibration dataset, and then to use that estimated relationship to impute ground truth data in the main sample, from which regression parameters and measures of uncertainty can then be recovered.

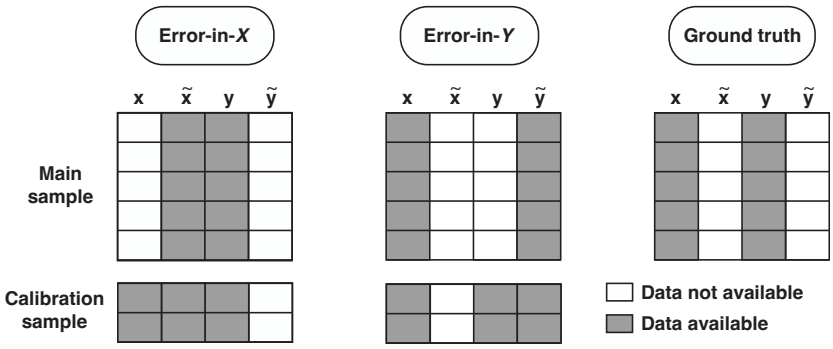

Figure 1: Three data availability regimes with different implications for parameter recovery and bias correction. Figure shows three data availability scenarios evaluated in this analysis. In the error-in-X case (left), the main analysis sample includes ground truth data for the dependent variable y, but only remotely sensed measurements for the independent variable x (denoted $\tilde{x}$). The calibration sample, which is generally smaller than the main sample, additionally includes ground truth observations of x. In contrast, the error-in-Y case (middle) 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, ground truth observations are available in the entire main sample.

The steps for implementing multiple imputation in the error-in-X case using linear models are as follows. The error-in-Y implementation is analogous.

1. **The imputation step:** First, generate K imputations of the "missing" ground truth observations in the main estimating sample. This is done through three stages.[7]

[7]There are many methods with which one could conduct the imputation step; here, we describe

(a) In the calibration sample, fit an imputation model by regressing ground truth x on the remotely sensed variable $\tilde{x}$ and y:

$$x = \delta + \gamma\tilde{x} + \psi y + e. \tag{6}$$

(b) Take K draws of $\hat{\delta}$, $\hat{\gamma}$, $\hat{\phi}$ and $\hat{\sigma}^2$ from their estimated posterior distributions, where $\hat{\sigma}^2$ is the estimated variance of e. This gives $\hat{\delta}^k$, $\hat{\gamma}^k$, $\hat{\phi}^k$ and $\hat{\sigma}^{2,k}$ for $k = 1, \ldots, K$.

(c) In the main sample, use these draws to predict the missing ground truth values of the remotely sensed variable K times: $\hat{x}^k = \hat{\delta}^k + \hat{\gamma}^k\tilde{x} + \hat{\psi}^k y + \hat{e}^k$, with $\hat{e}^k$ drawn from a normal distribution with mean 0 and variance $\hat{\sigma}^{2,k}$. Note that this step generates K versions of the main estimating sample, one for each imputation k.

2. **The estimation step:** Second, estimate the parameters of interest. In the main sample, estimate the model of interest K times using the K sets of imputed values as if they were ground truth. For example, a simple linear regression would be:

$$y = \alpha^k + \beta^k\hat{x}^k + u^k. \tag{7}$$

Store $\hat{\beta}^k$ and $\widehat{V}(\hat{\beta}^k)$ for each imputation k.

3. **The combining step:** Finally, obtain the multiple imputation estimate of the parameter of interest, denoted $\hat{\beta}_{MI}$, and its variance, $\widehat{V}(\hat{\beta}^{MI})$. To do so, pool the parameters estimated from K multiply imputed data sets using "Rubin's rules" (Rubin, 1987):

$$\begin{aligned}\hat{\beta}^{MI} &= \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}^k \\ \widehat{V}(\hat{\beta}^{MI}) &= \underbrace{\frac{1}{K}\sum_{k=1}^{K}\widehat{V}(\hat{\beta}^k)}_{\substack{\text{within-imputation}\\ \text{variance}}} + \underbrace{\frac{\sum_{k=1}^{K}(\hat{\beta}^k - \hat{\beta}^{MI})}{K-1}}_{\substack{\text{between-imputation}\\ \text{variance}}} + \underbrace{\frac{\frac{\sum_{k=1}^{K}(\hat{\beta}^k - \hat{\beta}^{MI})}{K-1}}{K}}_{\substack{\text{adjustment for}\\ \text{finite } K}}.\end{aligned} \tag{8}$$

Bayesian imputation under the normal linear model, following van Buuren (2012) and Rubin (1987). This algorithm is implemented in the `mice` package in `R`. See the function documentation for `mice` and, in particular, `mice.impute.norm`, for details. The algorithm, specifically the imputation step, is described in greater detail in Supplementary Materials Section A.6. An alternative, though very similar, implementation of multiple imputation (`mice.impute.norm.boot`) uses a bootstrapping approach for imputation. We show in Figure B.3 that our main results are consistent when using these and other implementations of the imputation step. We use the Bayesian normal linear model approach throughout the experiments due to its computational efficiency.

For an illustrative example of how this procedure works to correct biases induced by remotely sensed measures, see Figure A.5.

A few key features of multiple imputation are worth noting. First, multiple imputation removes biases by correcting structured errors in the remotely sensed measures. For example, when extreme values of x are routinely underestimated, as is common in remotely sensed estimates, $\hat{\gamma}$ in Equation 6 is greater than one, and predictions $\hat{x}^k$ resulting from the imputation step no longer reflect this underestimation. Second, during the estimation step, the calibration sample is commonly appended to the main sample – an approach known as "efficient" multiple imputation because it takes advantage of all available data. In contrast, "standard" multiple imputation uses just the main sample. As we discuss below, standard multiple imputation is appropriate when the true population parameter β is likely to differ between calibration and main samples. Third, a key feature of multiple imputation is that the final parameter estimates and standard errors account for uncertainty in the imputation step. In contrast, simpler methods of imputation such as linear or nonlinear regression calibration, which we analyze empirically below, use just a single imputation regression and result in variance underestimation and bias in finite samples (van Buuren, 2012; Freedman et al., 2008). Fourth, while we depict linear models for both the imputation and estimation steps, more flexible models and additional covariates or fixed effects can easily be used, including within canned software packages (van Buuren and Groothuis-Oudshoorn, 2011). This flexibility allows multiple imputation to complement existing analysis pipelines without users having to modify their main sample estimator; analysts can simply use the imputed values as if they were ground truth. We detail and implement multiple imputation in various such settings – including triple difference panel data estimators and matrix completion approaches – later in the paper. Finally, readers may note that multiple imputation resembles some instrumental variables (IV) approaches used to address measurement error. In Supplementary Materials Section A.3 we discuss both the similarities and differences between multiple imputation and IV, including split-sample implementations.

2.2.2 When does multiple imputation work? Prior theory and evidence

Multiple imputation, like other related statistical error correction methods (e.g., Keogh et al., 2020; Angelopoulos et al., 2023; Wang et al., 2020), relies on extrapolating estimated relationships from a calibration sample to impute data in the main sample of interest. The value of leveraging a calibration sample in this manner, of course, depends on whether the conditional distributions of the true values of the missing data align between the calibration and the main samples. A key benefit of multiple imputation is its demonstrated success – theoretically, in simulation studies, and in empirical applications

– at estimating parameters of interest given varied error structures and calibration data settings (e.g., Cole et al., 2006; Freedman et al., 2008; van Buuren, 2012; Little and Rubin, 2019; Keogh and White, 2014). We do not review the large imputation literature here and we leave formal proofs of multiple imputation in both Bayesian (Rubin, 1987) and frequentist (Meng, 1994) frameworks to prior literature. Instead, we briefly summarize the conditions under which multiple imputation will recover valid inferences and provide practical guidance for users.

Multiple imputation delivers consistent parameter estimates ($\hat{\beta}^{MI}$) and valid measures of uncertainty ($\widehat{V}(\hat{\beta})^{MI}$) under three core conditions. First, the regression model itself must be well-specified. That is, the ground truth point estimate $\hat{\beta}$ in the main sample must be a consistent estimator of the true population parameter β. Second, ground truth data must be "missing at random" (MAR) from the main sample, conditional on the observed data. For ground truth values to be MAR, missingness can depend on observed data but must not depend on the unobservable ground truth value once the observed data are accounted for. Third, the imputation model must be statistically compatible with the main sample model of interest, a property known as "congeniality" (Meng, 1994). Intuitively, this condition means that the imputation model should not impose assumptions that contradict or are more restrictive than those applied in the main sample estimating equation. In practice, when both the imputation and main sample models are linear regressions, a simple way to avoid issues with uncongeniality is to make the imputation model at least as rich as the analysis model (Xie and Meng, 2017; Carpenter et al., 2023). For example, while the imputation model can include additional controls or interactions that are not used in the main sample model, it must include the controls, interactions, or functional forms that enter the main estimating equation of interest. Congeniality is detailed in Carpenter et al. (2023, Ch. 2) and formalized in Meng (1994). Although these three conditions are sufficient for multiple imputation to deliver consistent parameter estimates and valid uncertainty measures, we note that multiple imputation has also been found to perform remarkably well even when these conditions are not fully satisfied (Carpenter et al., 2023).

In practice, the MAR assumption cannot be directly tested empirically because one would need the missing data to do so. Instead, analogous to causal identification assumptions in empirical studies, researchers must indirectly evaluate whether the MAR assumption is defensible in their setting of interest. In remote sensing applications, the calibration sample is generally one of convenience, where ground truth data could be spatially and/or temporally disjoint from the main estimating sample. In such settings, multiple imputation recovers unbiased parameter estimates if controls can be used in the imputation model to align the conditional distribution of the missing data values in the

imputation and main samples. This is similar to the classic selection problem and solution proposed by Heckman (1979), as the estimated imputation regression should include all variables correlated with both the measurement error and the probability of inclusion in the calibration sample. In addition to including any relevant and observable covariates in the imputation regression step to help ensure unbiasedness with non-random calibration samples, the imputation step can also include auxiliary variables that increase its predictive power, even if they do not correlate with selection into the calibration sample. Such variables will not affect parameter bias but can improve efficiency (Carpenter et al., 2023).

3 Methods: real data simulation experiment

We conduct a series of empirical experiments to quantify both the biases induced by remotely sensed variables and the ability of multiple imputation and related methods to correct these biases.

Data Our real data simulation experiment relies primarily on a benchmark dataset from Rolf et al. (2021), which includes remotely sensed predictions and corresponding ground truth labels for six variables across the continental United States: forest cover, population density, nighttime luminosity, average household income, elevation, and road length.[8] 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) for details). Importantly, MOSAIKS generates predictions with similar error magnitude and structure to other common methods, such as a convolutional neural network (see Figure A.1 and Supplementary Figure 17 from Rolf et al. (2021)), making these data generally representative of many modern remotely sensed predictions.

Figure 2 shows these labels and remotely sensed predictions for all six variables at 1 km×1 km resolution across 80,000 sampled locations. Ground truth observations are collected from ∼2010-2015 (see Table B.1 for details) and satellite data is from 2018. Temporal misalignment between imagery and ground truth values may contribute to

[8]Note that the "ground truth" measures of forest cover and nighttime luminosity that we use are derived from imagery, though from different satellites using different wavelengths than the fine-resolution RGB imagery used to make the "remotely sensed predictions" of all six variables. This represents a setting, somewhat common in practice (e.g., Baragwanath and Shinde (2025); Brown et al. (2025)), where there exists both a high and low-quality satellite-based measure of the same variable. Simulation results are consistent when dropping variable pairs that include either forest cover or nighttime luminosity.

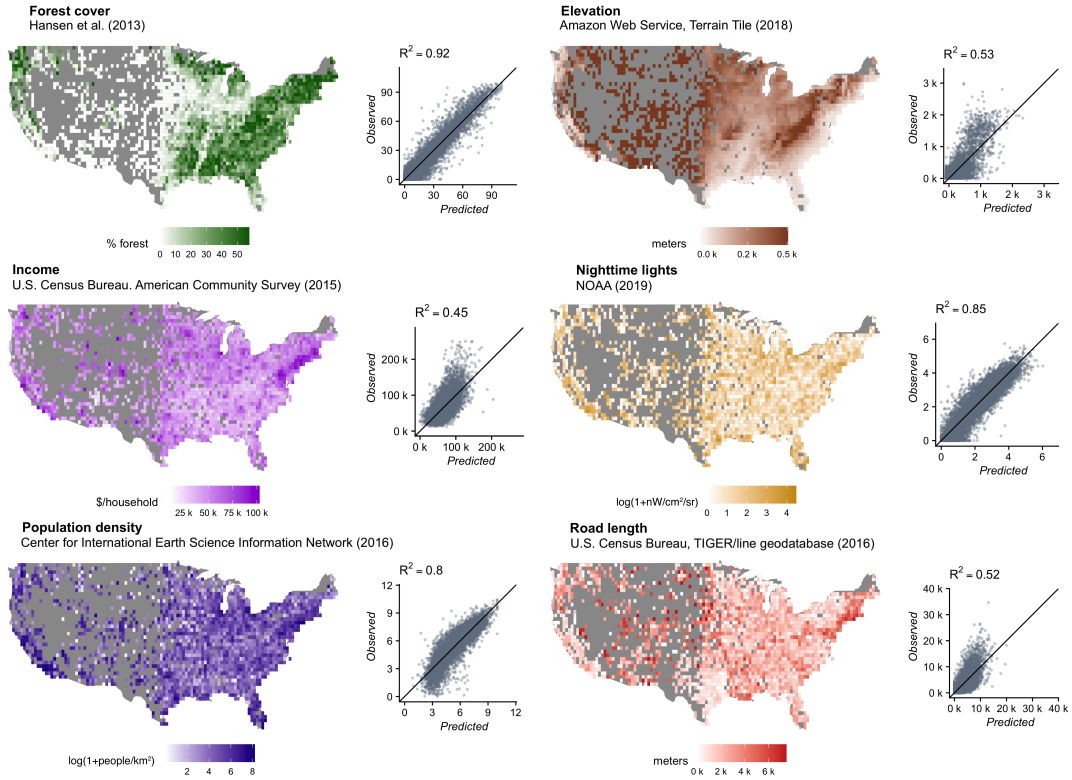

Figure 2: Ground truth labels and remotely sensed predictions of forest cover, elevation, income, nighttime lights, population density, and road length from Rolf et al. (2021). Maps show ground truth labels for 80,000 1 km×1 km grid cells which were sampled with a population-weighted uniform-at-random sampling scheme from across the continental United States and are aggregated to 20km×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.

measurement error in the remotely sensed estimates, though this is unlikely to be a dominant source of error. Of these 80,000 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.[9] We standardize all variables by the mean and standard deviation of the ground truth observations to facilitate comparisons across variables.

Methods To systematically evaluate bias induced by remotely sensed measurements across diverse empirical settings we use the co-located labeled data and predictions from Rolf et al. (2021) to create a set of 42 different ordered pairs of variables (e.g., population density and forest cover; nighttime luminosity and income, etc.).[10] Each of these pairs

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

[10]We use six variables from Rolf et al. (2021) and add two climate variables from PRISM, leading to eight total. Each variable can be paired with every other variable twice, once where it is the outcome

represents one potential relationship of interest. While it is impossible to create a fully representative sample of the many empirical settings where remotely sensed data could be used in regression analysis, these pairs provide a large and diverse set of regression models to evaluate the degree of bias introduced by measurement error.

We first consider a cross-sectional research design in which the unit of observation is one of 40,000 1 km×1 km grid cells across the U.S. We assume that the simple linear regression $y = \alpha + \beta x + \varepsilon$ is correctly specified, such that recovered coefficients $\hat{\alpha}$ and $\hat{\beta}$ and their standard errors represent unbiased estimates of the true parameters of interest when this regression is estimated using ground truth data. Since the "true" population parameters are unknown, we take parameters estimated using ground truth values as the target values in this experiment, and test whether parameters estimated using remotely sensed data match these target estimates. While the ground truth data may also be measured with error – a topic we address in Section 4.4 – we assume throughout that the ground truth data are well-measured and test the ability of remotely sensed data and multiple imputation to recover the ground truth relationship. For each pair of variables, we estimate three regressions, reflecting the three data availability regimes outlined in Figure 1: one where the ground truth labels are used for both variables, 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).

For each regression between each pair of variables, we calculate a distribution of recovered parameter estimates 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. We then partition each bootstrap sample 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 can be used for error correction. Performance metrics are calculated in the main sample. Coverage is calculated as the fraction of the 100 bootstrap runs in which the estimated confidence interval contains the ground truth point estimate.[11] 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 the absolute proportional bias in the regression coefficient (i.e., Equation S10) and proportional bias in standard errors (i.e., Equation S11) over all bootstrap-by-variable-pair combinations,

variable, and once where it is the independent variable. This gives $8 \times 7 = 56$ ordered pair combinations. We do not explore remotely sensed predictions of temperature or precipitation, which leaves 42 error-in-X and error-in-Y models with associated ground truth models.

[11] Traditionally, coverage is calculated using the population parameter of interest; we instead use the ground truth point estimate because the population parameter is unknown in this experiment. The same holds for power.

and the distribution of coverage and power over all variable pair combinations.[12] This experimental design is illustrated in Figure B.1.

4 Results: real data simulation experiment

4.1 Errors in remotely sensed measurements bias parameter estimates

The bias introduced by errors in remotely sensed measurements across all regressions between all pairs of variables is shown in Figure 3. 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.[13] 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.

Figure 3 shows that remotely sensed variables introduce substantial bias into linear regression coefficients; the median coefficient bias of the uncorrected estimates is 23% 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., nighttime 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 to high mean biases of 69% in the error-in-X case and 37% in the error-in-Y case. Figure B.2 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.

Figure 3, second row shows that errors in remotely sensed measurements also lead to biased estimates of parameter uncertainty. Standard errors are biased upward in the error-in-X case, with a median bias of 12% and a long right tail. Somewhat more concerning is the error-in-Y case, in which uncorrected standard errors are biased downward relative to ground truth values, with a median bias of -12%.

Importantly, this figure's third row shows that this large coefficient bias leads to exceptionally poor coverage. Mean coverage is below 25% in both error-in-X and error-in-Y cases, with recovered 95% confidence intervals rarely containing associated ground

[12] All performance metrics are calculated for only the 40 of the 42 variable pairs that have a statistically significant ($p < 0.05$) relationship in the ground truth data.

[13] Coverage and power are defined for each pair of variables using the distribution of results over bootstrap samples. Thus, the distributions shown in Figure 3 for coverage and power are only over the 40 pairs of variables.

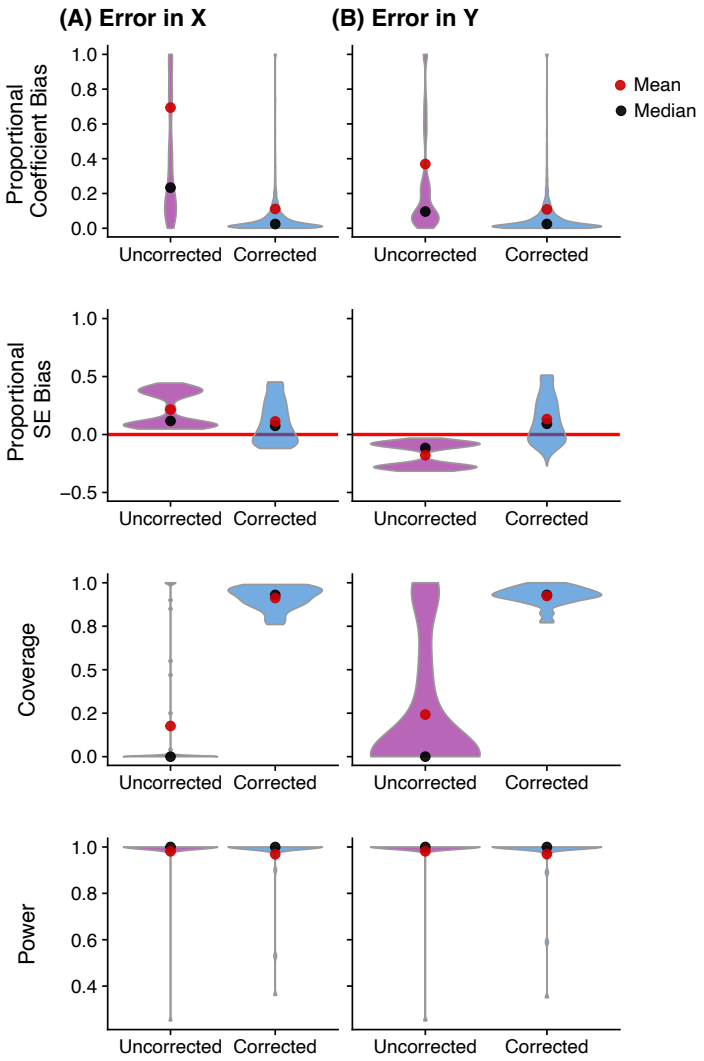

Figure 3: Bias, coverage, and power for regression models using remotely sensed variables both with and without correction via multiple imputation. Figure shows the distribution of absolute proportional coefficient bias, proportional standard error bias, coverage, and power over a set of 100 bootstrapped 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). For proportional biases, 0.25 indicates a 25% bias. Purple distributions indicate regression models in which remotely sensed variables are used without correction as either an independent (panel **(A)**, "error-in-X") or dependent (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. Data for violin plots has been winsorized for visual display purposes only.

truth parameter values. The bottom row of the figure reveals that power is not a concern in the uncorrected models in this experiment, with mean power >95% in both the error-in-X and error-in-Y cases. Sign reversal, while possible, is also not common, as parameter values have the same sign when estimated using remotely sensed measurements and ground truth measurements 99% of the time, despite the substantial parameter bias.

4.2 Differential measurement error is the largest source of bias

Multiple forms of measurement error could be responsible for these biases. It is clear from the substantial bias in the error-in-Y case and the exaggeration of coefficients in the error-in-X case that the assumptions of classical measurement error do not hold in this setting. Moreover, Figure A.2 documents substantial mean-reverting measurement error in nearly all remotely sensed variables (λ estimates from Equation 3 across our six remotely sensed variables range from 0.47 for income to 0.9 for forest cover) while Figure A.3 shows substantial differential measurement error (i.e., non-zero covariance between errors in one variable and levels of other variables). Thus, the second terms in both expressions in Equation 4 are non-zero and contribute to bias.[14]

To decompose overall bias into differential versus mean-reverting sources, Figure A.4 presents the results from an exercise in which we use the linear measurement error model from Section 2 to derive and compute coefficient biases under different assumed error structures. We then compare these hypothetical biases (e.g., what would the bias be if only mean-reverting measurement error were present) to observed coefficient biases across all variable pairs. These results show that while mean reversion, differential error, and classical measurement errors all play a role, differential measurement error is responsible for most of the biases we uncover. Specifically, a measurement error model that allows only for classical and mean-reverting measurement errors explains very little of the variation in observed biases across variable pairs. But, nearly all of the variation is explained when differential error is additionally accounted for. In contrast, 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.[15]

[14] Additional evidence for the presence of non-classical measurement error is found by computing Frisch bounds (Black et al., 2000), which are intervals that contain the true coefficient if the measurement error is classical in an error-in-X setting. Frisch bounds are constructed from forward and reverse regressions using the remotely sensed variable. In 29 out of the 40 variable combinations, the full-sample ground truth estimate falls outside the Frisch bounds for all 100 bootstraps. In an additional 7 combinations, at least one bootstrapped sample has Frisch bounds that do not contain the ground truth estimate.

[15] The biases in standard errors shown in Figure 3 are also consistent with differential, mean-reverting measurement error. We observe consistent downward bias in recovered standard errors for error-in-Y models, indicating the presence of mean-reverting and differential measurement error, but upward bias in standard errors for error-in-X models, consistent with mean-reverting measurement error. However, classical and differential measurement error can push the bias in standard errors in an error-in-X model in either direction, suggesting multiple factors are at play.

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

We find that multiple imputation is highly effective at correcting biases introduced by remotely sensed measurements. Figure 3 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, respectively. 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. When compared to the uncorrected model, multiple imputation mitigates the problem of overly precise standard errors uncovered in the error-in-Y case, while also reducing the upward bias in standard errors uncovered in the error-in-X case.

With low bias and standard errors that account for both sample and imputation uncertainty, Figure 3 shows that models estimated using multiple imputation tend to have good coverage: 95% confidence intervals include the point estimate from the ground truth labels >90% of the time for both error-in-X and error-in-Y cases. The additional uncertainty from imputation, however, lowers mean power slightly, from 98% to 97%, for both 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 very modest reductions in statistical power. Figure B.3 shows that this conclusion is consistent whether the imputation step is implemented using the linear regression based approach, as we have used throughout our main analysis, or alternative methods such as linear regression bootstrapping or nonparametric predictive mean matching.

4.4 Multiple imputation performs relatively well with calibration sets that are small, distant from the main sample, and imperfectly measured

A primary cost of implementing a correction method like multiple imputation is that a calibration set of ground truth labels must be obtained. The experiments shown above use a relatively large and randomly selected calibration sample. 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 when using calibration datasets of different size, spatial distribution, and degree of

measurement error to evaluate the method's generalizability to a wider range of real-world applications.

Sample size: Figure 4 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 employed. 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 all models.[16] As expected, coefficient bias in the multiple imputation models increases as the calibration set size falls. The rate of performance decline 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 generally 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 a calibration dataset that is just 180 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, however, the imputation procedure is poorly constrained and bias increases above the uncorrected model.

The second row of Figure 4 shows that estimated standard errors become increasingly inflated as calibration set size declines, due to increased uncertainty in the imputation step. Further, model coverage decreases slightly, though less than proportionally with the increase in coefficient bias due to increases in the estimated standard errors. Finally, small calibration samples decrease power, which falls to 74% for error-in-X and 73% for error-in-Y when the calibration set size is reduced to 180 observations.

Spatial proximity: As articulated in Section 2, multiple imputation's performance depends on the representativeness of the calibration sample. In practice, calibration data often come from convenience samples obtained in locations that may differ substantially from the main sample. For example, ground truth data may be available only in places that are more populated, exhibit higher incomes, or have higher crop yields. While the specific structure of the calibration data will differ in each empirical setting, here we use spatial proximity of the calibration and main samples to evaluate the implications

[16] In Figures 4 and 5 we report median coefficient and standard error bias across all 40 pairs of variables, 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.

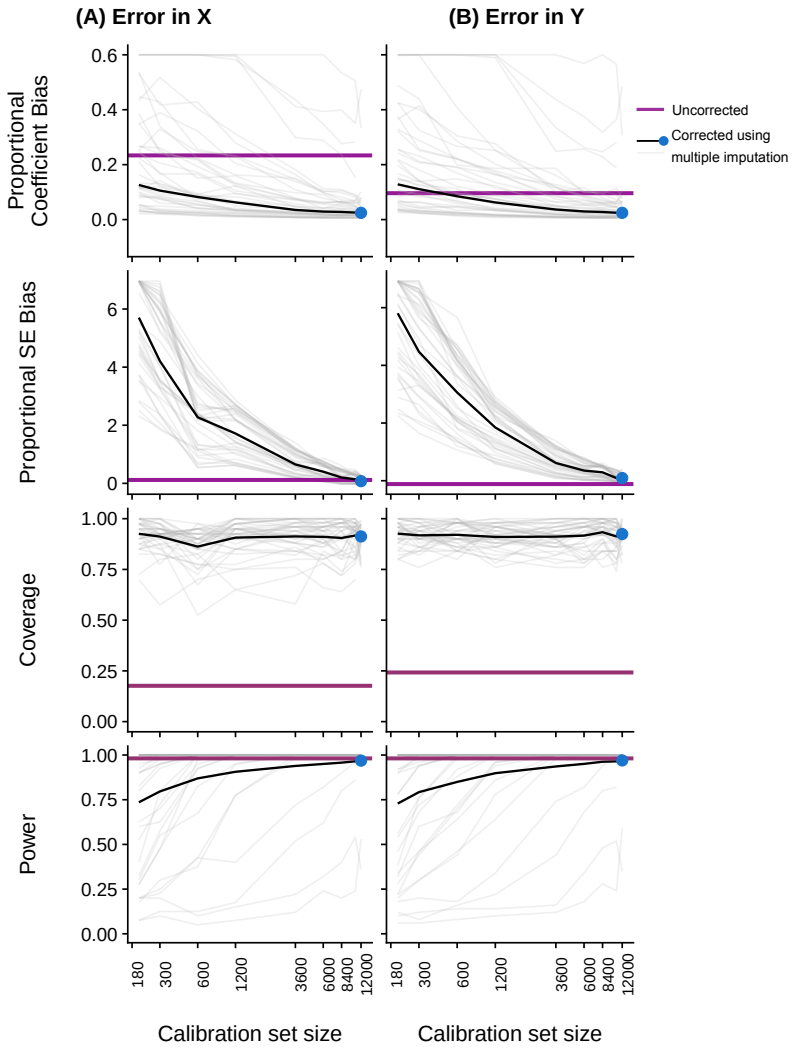

Figure 4: The effect of calibration set size on the ability of multiple imputation to correct biases introduced by remotely sensed variables. Figure shows median bias, mean coverage, and mean power as a function of the size of the dataset available for calibration in the multiple imputation procedure. 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. 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"). The top row shows absolute proportional coefficient bias, while the second row is proportional bias in standard errors.

of non-representative calibration data, given the strong spatial correlations observed in most environmental and economic phenomena. Specifically, 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. To do so, we divide the U.S. into a checkerboard pattern, sampling the main dataset from the

red squares and the calibration dataset from the blue squares in Figure A.6. Increasing the checkerboard square size from a side length of 0.2° to 16° latitude and longitude (∼ 20km to 1600km) increases the separation between the main and calibration datasets. Supplementary Materials Section A.5 provides more details on this procedure.

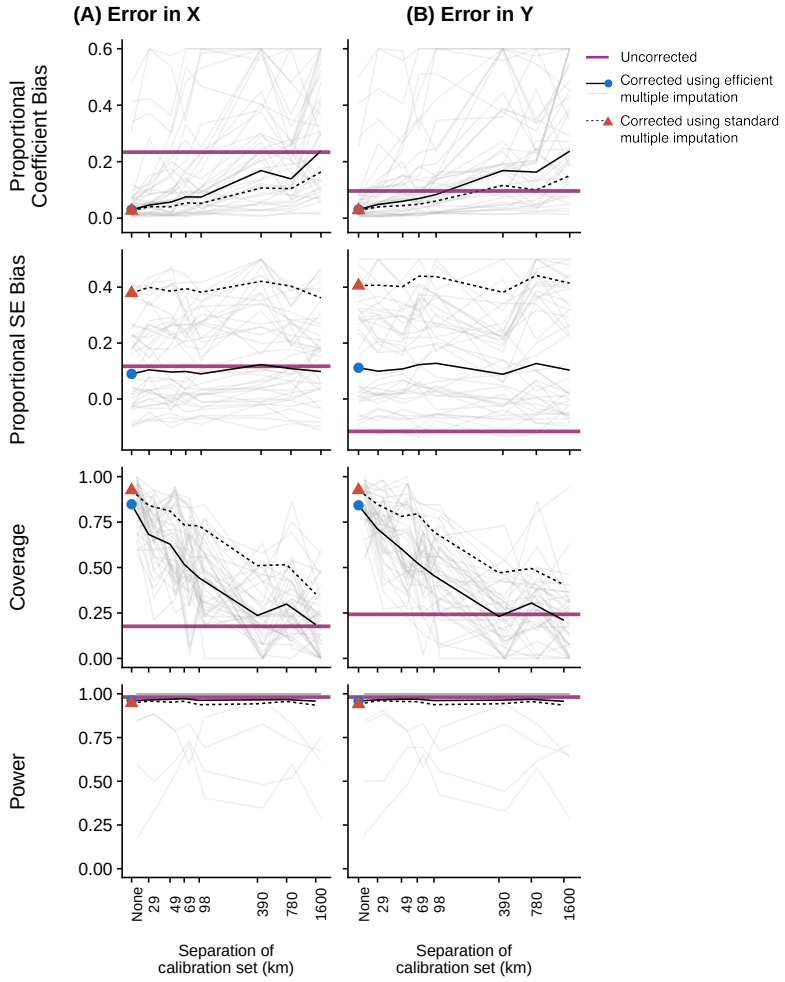

Figure 5: The effect of distance between calibration and main datasets on the ability of multiple imputation to correct biases introduced by remotely sensed variables. Figure shows median bias, mean coverage, and mean power as a function of the degree of separation between the calibration set and the main regression sample (distances indicate the side length of the checkerboard pattern used). 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. Unless noted otherwise, figure depictions are analogous to those in Figure 4.

Figure 5 shows the four performance metrics for multiple imputation regression models (vertical axes) plotted against the width of the checkerboard (horizontal axes). As in Figure 4, 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 geographic 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 at all evaluated distances between the main and calibration datasets, though median bias increases from 2% in the baseline experiment to 24%. Correspondingly, coverage gradually declines from a mean of 92% with a randomly distributed calibration sample to close to the uncorrected level of 19% at a checkerboard square width of ∼1600km. For the error-in-Y case, multiple imputation outperforms the uncorrected model only up to a checkerboard square width of roughly two hundred kilometers, on average. As in Figure 4, the loss of performance is similar in both cases, but the smaller baseline bias in the uncorrected model in the error-in-Y case leads to smaller 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 what is called "standard" in place of "efficient" multiple imputation) 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 5. 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 multiple imputation and the coverage is roughly double in both the error-in-X and error-in-Y cases. Intuitively, combining ground truth and satellite measures in the same regression using efficient multiple imputation is effective so long as the parameter of interest is similar in the main and calibration samples. When the true parameter of interest differs between the populations represented by these two samples, which can occur when the calibration and main datasets are spatially distant, efficient multiple imputation can increase bias and lower coverage. It is likely that optimal combination rules placing nonzero weights on both calibration and main samples could be beneficial in many cases, although exploring this is beyond the scope of this paper.

Errors in ground truth data: We have assumed throughout that ground truth measurements are fully accurate, but, of course, measurement error is also present in such data, potentially impacting parameter bias and the efficacy of multiple imputation. While the implications of measurement error for standard regression analyses have been well-

studied, here we assess how errors in ground truth data impact the performance of multiple imputation with remotely sensed variables. Specifically, we run an experiment in which we add Gaussian noise to the ground truth observations in the calibration dataset before conducting multiple imputation. We evaluate magnitudes of added noise that range from near zero to adding random draws from a distribution with the same variance as the ground truth values.

Figure B.5 shows the four performance metrics for multiple imputation regression models (vertical axes) plotted against the degree of noise added to the calibration data (horizontal axes) for both error-in-X (left column) and error-in-Y (right column). These results indicate that increased measurement error in the calibration data increases parameter bias in the error-in-X case, but not in the error-in-Y case. Intuitively, this occurs because error in the calibration data translates into added noise in the main sample estimation with multiply imputed values. To see this, consider the error-in-X case and recall that in the imputation step we estimate $x = \delta + \gamma\tilde{x} + \psi y + e$ in order to impute values of $\hat{x} = \hat{\delta} + \hat{\gamma}\tilde{x} + \hat{\psi}y + \hat{e}$, where $\hat{e} \sim N(0, \hat{\sigma}^2)$, in the main sample. Adding noise to x in the imputation step increases the variance of e, which increases the variance of $\hat{e}$, adding random noise to imputed values. The effect of this additional random error in the main sample is identical to the effects of classical measurement error; attenuation bias occurs for error-in-X models and estimated standard errors are inflated in the error-in-Y case, although no bias occurs in the error-in-Y case. Therefore, Figure B.5 shows that coverage falls for error-in-X models and that power very slightly declines for both error-in-X and error-in-Y models when ground truth data become highly error-prone.

These results indicate that multiple imputation is still valuable for standard cases of noisy ground truth data. For error-in-X settings, multiple imputation exhibits lower coefficient bias than using the remotely sensed values directly up until the variance of the error in ground truth data amounts to half of the overall variance in the ground truth values. Moreover, multiply imputed standard errors are largely unaffected by noisy ground truth data for error-in-X models. In the error-in-Y case, multiple imputation successfully removes bias from remotely sensed predictions even in the presence of substantial noise in ground truth data, although standard errors are inflated and power may slightly decline.

Synthesis: 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 calibration set ($\sim$ 1-2% the size of the main sample), a relatively distant calibration set (a few hundred kilometers from the main sample), or

a relatively mis-measured calibration data set (with errors as large as half the variance of the true values). In settings with extremely limited calibration data, however, biases can be amplified and power reduced when multiple imputation is used. Importantly, across all settings analyzed, the coverage of multiple imputation models exceeds that of the uncorrected models, with the one exception of using strongly mis-measured calibration data in error-in-X settings.

Across the three experiments shown above, we see that calibration set size and proximity have similar impacts in both error-in-X and error-in-Y cases. Interestingly, errors in calibration data break this symmetry, with mis-measurement of the calibration data inflating bias and degrading coverage in the error-in-X setting but inflating standard errors in only the error-in-Y setting. This indicates that different aspects of calibration data quality can have different effects on parameter recovery depending on whether the dependent or independent variable is corrected by multiple imputation. While the results from these experiments can guide practitioners when evaluating and selecting calibration datasets, the decision of whether calibration data are "good enough" for multiple imputation to improve parameter estimation will vary by context. For example, nearby calibration data that are physically close but come from across a political border may not affect the efficacy of multiple imputation for studies of natural systems but could strongly affect it for studies of human systems. Though we did not evaluate the sensitivity of multiple imputation performance to calibration sets collected in different time periods than the main sample of interest, the question the researcher must ask of the data is the same as when calibration data are from different locations: is the relationship between the variables used in the imputation model similar in the calibration data as in the main sample? If so, multiple imputation is likely to help mitigate bias.

4.5 Additional controls in the imputation step improve precision

Practitioners may ask whether the addition of external covariates – variables that are available in both the calibration and main samples but not used in the estimating equation – may improve the performance of multiple imputation. We show in Figure B.6 that including such additional control variables during the imputation step does not substantially change parameter bias and slightly improves precision in both error-in-X and error-in-Y settings. This is theoretically consistent with the additional controls improving imputation model fit, but not predicting missingness of the ground truth value, which is randomly assigned in this experiment (Carpenter et al., 2023). In empirical settings where the calibration sample systematically differs from the main sample, additional controls

in the imputation step may help address both bias and precision.

4.6 Multiple imputation performs well relative to other correction methods

We evaluate the performance of multiple imputation as compared to other common error correction methods that similarly rely on a calibration sample, including linear regression calibration, nonlinear regression calibration, prediction-powered inference, and complete case analysis. Each of these methods is described in Supplementary Materials Section A.6. We find, consistent with prior literature, that with large and randomly selected calibration datasets multiple imputation approaches outperform other error correction methods for most metrics, especially in the error-in-X setting where bias tends to be largest. Additionally, we show that multiple imputation is as good as or better than alternative error correction techniques when calibration data are limited or spatially distant. These settings have high practical relevance but have yet to be systematically evaluated in the literature (McNeish, 2017). These results are detailed in Supplementary Materials Section A.7.

5 Empirical applications

To evaluate the degree to which our findings generalize to real-world settings, we replicate and apply multiple imputation to four papers spanning diverse variables, geographic settings, and econometric designs (Brooks and Usmani, 2026; Baragwanath and Shinde, 2025; Deschenes et al., 2017; Ratledge et al., 2022). In each replication, we compare estimates made using satellite-based values directly to estimates corrected using multiple imputation. When data availability allows, we also compare to estimates based entirely on ground truth data. In general, these empirical applications indicate that in common empirical settings, measurement error from satellite-based predictions leads to substantial bias in parameter estimates and that multiple imputation can effectively reduce this bias – consistent with the simulation experiments shown above. Correcting measurement error in satellite-based predictions, we find, leads to quantitative and qualitative changes to the main results of all empirical analyses we evaluated.

Applying multiple imputation in these four datasets requires adapting its implementation to four distinct panel data settings. Below, we first develop a method to use multiple imputation with panel data and fixed effects. We then summarize the findings of our four replication analyses. Additional details on each replication are provided in Supplementary Discussion C.

5.1 Adapting multiple imputation for panel data settings

Panel datasets with rich controls are increasingly used in empirical studies to lend credibility to the causal interpretation of estimated coefficients of interest by accounting for potentially confounding factors. As discussed above, any controls – including fixed effects – included in the main sample estimation should also be included in the imputation step to ensure congeniality between imputation and main sample estimation. This can complicate estimation in some settings where it is not possible to estimate all controls in both stages of multiple imputation. For example, spatial unit fixed effects for some spatial units will not be estimable in the imputation step if there is not full spatial coverage of ground truth data. We show that in these settings, fixed effects can be projected out of the data, rather than being controlled for, using an approach motivated by the Frisch-Waugh-Lovell Theorem (Lovell, 2008). Here, we illustrate this approach in a general fixed effects framework, before applying it to a diversity of settings in the four replications below.

Consider a panel fixed effects regression in which an outcome, Y_{it}, is regressed on a treatment X_{it} with spatial unit, i, and temporal unit, t, fixed effects: $Y_{it} = \beta X_{it} + \lambda_i + \theta_t + \varepsilon_{it}$. For illustration, assume an error-in-Y setting. Suppose that while the calibration data contain some observations for all time periods, such that θ_t can be estimated in the calibration sample and then used to predict values in the main sample, only a subset of all counties are available in the calibration data. In this setting, λ_i can be estimated in the imputation step only for the counties with ground truth data, and thus imputations cannot be made for all of the counties in the main sample.

To address this challenge, we propose an approach in which all regression variables are residualized by λ_i (or, in general, any set of fixed effects that cannot be included as controls) prior to the multiple imputation analysis. This effectively controls for these fixed effects throughout the entire multi-step procedure. After residualization, multiple imputation can be conducted in the standard manner: first, we estimate the imputation model using the residualized data in the calibration sample, including the temporal fixed effect:

$$\ddot{\mathrm{Y}}_{it} = \gamma \ddot{\tilde{\mathrm{Y}}}_{it} + \ddot{\theta}_t + e_{it}., \tag{9}$$

where residualized variables are denoted with a double dot superscript. Second, we impute all residualized observations of Y in the main sample using Equation 9 K times, indicating each predicted observation for imputation k as $\ddot{\tilde{\mathrm{Y}}}^k_{it}$. Finally, we estimate K

iterations of the relationship of interest in the main sample:

$$\hat{\ddot{Y}}_{it}^{k} = \beta^{k}\ddot{X}_{it} + \ddot{\theta}_{t}^{k} + \epsilon_{it}^{k}. \tag{10}$$

Coefficients and standard errors are then computed using Rubin's Rules (Equation 8).

We apply this general procedure of controlling for all possible fixed effects and projecting out the others to each of the four empirical analyses below. In each, we compare recovered coefficients before and after applying multiple imputation. In the two empirical analyses where the coefficient of interest can be estimated directly from ground truth values available in the entire main sample, we show how multiple imputation performs relative to the same research design estimated using ground truth data.

5.2 Results: empirical applications

The effects of the U.S. NO_x Budget Program on $PM_{2.5}$. In Deschenes et al. (2017), the authors estimate the effects of the U.S. NO_x budget program on ambient air pollution using a "triple-difference" empirical design. In Table 1, column 4, we replicate the authors' main finding that the NO_x budget program reduced average county $PM_{2.5}$ by 1.03 $\mu g/m^3$, estimated using ground truth air pollution data measured from local monitors (this result matches the authors' estimate in their Table 2, column 5).

We then ask what this analysis would have recovered had satellite data been used instead of pollution monitor readings. Specifically, we repeat the analysis using satellite-based $PM_{2.5}$ concentrations from Van Donkelaar et al. (2021), a commonly used satellite-based estimate of particulate matter. Table 1, column 3 shows that using remotely sensed air pollution data substantially attenuates the estimated effect of the budget program on air pollution, lowering the point estimate by ∼50% and reducing standard errors by ∼35%. In turn, the 95% confidence interval estimated with remotely sensed data does not contain the authors' original point estimate. Using the linear measurement error model to estimate the magnitude of differential and mean-reverting measurement error with matching observations of monitor data and satellite data, we find that bias induced is largely driven by mean-reverting measurement error ($\lambda = 0.432$), which more than halves the ground truth estimate. A moderate amount of differential measurement error $\frac{\sigma_{xu_y}}{\sigma_x^2}$ = -0.132 somewhat offsets this attenuating effect.

Next, we evaluate whether multiple imputation can effectively correct for this bias using a calibration set composed of ground truth values for 30% of the studied counties. Column 6 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.[17] Moreover, the corrected 95% confidence interval now contains the original point estimate.

The efficacy of enforcement actions for regulating deforestation in Brazil. Baragwanath and Shinde (2025) estimate how strict environmental regulations and enforcement actions in Brazil, called the deforestation "Blacklisting" program, impact deforestation rates using a synthetic difference-in-differences estimator. Table 1, column 3 replicates the authors' finding that Brazil's 2008 municipal Blacklisting policy reduced deforestation by 0.502 percentage points (a 47.6% reduction in baseline rates) when using a coarse satellite-based measure of deforestation called *Prodes*, which is used by the Brazilian government to detect deforestation (this result matches authors' estimate in their Table 2 column 1).

In column 4, we also replicate the authors' finding that when using a more accurate finer-resolution satellite-based measure of deforestation, *MapBiomas*, the estimated effect of the policy is substantially less negative, falling by >20%. This change is due to the improved deforestation data capturing deforested patches smaller than the 6.25 hectare detection limit of the inferior *Prodes* data. In this setting we view the *Prodes* data as the error-prone measure and the more accurate *Mapbiomas* data as ground truth; though of course neither measurement is fully accurate.

This is a clear example of differential measurement error (in the *Prodes* data) leading to bias in the estimated coefficient of interest. The treatment (Blacklisting) is negatively correlated with errors in measurement of the outcome variable (deforestation) because after treatment, regulated agents increase cutting of forest patches smaller than 6.25 hectares to strategically avoid detection by the government's inferior *Prodes*-based monitoring system. This means that measured deforestation is systematically lower for *treated groups* than it should be after treatment when measured using the *Prodes* data, leading to exaggerated point estimates.[18] Using the linear measurement error model and estimating error correlations in a matched sample of *Prodes* and *MapBiomas* data, we find that nearly all of the bias in the coefficient estimated using the *Prodes* data is driven by differential measurement error.

Next, we test the ability of multiple imputation to correct for errors in the *Prodes* data using a 30% calibration sample of the more accurate *MapBiomas* data. While *MapBiomas* data is available across the entire study area, this exercise mimics a situation where

[17] Note that the $PM_{2.5}$ 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.

[18] From Equation 4, $\mathbb{E}[\hat{\beta}_{\tilde{y}}] = \lambda_y \beta + \frac{\sigma_{xu_y}}{\sigma_x^2}$, and we observe that $\sigma_{xu_y} < 0$, such that the differential measurement error adds a negative bias.

researchers are able to collect only a relatively small sample of accurate measures due to resource or accessibility constraints. Table 1, column 5 shows that multiple imputation reduces bias by 83% and leads to an estimated effect of Blacklisting of -0.416 percentage points. Consistent with the real data simulation experiments above, this decrease in bias comes with an increase in the uncertainty of the parameter estimate.

The impacts of electrification on wealth in Uganda. Ratledge et al. (2022) estimate the impact of electricity access on wealth in Uganda using a machine-learning based method for causal inference called matrix completion (MC), along with other approaches. Their wealth outcome variable is predicted using a deep learning method trained on daytime satellite images. In Table 1, column 3, we replicate the authors' main finding that electricity access increases their predicted asset-based wealth metric by 0.23 units, which is an increase of 0.15σ (this result matches the authors' estimate in Figure 3b).

We then apply multiple imputation using survey-based wealth measures from the Demographic and Health Surveys (DHS) in Uganda as a calibration dataset ($N = 964$), noting that the authors also train their original predictive model using DHS data from across multiple countries. Table 1, column 5 shows that correcting the satellite-based estimates using these survey data nearly doubles the estimated effect of electricity access to 0.41. Comparison of the satellite-based and survey-based wealth estimates in the calibration sample indicates that the satellite-based estimates have both mean-reverting measurement error ($\lambda = 0.56$), and differential measurement error ($\frac{\sigma_{xuy}}{\sigma_x^2} =$ -0.27, implying that wealth is under-estimated in electrified communities). Both of these bias the authors' coefficient downward. We note that because the DHS data represent a repeated cross-section, rather than a panel, calibration and quantification of error structures could not be conducted using the full set of controls used in the model.[19] The findings from this replication highlight both the first-order implications that measurement error – and its correction – can have on policy-relevant parameter estimates, and the need for additional ground data to produce, evaluate, and correct remote sensing data for use in scientific inference.

The effects of $PM_{2.5}$ on athletic performance. Brooks and Usmani (2026) quantify the impact of air pollution on the performance of professional Indian cricket players using a two-way fixed effects specification and remotely sensed $PM_{2.5}$. Table 1, column 4 replicates their main finding that a one-unit increase in remotely sensed $PM_{2.5}$ during the day of the cricket match is associated with a 0.41 percentage point increase in run

[19]Specifically, we could not include location specific fixed effects because that would remove all of the variation in the DHS data; see Supplementary Discussion C for details.

probability (this result matches the authors' estimate in their Table 3, column 3). The mechanism they use to explain this phenomenon is the greater physical demands of the bowling (defensive) team than the batting (run-scoring) team.

We then apply multiple imputation using a calibration sample of stadium-level $PM_{2.5}$ ground measurements available for 27% of the 773 cricket matches analyzed. Table 1, column 6 shows a corrected estimate that is roughly half the size of the satellite-based estimate. Using the decomposition from the linear measurement error model, we find that the bias is due to substantial differential measurement error in the satellite-based measures, which inflates the uncorrected coefficient. This inflation is somewhat offset by mean reverting measurement error, which modestly attenuates the estimated coefficient.[20]

Synthesis Consistent with the results from the real data simulation experiments, these four empirical applications show the substantial impact that measurement error in remotely sensed measurements can have on parameter recovery in empirical analysis. Uncorrected parameters of interest are commonly double or half the estimates that we recover using multiple imputation. When ground truth data are available for the main estimating sample, biases are similarly large. These analyses show the practical usefulness of multiple imputation to correct for error in satellite measurements and to reduce bias in estimated parameters. With quantification of both mean-reverting and differential measurement error, they also build intuition for *why* the estimates change when corrected, enabling researchers to judge the plausibility of different error sources and structures in their own context. The diversity of study settings (United States, Brazil, India, Uganda) and estimation methods employed (two-way fixed effects, synthetic differences-in-differences, and matrix completion) demonstrates the generalizability of our findings.

[20] Using Equation 4, we find a relatively small multiplicative influence of mean-reverting measurement error ($\frac{\lambda_x \sigma_x^2}{\lambda_x^2 \sigma_x^2 + \sigma_{u_x}^2} = 0.8$, $\lambda_x = 0.26$) and a substantial additive contribution to bias from differential measurement error ($\frac{\sigma_{y u_x}}{\lambda_x^2 \sigma_x^2 + \sigma_{u_x}^2} = 0.0014$).

Y	X	**Uncorrected** ($\tilde{\beta}$)	**Truth** (β)	**MI** ($\hat{\beta}_{MI}$)	% $\Delta^{\tilde{\beta}\to\hat{\beta}_{MI}}$	**Source**
$\underline{PM_{2.5}}$	NBP	−0.52 (0.176)	**−1.03** (0.27)	−0.819 (0.22)	+58%	Deschenes et al. (2017)
Deforestation	Blacklisting	**−0.502** (0.107)	**−0.398** (0.099)	−0.416 (0.388)	−17%	Baragwanath and Shinde (2025)
Wealth	Electrification	**0.225** (0.0689)	–	0.405 (0.168)	+80%	Ratledge et al. (2022)
Athlete's performance	$\underline{PM_{2.5}}$	**0.0041** (0.00136)	–	0.0021 (0.00106)	−49%	Brooks and Usmani (2026)

Table 1: Application of multiple imputation to recent papers in a variety of settings using remotely-sensed variables for causal inference. This table presents results of applying multiple imputation (MI) to four published and working papers. All applications use panel data with a variety of estimation strategies and specifications detailed in the main text. The dependent (Y) and independent (X) variables are indicated in the first two columns, with the remotely sensed variable underlined. The *Uncorrected* column presents the estimated effect of X on Y using the remotely sensed variable directly, the *Truth* column presents the estimated effect using ground truth data, when it is available in the main estimating sample. Bolded values indicate coefficients reported in the original manuscript. The *MI* column reports the corrected coefficient using efficient multiple imputation and column 6 reports the respective percentage change in magnitude induced by correction.

6 Discussion

As the applications of remotely sensed data continue to grow in 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 evaluate the risks that measurement errors in remotely sensed data pose for parameter recovery in regression analyses. First, we quantify the biases introduced by remotely sensed variables when used in regression analysis. We uncover substantial and economically relevant bias in regression coefficients and associated measures of uncertainty when using remotely sensed variables across a diversity of empirical settings and estimation strategies. Second, we demonstrate that a standard statistical technique for imputation of missing data, multiple imputation, performs well at mitigating these biases in a range of contexts, as long as researchers have access to some amount of ground truth data for calibration. We synthesize our results into a practical guide for researchers using remotely sensed data in regression analysis in Figure 6. These findings apply most directly to studies using remotely sensed variables, but are relevant more broadly to analyses relying on machine learning predictions in downstream regressions.

There are a few important features and limitations of our analysis to keep in mind. First, the empirical returns to using multiple imputation as a correction method are higher for remotely sensed measurements with larger errors. For example, Figure B.4 shows that within the Rolf et al. (2021) benchmark dataset, variables with higher R^2 in the underlying remote sensing model exhibit lower bias when used in regression analysis without correction. 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, 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. 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 A.7 and A.8 that multiple imputation becomes less effective in this setting.

Third, as discussed earlier, multiple imputation performs best when calibration data are representative of the main sample, conditional on the model controls in the estimation and imputation steps. While we explore how geographic distance between the main and

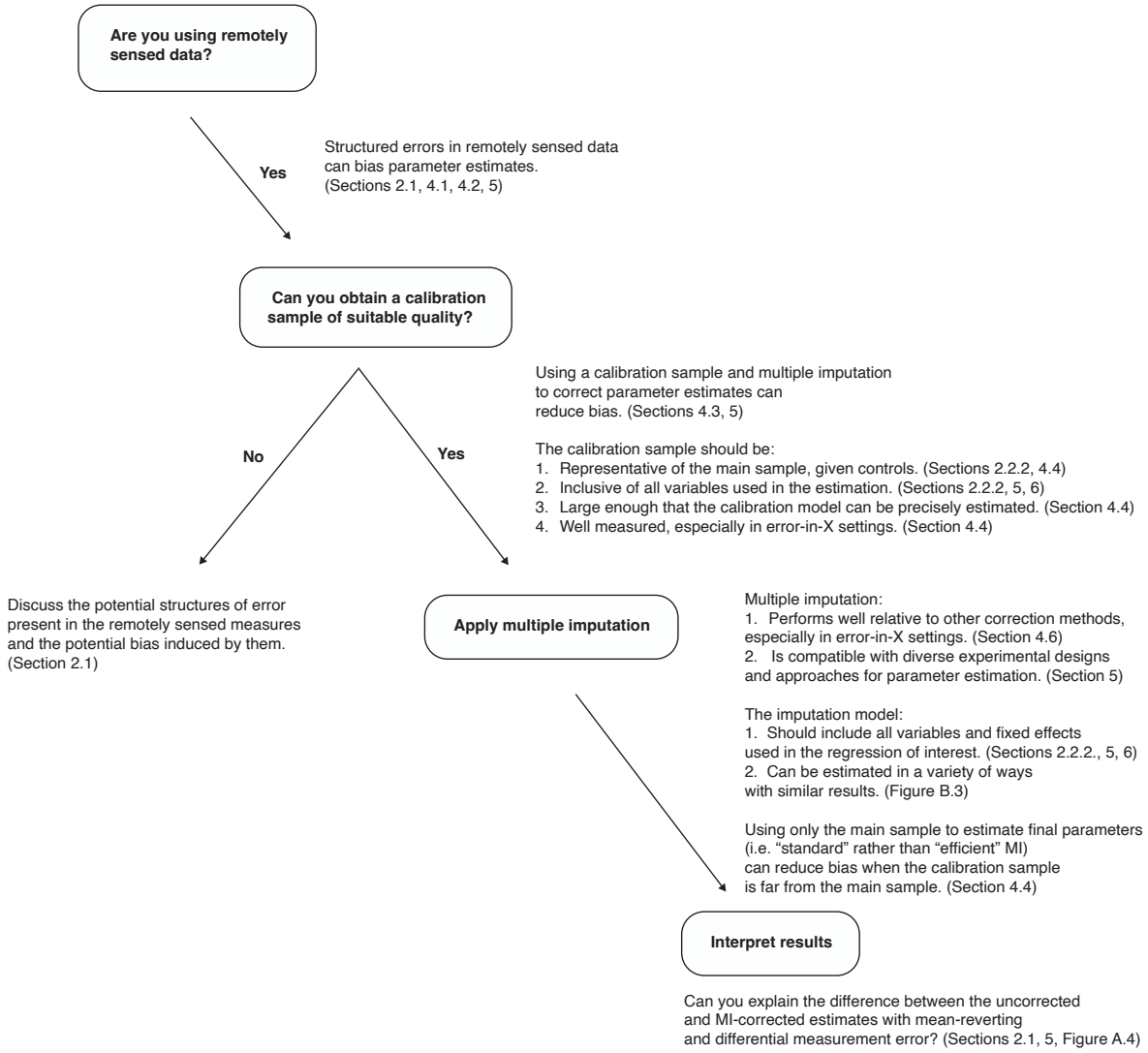

Figure 6: A practical guide for the use of satellite data in parameter estimation and the implementation of multiple imputation.

calibration set can lead to distribution shift between them, other factors (e.g., income, urbanization, or institutional infrastructure) could also lead to systematic differences. The value of a conditionally representative calibration set also highlights the importance of controlling for the same variables in both the imputation step and estimation step – residual variation after accounting for controls may differ from non-residual variation. This is especially important in specifications with large numbers of controls, such as fixed-effects models. For other tips and answers to frequently asked questions about using multiple imputation in practice see Carpenter et al. (2023) as well as Little and Rubin (2019); Keogh and Bartlett (2021); Schafer (1999) and van Buuren (2012).

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`,[21] and generalizes well across a wide range of empirical contexts, including when calibration data are limited and when regression models leverage panel data and standard fixed effects research designs.

References

Alix-Garcia, Jennifer, and Daniel Millimet. 2023. Remotely incorrect? accounting for nonclassical measurement error in satellite data on deforestation. *Journal of the Association of Environmental and Resource Economists* .

Angelopoulos, Anastasios N, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science* 382 (6671): 669–674.

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*, volume 111. 415–19.

Baragwanath, Kathryn, and Nilesh Shinde. 2025. Beyond the canopy: How satellite data detection thresholds influence policy evaluation and deforestation behavior. *Journal of Environmental Economics and Management* : 103219.

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.

[21] 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`.

Black, Dan A, Mark C Berger, and Frank A Scott. 2000. Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association* 95 (451): 739–748.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement error in survey data. In *Handbook of Econometrics*, volume 5. Elsevier, 3705–3843.

Brooks, Matthew S., and Faraz Usmani. 2026. The pollution–productivity curve: Non-linear effects and adaptation in high-pollution environments. Working paper. Available at: https://mspitzerbrooks.github.io/files/Brooks_PollutionProductivityCurve.pdf.

Brown, Christopher F, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, et al. 2025. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291* .

Carpenter, James R, Jonathan W Bartlett, Tim P Morris, Angela M Wood, Matteo Quartagno, and Michael G Kenward. 2023. *Multiple imputation and its application*. John Wiley & Sons.

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

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21 (1): C1C68.

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).

Christensen, Darin, Tamma Carleton, Esther Rolf, Cullen Molitor, Shopnavo Biswas, Karena Yan, and Graeme Blair. 2025. Estimating the footprint of artisanal mining in africa. Technical report, National Bureau of Economic Research.

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

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.

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*, volume 109. 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. *Statistics in Medicine* 27 (25): 5195–5216.

Garcia, Alberto, and Robert Heilmayr. 2024. Impact evaluation with nonrepeatable outcomes: The case of forest conservation. *Journal of Environmental Economics and Management* 125: 102971.

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

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society* : 153–161.

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.

Keogh, Ruth H, and Jonathan W Bartlett. 2021. Measurement error as a missing data problem. In *Handbook of Measurement Error Models*. Chapman and Hall/CRC, 429–452.

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

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.

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*, volume 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.

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

Lu, Kerri, Dan M Kluger, Stephen Bates, and Sherrie Wang. 2025. Regression coefficient estimation from remote sensing maps. *Remote Sensing of Environment* 330: 114949.

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.

Meng, Xiao-Li. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* : 538–558.

Millimet, Daniel L. 2011. The elephant in the corner: a cautionary tale about measurement error in treatment effects models. In *Missing data methods: Cross-sectional Methods and Applications*, volume 27. Emerald Group Publishing Limited, 1–39.

Pelletier, Johanne, Mira Korb, Solomon Alemu, Manex Bule Yonis, Travis J Lybbert, and Matthieu Stigler. 2025. Causal inference with predicted outcomes: Correcting prediction error bias in satellite-based impact evaluation. *Journal of Development Economics* : 103655.

Potapov, Peter, Svetlana Turubanova, Matthew C Hansen, Alexandra Tyukavina, Viviana Zalles, Ahmad Khan, Xiao-Peng Song, et al. 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.

Rambachan, Ashesh, Rahul Singh, and Davide Viviano. 2024. Program evaluation with remotely sensed outcomes. *arXiv preprint arXiv:2411.10959* .

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, et al. 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*, volume 81. John Wiley & Sons.

Sanford, Luke C., Megan Ayers, Matthew Gordon, and Eliana Stone. 2025. Adversarial debiasing for unbiased parameter recovery.

Schafer, Joseph L. 1999. Multiple imputation: a primer. *Statistical methods in medical research* 8 (1): 3–15.

Sherman, Luke, Jonathan Proctor, Hannah Druckenmiller, Heriberto Tapia, and Solomon Hsiang. 2026. Global high-resolution estimates of the un human development index using satellite imagery and machine learning. *Nature Communications* 17 (1): 1315.

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.

Torchiana, Adrian L, Ted Rosenbaum, Paul T Scott, and Eduardo Souza-Rodrigues. 2023. Improving estimates of transitions from satellite data: a hidden markov model approach. *Review of Economics and Statistics* : 1–45.

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, 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.

Wardle, Arthur R. 2025. Addressing bias from misclassification in predicted data: Theory and application to satellite crop data .

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

Xie, Xianchao, and Xiao-Li Meng. 2017. Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica* : 1485–1545.