NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

02-17

Inference for Errors-in-Variables Models in the Presence of Spatial and Temporal Dependence with an Application to a Satellite Remote Sensing Campaign

Bohai Zhang, Noel Cressie, and Debra Wunch

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Inference for Errors-in-Variables Models in the Presence of Spatial and Temporal Dependence with an Application to a Satellite Remote Sensing Campaign

Bohai Zhang

National Institute for Applied Statistics Research Australia

University of Wollongong, NSW 2522, Australia

email: bohai@uow.edu.au

and

Noel Cressie

National Institute for Applied Statistics Research Australia,

University of Wollongong, NSW, 2522, Australia

email: ncressie@uow.edu.au

and

Debra Wunch

Department of Physics and School of the Environment,

University of Toronto, Toronto, ON, M5S 3E8, Canada

email: dwunch@atmosp.physics.utoronto.ca

Abstract

Motivated by a satellite remote sensing mission, this article proposes an errors-in-variables (EIV) multiple-regression model with heteroscedastic errors for relating the satellite data products to similar products from a well characterized but globally sparse ground-based dataset. In the remote sensing setting, the model is used to remove the global offset of the satellite data from the more-accurate ground-based data. The error structure of the proposed EIV model comprises two components: A random-error component whose variance is inversely proportional to the underlying sample size for computing the regression data, and a systematic-error component whose variance remains the same as the underlying sample size increases. In this article, we discuss parameter identifiability for the proposed model and provide a two-stage parameter-estimation procedure. In Stage 1, the random-error variances are estimated by spatial/temporal-process modeling of individual observations that, when aggregated, yield the regression data. In Stage 2, the regression data are used to estimate the remaining variance-component parameters as well as the regression coefficients from adjusted estimating equations that yield asymptotically unbiased estimators. We illustrate the proposed procedure through both simulation studies and an application to validating measurements of atmospheric column-averaged CO_2 from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite, using corresponding ground-based measurements from the Total Carbon Column Observing Network (TCCON).

Keywords: Calibration, estimating equations, OCO-2, regression analysis, systematic error, TCCON

1. INTRODUCTION

It is a common problem in the environmental sciences that plentiful space-borne measurements have imperfect bias (accuracy) and variance (precision) properties. More generally, an important part of many scientific studies is a validation component, where less-plentiful but higher-precision and greater-accuracy measurements are matched to the more-plentiful measurements. For example, soil scientists measure the mineral content of soil cheaply using proximal γ-ray sensors, but they also take costly *in situ* soil samples that are sent to the laboratory for analysis (Viscarra Rossel et al., 2007). Another example, which we return to later, is remote sensing of trace gases in the atmosphere by satellites. Polar-orbiting satellites give plentiful measurements with global coverage in a matter of days (e.g., Crisp et al., 2004). In contrast, concomitant ground-based monitoring stations are sparsely distributed on Earth's surface (e.g., Wunch et al., 2011).

These examples are motivation for the methodological problem that we address in this article. Let Y represent a measurement of a phenomenon of interest, which is coincident with $p \geq 1$ covariate measurements, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$, that are available. There are many more measurements of Y than of $\mathbf{X}$, but there are enough simultaneous measurements of $(\mathbf{X}, Y)$ to allow a calibration equation to be fitted, from which all the measurements Y are adjusted. The covariates include more precise measurements of the same phenomenon and could include physical variables, for example, latitude (spatial), month (temporal), or solar zenith angle (geometric) in the remote sensing application.

Let $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$ denote the regression data, to which the calibration line,

$$Y_i = a + \mathbf{b}^T \mathbf{X}_i + \text{error}_i \ , \tag{1}$$

is fitted. An ordinary-least-squares-fit results in

$$(\hat{a}_{ols}, \hat{\mathbf{b}}_{ols}^T)^T = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y},$$

where

$$\mathbf{Y} \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}; \quad \tilde{\mathbf{X}} \equiv \begin{pmatrix} 1 \ \mathbf{X}_1^T \\ \vdots \quad \vdots \\ 1 \ \mathbf{X}_N^T \end{pmatrix}.$$

The resulting fitted line is $Y_{ols} = \hat{a}_{ols} + \hat{\mathbf{b}}_{ols}^T \mathbf{X}$.

The estimates $\hat{a}_{ols}$ and $\hat{\mathbf{b}}_{ols} \equiv (\hat{b}_{ols,1}, \ldots, \hat{b}_{ols,p})^T$ might be used for calibration. That is, suppose that Y^0 is observed along with $\mathbf{X}^0$ and, without loss of generality, that the first covariate measures the same phenomenon with a higher precision and accuracy. Then a straightforward correction is to adjust Y^0 with $Y_{ols}^0 \equiv (Y^0 - \hat{a}_{ols} - \hat{\mathbf{b}}_{ols,-1}^T \mathbf{X}_{-1}^0)/\hat{b}_{ols,1}$, where $\mathbf{X}_{-1}^0$ contains the covariate measurements in $\mathbf{X}^0$ except for the first covariate, and $\hat{\mathbf{b}}_{ols,-1}$ is the vector of the least-squares estimates of regression coefficients corresponding to $\mathbf{X}_{-1}^0$. We show below that Y_{ols}^0 is a naive adjustment that is generally biased and inefficient.

In general, both the response Y and the covariates $\mathbf{X}$ are measured with errors, which might be made up of both fixed effects and random effects. Thus, (1) should be modified to an errors-in-variables (EIV) model (e.g., Fuller, 1987). For $i = 1, \ldots, N$, assume

$$\begin{aligned} \mathbf{X}_i &= \mathbf{x}_i + \boldsymbol{\delta}_{x,i}, \\ Y_i &= y_i + \delta_{y,i}, \\ y_i &= a + \mathbf{b}^T \mathbf{x}_i, \end{aligned} \tag{2}$$

where $\mathbf{x}_i$ is a p-dimensional (fixed but unknown) vector of the mean of $\mathbf{X}_i$, and likewise y_i is a (fixed but unknown) scalar mean that is linearly related to the covariates' mean, $\mathbf{x}_i$. The errors $\boldsymbol{\delta}_{x,i}$ and $\delta_{y,i}$ are often assumed to be independent Gaussian vectors/variables $\mathcal{N}(\mathbf{0}, \Sigma_{x,i})$ and $\mathcal{N}(0, \sigma_{y,i}^2)$, respectively. The third equation in (2) expresses the linear relationship on the means $\{(\mathbf{x}_i, y_i)\}$, and all the error structure is absorbed into the first two equations. Notice that (2) reduces to (1) when $\boldsymbol{\delta}_{x,i} = \mathbf{0}$ for all $i = 1, \ldots, N$ (i.e., when there are no "errors in the variables" $\{\mathbf{X}_i\}$). Inference for such EIV models has been discussed extensively in a number of books (e.g., Fuller, 1987; Cheng and Van Ness, 1999; Carroll et al., 2006).

Now, consider the error terms in (2) to include both fixed effects and random effects; that is, for $i = 1, \ldots, N$,

$$\boldsymbol{\delta}_{x,i} = \boldsymbol{\mu}_{x,i} + \boldsymbol{\epsilon}_{x,i}, \quad \delta_{y,i} = \mu_{y,i} + \epsilon_{y,i},$$

where the μ-terms are fixed and represent bias, and the ϵ-terms are random and represent mean-zero noise. That is, $\boldsymbol{\delta}_{x,i} \sim \mathcal{N}(\boldsymbol{\mu}_{x,i}, \Sigma_{x,i})$ and $\delta_{y,i} \sim \mathcal{N}(\mu_{y,i}, \sigma^2_{y,i})$ in (2).

Closer inspection reveals that this is an over-parameterized model for which estimation of a and $\mathbf{b}$ would be problematic. A way out of this difficulty is to replace the fixed effects (i.e., μ-terms) with random effects whose distributions depend on far fewer parameters. That is, combine the EIV model (2) with the random-effects model,

$$\boldsymbol{\delta}_{x,i} = \boldsymbol{\eta}_{x,i} + \boldsymbol{\epsilon}_{x,i}, \quad \delta_{y,i} = \eta_{y,i} + \epsilon_{y,i}, \text{ for } i = 1, \ldots, N, \tag{3}$$

where $\{\boldsymbol{\eta}_{x,i}\}$ and $\{\eta_{y,i}\}$ are mutually independent. Further, assume that $\{\boldsymbol{\eta}_{x,i} : i = 1, \ldots, N\}$ are independent and identically distributed (iid) and follow a Gaussian distribution, $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\tau^2_{x,1}, \ldots, \tau^2_{x,p}))$; and $\{\eta_{y,i} : i = 1, \ldots, N\}$ are iid and follow a Gaussian distribution, $\mathcal{N}(0, \tau^2_y)$. Notice that the parameterization in (3) replaces $N(p+1)$ fixed-effects parameters $\{(\boldsymbol{\mu}_{x,i}, \mu_{y,i}) : i = 1, \ldots, N\}$ with just $(p+1)$ random-effects parameters $\{\tau^2_{x,1}, \ldots, \tau^2_{x,p}, \tau^2_y\}$. Here we use the random effects $\{\boldsymbol{\eta}_{x,i}\}$ and $\{\eta_{y,i}\}$ to model the systematic error in $\{\mathbf{X}_i\}$ and $\{Y_i\}$, respectively. We propose to use the error structures in (3) to fully account for uncertainties present in $\{\mathbf{X}_i\}$ and $\{Y_i\}$; that is, the "errors" in the EIV model are given by (3).

The leading application in this article is to a validation component of a remote sensing campaign. Satellite remote sensing measurements of Earth's atmosphere and surface can collect measurements on a global scale within a matter of days, which helps scientists understand the spatial distributions and temporal dynamics of environmental variables. Many remote sensing instruments rely on reflected sunlight from Earth's surface, which can be affected by clouds and aerosols in the atmosphere. The reflectivity of Earth's surface (albedo), aerosols, surface pressure, and a multitude of other variables affect the retrieval of the primary variable of interest (e.g., atmospheric carbon dioxide). All retrieved parameters possess uncertainty, but some (e.g., aerosols) are more uncertain than others. Consequently, the resulting data products typically have variabilities that potentially contain biases. In order to validate satellite remote sensing measurements, they are matched with more precise and accurate (usually ground-based) datasets to identify the biases. This is often achieved by fitting a linear-regression relationship with a small subset of the plentiful satellite data as the response (Y) and a coincident set of the more accurate and precise data sources as the covariates ($\mathbf{X}$). There are potentially systematic and random errors in all variables, which leads us to choose the EIV regression model (2) and (3) to obtain a best fit.

In this paper, our application focuses on the validation of column-averaged dry-air mole fractions of atmospheric carbon dioxide (CO_2) concentrations, collected by the Orbiting Carbon Observatory-2 (OCO-2) satellite (Crisp et al., 2017). These column-averaged dry-air mole fractions of CO_2 measurements are called "XCO_2," in units of parts per million (ppm). The OCO-2 satellite provides an unprecedented opportunity for observing XCO_2 in the atmosphere, which is a key component of Earth's carbon cycle. The resulting OCO-2 data products need to have high accuracies and precisions that tie to a scale set by the World Meteorological Organization (WMO), in order to identify regional sources and sinks of CO_2 (Wunch et al., 2017).

To tie the OCO-2 observations to the WMO scale, more accurate and precise observations from the Total Carbon Column Observing Network (TCCON) (e.g., Wunch et al., 2011, 2015) are used. The TCCON observations are collected from around 25 ground-monitoring stations, which provide spatially sparse but temporally rich data products. The OCO-2 satellite has a special observation mode, referred to as "target mode," in which the OCO-2 spacecraft "stares" at a ground location (usually, a TCCON station) as it passes overhead (Wunch et al., 2017). The target-mode observations, which cover a small ($0.2°$ longitude $\times$ $0.2°$ latitude) geographic region, are retrieved over a period of a few minutes. They are considered to be coincident with the TCCON time series generated in a 2-hour time window centered at the mean target time of the OCO-2 observations. For each of these target-mode maneuvers, thousands of individual OCO-2 observations are aggregated to form one value

Y, and around 65 of their coincident individual TCCON observations are aggregated to form one value X, resulting in a point (X, Y) in a regression analysis. This is repeated at a number of TCCON locations and times, resulting in regression data $\{(X_i, Y_i) : i = 1, \ldots, N\}$; for example, N was equal to 66 for obtaining Version 7 of the OCO-2 data product (Section 5).

After an initial adjustment of the OCO-2 data using environmental variables such as the surface-pressure offsets and the abundance of aerosols (Mandrake et al., 2015), a regression relationship between the (aggregated) OCO-2 target-mode data (Y) and the (aggregated) TCCON data (X) was computed. This provided a global offset between the retrieved XCO_2 from OCO-2 and TCCON, resulting in the Version 7 OCO-2 data product (Mandrake et al., 2015). In this article, we shall re-do the calculation of the global offset based on the statistical methodology we develop.

The rest of paper is organized as follows. In Section 2, we motivate our proposed methodology by first reviewing the EIV model that was used to produce OCO-2's Version 7. Then a multivariate EIV model is defined with $p \geq 1$ covariates, a special case of which is $p = 1$. In Section 3, we elaborate our parameter-estimation procedure for the multivariate EIV model, including a discussion of parameter identifiability. In Section 4, our proposed methodology is illustrated through simulation studies. In Section 5, we re-consider the Version 7 calculation between the OCO-2 data and the TCCON data; we use our proposed EIV model in both univariate-regression and multiple-regression settings to re-do the calculations. Conclusions follow in Section 6, and a technical Appendix and Supplementary Material complete the paper.

2. STATISTICAL MODELING FOR VALIDATION

Validation of remote sensing data is used to motivate the EIV regression models presented in this section. Here, the OCO-2 measurements of XCO_2 are regressed on the ground-based measurements of XCO_2 from TCCON, to determine an overall slope between them. The XCO_2 values from TCCON are often treated as the "true" values, but our model recognizes their uncertainties.

2.1 The Current Errors-in-Variables Model for Validating the OCO-2 Data

Let $\{(X_i, Y_i) : i = 1, \ldots, N\}$ be N pairs of remote sensing data and their coincident ground-based data (where "coincident" is defined through both spatial and temporal criteria; see Section 5). Version 7 of the OCO-2 retrieved data products used an errors-in-variables model for adjusting the global biases in the OCO-2 data (Mandrake et al., 2015), although it was not called that. The model used for this calibration was:

$$X_i = x_i + \epsilon_{x,i}, \quad Y_i = y_i + \epsilon_{y,i}, \quad y_i = a + bx_i, \tag{4}$$

where for $i = 1, \ldots, N$, $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are mutually independent measurement errors with mean zero and variances $\sigma^2_{x,i}$ and $\sigma^2_{y,i}$, respectively. In (4), x_i is a fixed but unknown mean parameter; Fuller (1987) referred to this as a functional model. (In contrast, if $\{x_i\}$ are iid random variables, Fuller (1987) referred to the model (4) as a structural model.)

For the functional model, $\{x_i\}$ are parameters but their estimation is not of primary interest. In York (1968), it was proposed to estimate a and b through minimizing a sum-of-weighted-squares criterion:

$$S(a, b) = \sum_{i=1}^{N} \frac{w_{x,i} w_{y,i}}{b^2 w_{y,i} + w_{x,i}} (Y_i - a - bX_i)^2, \tag{5}$$

where $\{w_{x,i}\}$ and $\{w_{y,i}\}$ are pre-specified regression weights. This was the approach taken in obtaining the Version-7 slope, with zero intercept, in a regression relating OCO-2 and TCCON XCO_2 measurements; see Section 5 for an explanation of the regression weights that were used and a justification for prespecifying a to be zero.

To ensure consistent estimation of regression coefficients a and b, the weights $\{w_{x,i}\}$ and $\{w_{y,i}\}$ should be chosen as the reciprocals of the variances of $\{\epsilon_{x,i}\}$ and $\{\epsilon_{y,i}\}$, respectively (e.g., Carroll and Ruppert, 1996; Zhang et al., 2017). It can be shown that each summation term in $S(a, b)$ is the weighted perpendicular distance from (X_i, Y_i) to the regression line. Thus, the least-sum-of-weighted-squares (LWS) estimators of a and b based on minimizing (5) have the interpretation of minimizing the sum of weighted perpendicular distances from $\{(X_i, Y_i)\}_{i=1}^N$ to the regression line. The LWS estimators have also been called generalized least squares (GLS) estimators (e.g., Sprent, 1966; Cheng and Riu, 2006). Further, Titterington and Halliday (1979) and Cheng and Riu (2006) showed that LWS estimators are also maximum (profile) likelihood estimators, provided that $\{\epsilon_{x,i}\}$ and $\{\epsilon_{y,i}\}$ are mean-zero Gaussian random variables and the weights are specified as the reciprocals of their respective variances.

The score equations for a and b based on (5) are unbiased when $w_{x,i}/w_{y,i} = \sigma^2_{y,i}/\sigma^2_{x,i}$, for $i = 1, \ldots, N$ (e.g., Zhang et al., 2017). Unbiasedness of estimating equations is a desirable property, since unbiased estimating equations lead to consistent parameter estimators under regularity conditions (Godambe, 1960; Yi and Reid, 2010). Asymptotic unbiasedness can be achieved by substituting consistent estimators for the theoretical variances of X_i and Y_i, for $i = 1, \ldots, N$. Since $\{X_i\}$ and $\{Y_i\}$ are typically aggregated data obtained from datasets of individual observations (in our case, from the TCCON time series and the OCO-2 spatial observations, respectively), Zhang et al. (2017) proposed to estimate error variances of the TCCON ($\{X_i\}$) and OCO-2 ($\{Y_i\}$) regression data from the corresponding datasets of individual observations. Importantly, account was taken of the temporal dependence between the individual TCCON observations and of the spatial dependence between the individual OCO-2 observations.

In this paper, we extend the OCO-2 validation model given by (4) to the validation model given by (2) and (3), where the error structure in (3) has additional random-effect components in each of X_i and Y_i, for $i = 1, \ldots, N$. The random-effect components capture the systematic error that may be present in TCCON and OCO-2 data: Systematic errors have variances that remain the same as sample size increases. In the proposed statistical model (2) and (3), these systematic errors are added to random errors (whose variances are inversely proportional to sample size).

Incorporating additional variance components can help fully capture the variance of the term $(Y_i - a - bX_i)$ in (5). It provides a way to estimate consistently and efficiently the regression coefficients a and b with the use of only a few extra parameters that account for often-present systematic errors. In the next subsection, we generalize this proposed model and new methodology from the univariate regression setting to the multiple-regression setting.

2.2 The Proposed Multivariate Errors-in-Variables Model

Here we generalize the univariate EIV model to the multivariate setting where the calibration equation has $p \geq 1$ covariates. For the i-th regression point, let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})^T$ be a p-dimensional covariate vector coincident with the response Y_i. Recall the model described by equations (2) and (3), which is generalized here to include more than one covariate. The multivariate EIV model considered in this paper is: For $i = 1, \ldots, N$,

$$\begin{aligned} \mathbf{X}_i &= \mathbf{x}_i + \boldsymbol{\eta}_{x,i} + \boldsymbol{\epsilon}_{x,i}, \\ Y_i &= y_i + \eta_{y,i} + \epsilon_{y,i}, \\ y_i &= a + \mathbf{b}^T \mathbf{x}_i, \end{aligned} \tag{6}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is a fixed but unknown vector of mean parameters; a and $\mathbf{b}$ are unknown regression coefficients to be estimated; $\boldsymbol{\eta}_{x,i} \equiv (\eta_{x,i,1}, \ldots, \eta_{x,i,p})^T$ is a vector of systematic errors (i.e., random effects) in $\mathbf{X}_i$ with mean $E(\boldsymbol{\eta}_{x,i}) = \mathbf{0}_{p\times 1}$ and covariance matrix $\text{var}(\boldsymbol{\eta}_{x,i}) \equiv T_x = \text{diag}(\tau^2_{x,1}, \ldots, \tau^2_{x,p})$; $\eta_{y,i}$ is the systematic error (i.e., random effect) in Y_i with mean $E(\eta_{y,i}) = 0$ and variance τ^2_y; $\boldsymbol{\epsilon}_{x,i} = (\epsilon_{x,i,1}, \ldots, \epsilon_{x,i,p})^T$ is the vector of random errors in $\mathbf{X}_i$ with mean $E(\boldsymbol{\epsilon}_{x,i}) = \mathbf{0}_{p\times 1}$ and covariance $\Sigma_{\epsilon,x,i}$; and $\epsilon_{y,i}$ is the random error in Y_i with mean $E(\epsilon_{y,i}) = 0$ and variance $\sigma^2_{\epsilon,y,i}$. We further assume that all the errors, $\{\boldsymbol{\eta}_{x,i}\}$, $\{\eta_{y,i}\}$, $\{\boldsymbol{\epsilon}_{x,i}\}$, and $\{\epsilon_{y,i}\}$, are mutually independent. Initially, it may appear that the model is over-parameterized, but we shall see below that this is not the case when regression data are aggregated from individual observations and the individual observations are available.

The systematic-error component, $\{\eta_{y,i}\}$, can also be interpreted as the equation-error term in the functional linear model (e.g., Fuller, 1987; Carroll and Ruppert, 1996), which can be used to model effects of unaccounted-for covariates. It is well known that the equation error can dramatically affect the parameter estimation of regression coefficients (Carroll and Ruppert, 1996; Cheng and Riu, 2006), and ignoring it could typically lead to over-calibration of the Y-values (Carroll and Ruppert, 1996). It is also well known that covariates not observed precisely and not accounted for in an EIV regression model can bias the parameter estimation of the regression coefficients (Fuller, 1987; Carroll et al., 2006) and hence the corrected Y-values.

Let $\Sigma_{x,i} = \Sigma_{\epsilon,x,i} + T_x$ be the covariance matrix of $\mathbf{X}_i$, and denote all the parameters (except the means $\{\mathbf{x}_i\}$) by

$$\boldsymbol{\theta} \equiv \{a, \mathbf{b}, \Sigma_{\epsilon,x,i}, \sigma^2_{\epsilon,y,i}, T_x, \tau^2_y\}.$$

When all errors in (6) are Gaussian, the joint log-likelihood of $\boldsymbol{\theta}$ and $\{\mathbf{x}_i\}$ is given by

$$\begin{aligned}\ell(\{\mathbf{x}_i\}, \boldsymbol{\theta}) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{X}_i - \mathbf{x}_i)^T \Sigma_{x,i}^{-1}(\mathbf{X}_i - \mathbf{x}_i) - \frac{1}{2}\sum_{i=1}^{N}\frac{(Y_i - a - \mathbf{b}^T\mathbf{x}_i)^2}{\sigma^2_{\epsilon,y,i} + \tau^2_y} \\ & -\frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_{x,i}| - \frac{1}{2}\sum_{i=1}^{N}\log(\sigma^2_{\epsilon,y,i} + \tau^2_y) + \text{constant}. \end{aligned} \tag{7}$$

Since $\{\mathbf{x}_i\}$ are nuisance parameters, we substitute their maximum likelihood estimators (MLEs) into (7) and obtain a profile log-likelihood given by (see Appendix A),

$$\begin{aligned}\ell(\boldsymbol{\theta}) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}\frac{(Y_i - a - \mathbf{b}^T\mathbf{X}_i)^2}{\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau^2_y} - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_{x,i}| \\ & -\frac{1}{2}\sum_{i=1}^{N}\log(\sigma^2_{\epsilon,y,i} + \tau^2_y) + \text{constant}. \end{aligned} \tag{8}$$

As we have already remarked, estimation of $\boldsymbol{\theta}$ based on only the regression data $\{(\mathbf{X}_i, Y_i)\}$ in (8) is problematic. First, the random errors have heterogeneous variances, which has been widely discussed in many contexts (e.g., Cheng and Riu, 2006; Riu and Rius, 1995; Cheng and Tsai, 2015). Nevertheless, estimation of their variances is possible when each regression datum is in fact an average, or similar aggregation, of individual observations and the individual observations are available. Then, the square of the sample standard error is a natural choice for estimating the random-error variances, assuming that the individual observations are iid. However, when correlation is present, the square of the sample standard error is a biased estimator of the random-error variances. To account for correlations between individual observations, we propose to use spatial/temporal-process modeling, where the correlations and variances are characterized by a spatial/temporal covariance function; see Zhang et al. (2017).

Second, estimation of regression coefficients jointly with the systematic-error variance, τ_y^2, is not trivial, even for functional EIV models that assume $\mathbf{x}_i$ is a fixed effect without systematic errors. (It is more straightforward for structural EIV models; see Kulathinal et al., 2002; Cheng and Riu, 2006). In Section 3, we show how to estimate the regression coefficients jointly with τ_y^2 (which can also be seen as the equation-error variance) from adjusted estimating equations.

Third, the model (6) has parameter-identifiability issues, and in Section 3 we show how a two-stage parameter-estimation approach resolves these issues.

2.3 The Single Covariate ($p = 1$) Case

Let us return to the EIV model given in Section 2.1 and augment it to account for systematic errors through extra random effects. The case of $p = 1$ in Section 2.2 (single covariate) is immediately applicable to the validation of remote sensing data from accurate ground-based measurements. We now write the model for this special case in detail. For $i = 1, \ldots, N$,

$$X_i = x_i + \eta_{x,i} + \epsilon_{x,i}, \quad Y_i = y_i + \eta_{y,i} + \epsilon_{y,i}, \quad y_i = a + bx_i, \tag{9}$$

where $\eta_{x,i}$ and $\eta_{y,i}$ are mean-zero systematic errors with variances τ_x^2 and τ_y^2, respectively; $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are mean-zero random errors with variances $\sigma^2_{\epsilon,x,i}$ and $\sigma^2_{\epsilon,y,i}$, respectively; and all errors $\{\eta_{x,i}\}$, $\{\eta_{y,i}\}$, $\{\epsilon_{x,i}\}$, and $\{\epsilon_{y,i}\}$ are mutually independent.

It is straightforward to show that the profile log-likelihood function is:

$$\begin{aligned}\ell(\boldsymbol{\theta}) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}\frac{(Y_i - a - bX_i)^2}{b^2(\sigma^2_{\epsilon,x,i} + \tau_x^2) + \sigma^2_{\epsilon,y,i} + \tau_y^2} - \frac{1}{2}\sum_{i=1}^{N}\log(\sigma^2_{\epsilon,x,i} + \tau_x^2) \\ & -\frac{1}{2}\sum_{i=1}^{N}\log(\sigma^2_{\epsilon,y,i} + \tau_y^2) + \text{constant},\end{aligned} \tag{10}$$

where the fixed but unknown mean parameters $\{x_i\}$ are "profiled out" of the log-likelihood (Titterington and Halliday, 1979). This single-covariate model will be applied to validation of the XCO_2 data from OCO-2 in Section 5.

3. PARAMETER ESTIMATION

The multivariate EIV model in (6) has an important application to validating remote sensing data, provided that the parameters $\boldsymbol{\theta}$ can be estimated. While the regression coefficients a and $\mathbf{b}$ are of primary importance, it is necessary to provide a practical solution for estimating variance-component parameters, since they will be substituted into the estimating equations of the regression coefficients.

3.1 Parameter Identifiability

We discuss here parameter identifiability for the proposed model given by equation (6). First, based on data $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$ and because of the additivity of random effects, the random-error variances, $\{\mathit{\Sigma}_{\epsilon,x,i}\}$ and $\{\sigma^2_{\epsilon,y,i}\}$ are not identifiable in the presence of the respective unknown systematic-error variances, T_x and τ_y^2. We resolve this identifiability issue in situations where, behind each $X_{i,1}, \ldots, X_{i,p}$ and Y_i, there are sets of individual observations, $\mathcal{D}_{x,i,j} \equiv \{\tilde{X}_{i,j,k} : k = 1, \ldots, n_{x,i,j}\}$ for $j = 1, \ldots, p$, and $\mathcal{D}_{y,i} \equiv \{\tilde{Y}_{i,\ell} : \ell = 1, \ldots, n_{y,i}\}$, respectively, that are available. This is the case for the remote sensing data analysis in Section 5. Section 3.2 gives details about the estimation of $\{\mathit{\Sigma}_{\epsilon,x,i}\}$ and $\{\sigma^2_{\epsilon,y,i}\}$ from datasets $\{\mathcal{D}_{x,i,j}\}$ and $\{\mathcal{D}_{y,i}\}$ of individual observations.

For the moment, we assume that $\{\mathit{\Sigma}_{\epsilon,x,i}\}$ and $\{\sigma^2_{\epsilon,y,i}\}$ are known, and we move onto estimating the remaining parameters from the log-likelihood (8). Consider $\ell(\boldsymbol{\theta})$ as a function of τ_y^2 and $\{\tau^2_{x,j}\}$. The score equation for τ_y^2 is:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \tau_y^2} = \frac{1}{2} \sum_{i=1}^{N} \frac{(Y_i - a - \mathbf{b}^T \mathbf{X}_i)^2}{(\mathbf{b}^T \mathit{\Sigma}_{x,i} \mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2)^2} - \frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma^2_{\epsilon,y,i} + \tau_y^2} = 0, \tag{11}$$

where recall that for $i = 1, \ldots, N$, $\mathit{\Sigma}_{x,i} = \mathit{\Sigma}_{\epsilon,x,i} + T_x$, and $T_x = \mathrm{diag}(\tau^2_{x,1}, \ldots, \tau^2_{x,p})$. Also, for

$j = 1, \ldots, p$, the score equations for $\{\tau_{x,j}^2\}$ are

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \tau_{x,j}^2} &= \frac{1}{2}\sum_{i=1}^{N} \frac{(Y_i - a - \mathbf{b}^T\mathbf{X}_i)^2}{(\mathbf{b}^T \Sigma_{x,i}\mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2)^2} \cdot (\mathbf{b}^T E_{j,j}\mathbf{b}) - \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}(\Sigma_{x,i}^{-1} E_{j,j}) \\ &= \frac{1}{2}\sum_{i=1}^{N} \frac{(Y_i - a - \mathbf{b}^T\mathbf{X}_i)^2 b_j^2}{(\mathbf{b}^T \Sigma_{x,i}\mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2)^2} - \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}(\Sigma_{x,i}^{-1} E_{j,j}) = 0, \end{aligned} \quad (12)$$

where $E_{j,j}$ is a $p \times p$ matrix with (j, j)-th entry equal to 1 and all other entries equal to 0.

Upon substituting (11) into (12), for $b_j \neq 0$, we obtain:

$$b_j^2 = \left\{ \sum_{i=1}^{N} \mathrm{tr}(\Sigma_{x,i}^{-1} E_{j,j}) \right\} \Big/ \left\{ \sum_{i=1}^{N} (\sigma_{\epsilon,y,i}^2 + \tau_y^2)^{-1} \right\}. \quad (13)$$

Since b_j given by (13) is not a solution of $\partial \ell(\boldsymbol{\theta})/\partial b_j = 0$, the parameters b_j, $\tau_{x,j}^2$, and τ_y^2 are not jointly estimable from the score equations of the profile log-likelihood. However, there is a way through this by using both the regression data $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$ and the individual observations $\{\mathcal{D}_{x,i,1}\}, \ldots, \{\mathcal{D}_{x,i,p}\}$, and $\{\mathcal{D}_{y,i}\}$ (whose average, or similar aggregation, results in $\{\mathbf{X}_i\}$ and $\{Y_i\}$, respectively).

Estimation of the systematic-error variances $\{\tau_{x,j}^2 : j = 1, \ldots, p\}$ requires additional validation datasets that are coincident with $\{\mathbf{X}_i\}$ and do not contain systematic errors. In the OCO-2 remote sensing application, this would involve measurements of CO_2 from aircraft in the atmosphere above TCCON sites. Assuming that the systematic-error variances $\{\tau_{x,j}^2\}$ are known, we now focus on estimating the remaining model parameters.

We propose a two-stage estimation procedure: In the first stage (Section 3.2), the variances of random errors in $\mathbf{X}_i$ and Y_i are estimated from datasets of individual observations. In the second stage (Section 3.3), regression coefficients a and $\mathbf{b}$ and the systematic-error variance τ_y^2 are estimated from the regression data $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$. At each estimation stage, we substitute parameter estimates obtained from the previous stage into the estimating equations as if they were known. Thus, estimation at the second stage can be seen as pseudo maximum likelihood estimation (Gong and Samaniego, 1981).

3.2 Stage 1: Estimation of the Random-Error Variances in $\mathbf{X}_i$ and Y_i

In practice, $\mathbf{X}_i$ and Y_i are often averages, or similar aggregations, obtained from datasets of individual observations, $\mathcal{D}_{x,i,j}$ and $\mathcal{D}_{y,i}$. For notational simplicity, in this section we omit the subscript i for individual observations, $\{\tilde{X}_{i,j,k}\}$ and $\{\tilde{Y}_{i,\ell}\}$. It should be understood that in the methodology that follows, the regression point i is fixed at a given value in $\{1, \ldots, N\}$.

Since for a given regression point individual observations are usually generated under homogeneous atmospheric conditions, we make constant-mean and homogeneous-variance assumptions and model them as follows: For $j = 1, \ldots, p$,

$$\tilde{X}_{j,k} = x_j + \eta_{x,j} + \tilde{\epsilon}_{x,j,k}, \quad \tilde{Y}_\ell = y + \eta_y + \tilde{\epsilon}_{y,\ell}, \tag{14}$$

where $\tilde{\epsilon}_{x,j,k}$ and $\tilde{\epsilon}_{y,\ell}$ are mean-zero random errors with homogeneous variances $\tilde{\sigma}^2_{x,j}$ and $\tilde{\sigma}^2_y$, respectively, for $k = 1, \ldots, n_{x,j}$ and $\ell = 1, \ldots, n_y$; and $\eta_{x,j}$ and η_y are mean-zero systematic errors with variances $\tau^2_{x,j}$ and τ^2_y, respectively. Recall that the means $\mathbf{x} \equiv (x_1, \ldots, x_p)^T$ and y are fixed but unknown and related by $y = a + \mathbf{b}^T\mathbf{x}$.

If individual observations are independent, the sample variance provides a consistent estimator for $\tilde{\sigma}^2_{x,j}$ and $\tilde{\sigma}^2_y$. Furthermore, if the regression data are simple averages, $X_j = \sum_{k=1}^{n_{x,j}} \tilde{X}_{j,k}/n_{x,j}$, for $j = 1, \ldots, p$, and $Y = \sum_{\ell=1}^{n_y} \tilde{Y}_\ell/n_y$, then under independence the variances of the random errors in X_j and Y are $\sigma^2_{\epsilon,x,j} = \tilde{\sigma}^2_{x,j}/n_{x,j}$ and $\sigma^2_{\epsilon,y} = \tilde{\sigma}^2_y/n_y$, respectively. However, correlations between observations may be nonnegligible, and hence they need to be accounted for when estimating $\tilde{\sigma}^2_{x,j}$ and $\tilde{\sigma}^2_y$. In the OCO-2 validation from TCCON, Zhang et al. (2017) discussed how to apply spatial/temporal-process modeling of the individual OCO-2 observations (observed by the satellite within a small geographic region within 5 minutes of each other) and the individual TCCON observations (observed at point locations from the ground over longer periods of time), where dependence was modeled through spatial/temporal covariance functions. For example, the Matérn covariance function (Stein, 1999; Matérn, 2013) can be used in both space and time to model variances and correlations:

$$\mathcal{C}(h; \boldsymbol{\psi}) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (h/\phi)^\nu \mathcal{K}_\nu (h/\phi), \tag{15}$$

where $h \geq 0$ is a spatial/temporal distance between two individual observations. In (15), $\boldsymbol{\psi} \equiv \{\sigma^2, \phi, \nu\}$, where $\sigma^2 > 0$ is the variance parameter, $\phi > 0$ is the range parameter modeling how correlations decay with increasing h, and $\nu > 0$ is the smoothness parameter; $\Gamma(\cdot)$ is the gamma function; and $\mathcal{K}_\nu(\cdot)$ is a modified Bessel function of the second kind of order ν.

When individual observations are Gaussian and free of outliers, Restricted Maximum Likelihood (REML) estimation (Patterson and Thompson, 1971; Harville, 1977) can be used to estimate covariance-function parameters (e.g., $\boldsymbol{\psi}$ in (15)). In the case of OCO-2 validation, this is reasonable for TCCON, but individual OCO-2 observations appear to contain outliers that require its covariance function to be estimated from semiparametric robust methodology (e.g., based on robust variogram estimators in Cressie and Hawkins, 1980). We remark that it is a property of both REML estimation and semiparametric estimation based on variograms that the systematic-error terms in (14) are not present in the objective function to be optimized, and hence the systematic-error variances will not affect the estimation of the random-error variances.

After we obtain the estimated covariance-function parameters, $\hat{\boldsymbol{\psi}}_{x,j} \equiv \{\hat{\tilde{\sigma}}^2_{x,j}, \hat{\phi}_{x,j}, \hat{\nu}_{x,j}\}$ and $\hat{\boldsymbol{\psi}}_y \equiv \{\hat{\tilde{\sigma}}^2_y, \hat{\phi}_y, \hat{\nu}_y\}$, random-error variances in $X_{i,j}$ and Y_i are estimated by

$$\hat{\sigma}^2_{\epsilon,x,j} \equiv \hat{\tilde{\sigma}}^2_{x,j}/\tilde{n}_{x,j}; j = 1, \ldots, p, \quad \hat{\sigma}^2_{\epsilon,y} \equiv \hat{\tilde{\sigma}}^2_y/\tilde{n}_y,$$

where $\tilde{n}_{x,j}$ and $\tilde{n}_y$ are effective sample sizes, respectively; see Zhang et al. (2017). Positive correlations between individual observations lead to an effective sample size smaller than the actual sample size. When X_j and Y are the sample means or the sample medians of individual observations $\{\tilde{X}_{j,k}\}$ and $\{\tilde{Y}_\ell\}$, respectively, $\mathrm{var}(X_j)$ and $\mathrm{var}(Y)$ can be estimated under dependence within the sets of individual observations (Zhang et al., 2017, Sections 4 and 5).

3.3 Stage 2: Unbiased Estimation of Regression Coefficients a and $\mathbf{b}$ and the Systematic-Error Variance τ_y^2

Here, we re-introduce the subscript i indicating the i-th regression point. At this stage, parameters from Stage 1 have been estimated which, for the purposes of the analysis to follow, we assume are now fixed and known (Gong and Samaniego, 1981). We first calculate the score equations from the log-likelihood equation (8) and determine if they are unbiased. Since the other parameters are assumed fixed, with a slight abuse of notation, let $\boldsymbol{\theta} = (a, \mathbf{b}^T, \tau_y^2)^T$. Then the estimating equations from the score functions of $\boldsymbol{\theta}$ are:

$$U_a(\boldsymbol{\theta}) \equiv \partial\ell(\boldsymbol{\theta})/\partial a = 0, \;\; \mathbf{U_b}(\boldsymbol{\theta}) \equiv \partial\ell(\boldsymbol{\theta})/\partial \mathbf{b} = \mathbf{0}, \;\text{ and } U_{\tau_y^2}(\boldsymbol{\theta}) \equiv \partial\ell(\boldsymbol{\theta})/\partial \tau_y^2, \tag{16}$$

which are unbiased if we can show that $E(U_a(\boldsymbol{\theta})) = 0$, $E(\mathbf{U_b}(\boldsymbol{\theta})) = \mathbf{0}$, and $E(U_{\tau_y^2}(\boldsymbol{\theta})) = 0$, for all $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^p$ and $\tau_y^2 \geq 0$. As discussed above, unbiasedness of estimating equations is a desirable property, since it results in consistent parameter estimators under regularity conditions given, for example, in Godambe (1960).

Recall that the regression data are $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$. According to assumptions of the model in equation (6), we have $E(Y_i^2) = (a+\mathbf{b}^T\mathbf{x}_i)^2 + \sigma_{\epsilon,y,i}^2 + \tau_y^2$, $E(\mathbf{X}_i\mathbf{X}_i^T) = \mathbf{x}_i\mathbf{x}_i^T + \varSigma_{x,i}$, and $E(\mathbf{X}_iY_i) = \mathbf{x}_i(a + \mathbf{x}_i^T\mathbf{b})$, for $i = 1, \ldots, N$. Then from (8) and (16),

$$E(U_a(\boldsymbol{\theta})) \;=\; \sum_{i=1}^{N} \frac{1}{\omega_i} E(Y_i - a - \mathbf{X}_i^T\mathbf{b}) = 0,$$

and

$$\begin{aligned} E(\mathbf{U_b}(\boldsymbol{\theta})) &= \sum_{i=1}^{N} \frac{1}{\omega_i^2} E\left((Y_i - a - \mathbf{X}_i^T\mathbf{b})\mathbf{X}_i\omega_i + (Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 \varSigma_{x,i}\mathbf{b}\right) \\ &= \sum_{i=1}^{N} \frac{1}{\omega_i^2} (-\varSigma_{x,i}\mathbf{b}\omega_i + \varSigma_{x,i}\mathbf{b}\omega_i) = \mathbf{0}, \end{aligned}$$

where $\omega_i \equiv \mathbf{b}^T\varSigma_{x,i}\mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2 = \mathrm{var}(Y_i - a - \mathbf{X}_i^T\mathbf{b})$. That is, the estimating equations for

a and $\mathbf{b}$ are unbiased. However,

$$\begin{aligned} E(U_{\tau_y^2}(\boldsymbol{\theta})) &= \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\omega_i^2}E(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_{\epsilon,y,i}^2 + \tau_y^2} \\ &= \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2} - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_{\epsilon,y,i}^2 + \tau_y^2} < 0, \end{aligned}$$

and hence the estimating equation, $U_{\tau_y^2}(\boldsymbol{\theta}) = 0$, is biased and requires modification.

Bias-correction of score functions of the profile log-likelihood has been studied by McCullagh and Tibshirani (1990), who proposed a simple method for adjusting the profile log-likelihood so that its score function has mean zero and variance equal to the negative expected Hessian matrix. Their method results in modified score functions that are both unbiased and information unbiased. Motivated by their approach, we apply a mean adjustment and obtain the following unbiased estimating function for τ_y^2:

$$\begin{aligned} \tilde{U}_{\tau_y^2}(\boldsymbol{\theta}) &= U_{\tau_y^2}(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}}(U_{\tau_y^2}(\boldsymbol{\theta})) \\ &= \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\omega_i^2}E(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\omega_i}. \end{aligned}$$

Other approaches are possible, such as in Yi and Reid (2010), where the roots of the biased estimating equations are transformed.

Let $\mathbf{U}(\boldsymbol{\theta}) \equiv (U_a(\boldsymbol{\theta}), \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})^T, \tilde{U}_{\tau_y^2}(\boldsymbol{\theta}))^T$ denote the unbiased-estimating-function vector and $V(\boldsymbol{\theta}) \equiv \partial\mathbf{U}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ denote its Hessian matrix. Then under regularity conditions (Godambe, 1960), the root of the estimating equation, $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$, denoted by $\hat{\boldsymbol{\theta}}$, is consistent and asymptotically Gaussian, with asymptotic covariance matrix:

$$E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta}))^{-1}E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})^T)^{-1}, \tag{17}$$

which is the inverse of the Godambe information matrix (Godambe, 1960). This asymptotic covariance matrix is calculated in Appendix B.

The Fisher-scoring algorithm can be used to estimate $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - V(\boldsymbol{\theta}^{(k)})^{-1}\mathbf{U}(\boldsymbol{\theta}^{(k)}).$$

In the simulations of Section 4 and the remote sensing analysis in Section 5, convergence of the Fisher-scoring algorithm typically occurs in about six iterations, with initial values chosen as follows: For the regression coefficients a and $\mathbf{b}$, we use ordinary-least-squares estimates that are given following (1). For τ_y^2, we use the Method-of-Moments estimate,

$$\max\left\{\frac{1}{N}\sum_{i=1}^{N}((Y_i - \hat{a} - \mathbf{X}_i^T\hat{\mathbf{b}})^2 - \hat{\mathbf{b}}^T\hat{\Sigma}_{x,i}\hat{\mathbf{b}} - \hat{\sigma}^2_{\epsilon,y,i}), 0\right\},$$

where $\hat{\sigma}^2_{\epsilon,y,i}$ and $\hat{\Sigma}_{x,i} \equiv \hat{\Sigma}_{\epsilon,x,i} + T_x$ are obtained from Stage 1 (Section 3.2).

4. SIMULATION STUDIES

In this section, we consider multiple regression with two covariates (i.e., $p = 2$) and show by simulation that ignoring the nonzero systematic errors in $\{Y_i\}$ will result in biased estimates of regression coefficients a and $\mathbf{b}$. Here, we have chosen representative studies and have not implemented a full simulation experiment.

The simulation requires specification of mean parameters $\{x_{i,1} : i = 1, \ldots, N\}$ and $\{x_{i,2} : i = 1, \ldots, N\}$, and we chose values scattered across the domains $[3, 16]$ for $\{x_{i,1}\}$ and $[2, 8]$ for $\{x_{i,2}\}$. To avoid any prejudice in their choice, we generated the N values uniformly on their respective domains and, once generated, they remained fixed in the simulation. To see the effect of N, the sample sizes, $N = 150$ and $N = 600$, were chosen. In what follows, the simulation was repeated $L = 500$ times, so that bias and variance of parameter estimates could be ascertained. The observed covariate vector, $\mathbf{X}_i = (X_{i,1}, X_{i,2})^T$, has errors and, for $i = 1, \ldots, N$, they were randomly generated from a Gaussian distribution with mean $\mathbf{x}_i = (x_{i,1}, x_{i,2})^T$ and covariance matrix $\Sigma_{x,i} = \Sigma_{\epsilon,x,i} + T_x$; recall that $\Sigma_{\epsilon,x,i}$ is the random-error covariance matrix and T_x is the systematic-error covariance matrix. Heterogeneous random-error covariance matrices $\{\Sigma_{\epsilon,x,i}\}$ over $i = 1, \ldots, N$ were specified as follows: $\sigma^2_{\epsilon,x,i,1} \equiv \Sigma_{\epsilon,x,i}(1,1) = (0.1x_{i,1})^2$, $\sigma^2_{\epsilon,x,i,2} \equiv \Sigma_{\epsilon,x,i}(2,2) = (0.1x_{i,2})^2$, and $\Sigma_{\epsilon,x,i}(1,2) = \Sigma_{\epsilon,x,i}(2,1) =$

$\rho\sqrt{\sigma^2_{\epsilon,x,i,1}\sigma^2_{\epsilon,x,i,2}}$, with $\rho = 0.5$. The systematic-error covariance matrix was specified as $T_x = \mathrm{diag}(0.5, 0.5)$. The responses $\{Y_i\}_{i=1}^N$ were then randomly generated from $\mathcal{N}(y_i, \sigma^2_{\epsilon,y,i} + \tau^2_y)$, where $y_i = a + \mathbf{x}_i^T\mathbf{b}$, and the following values were specified: $a = 1$, $\mathbf{b} = (b_1, b_2)^T = (0.5, 1)^T$, heterogeneous $\sigma^2_{\epsilon,y,i} = (0.25y_i)^2$ over $i = 1, \dots, N$, and $\tau^2_y = 2$.

Since $\frac{1}{N}\sum_{i=1}^N(y_i^2/(\sigma^2_{\epsilon,y,i} + \tau^2_y))$ can be interpreted as a signal-to-noise ratio (SNR) that can affect the estimation of regression parameters, we chose different SNR levels in our simulation study. For the EIV-model parameters specified above, the SNR is about 12, corresponding to a relatively high-SNR scenario (denoted by "HI"). We also chose a low-SNR scenario (denoted by "LO") with $a = 1/3$, $\mathbf{b} = (b_1, b_2)^T = (1/6, 1/3)^T$, $\sigma^2_{\epsilon,y,i} = (0.75y_i)^2$, and other model parameters the same as those for the HI scenario; for the LO scenario, the SNR value is about 1.3.

We focus here on Stage 2 of the parameter estimation, arguably the most important part where regression coefficients a and $\mathbf{b}$ are estimated along with the systematic-error variance τ^2_y, with other parameters fixed at their respective values specified above. For each scenario, we simulated 500 realizations of $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$ and compared different parameter estimates to the true values of a, $\mathbf{b}$, and τ^2_y. Estimates of a, $\mathbf{b}$, and τ^2_y from the unbiased estimating equations (denoted by "UEE") are compared to the "gold standard," namely estimates of a and $\mathbf{b}$ when the true systematic-error variance $\tau^2_y = 2$ is used (denoted by "TRU"). The misspecified value of $\tau^2_y = 0$ yields estimates of a and $\mathbf{b}$ (denoted by "MSP") that are compared to UEE and TRU. Averages over the 500 simulations give Monte Carlo approximations to the estimates' means and standard errors.

For the HI scenario, Table 1 gives the parameter-estimation results under different specifications of τ^2_y. It can be seen that the proposed model with unbiased estimating equations (UEE) produces parameter estimates of a and $\mathbf{b}$ that are close to their respective true values, with comparable standard errors to those obtained by using the true value $\tau^2_y = 2$ (TRU); the estimates of τ^2_y have negative biases for a small sample size $N = 150$, and this remains the case although they are much closer to the true value, 2, for the larger sample size $N = 600$. In contrast, the case where the systematic errors in $\{Y_i\}$ are ignored (i.e., MSP) results in biased estimates of a and $\mathbf{b}$, especially for the intercept a (with significant negative biases),

Table 1: Parameter-estimation results for simulations with $p = 2$, where the true values of the parameters are $a = 1$, $b_1 = 0.5$, $b_2 = 1$, and $\tau_y^2 = 2$; that is, the HI signal-to-noise-ratio scenario. UEE: Regression coefficients are jointly estimated with τ_y^2 from unbiased estimating equations; TRU: Regression coefficients are estimated with τ_y^2 fixed at its true value of 2; MSP: Regression coefficients are estimated with τ_y^2 fixed at zero. The means of parameter estimates are reported, with standard errors in parentheses. Results were obtained based on $L = 500$ simulated regression datasets.

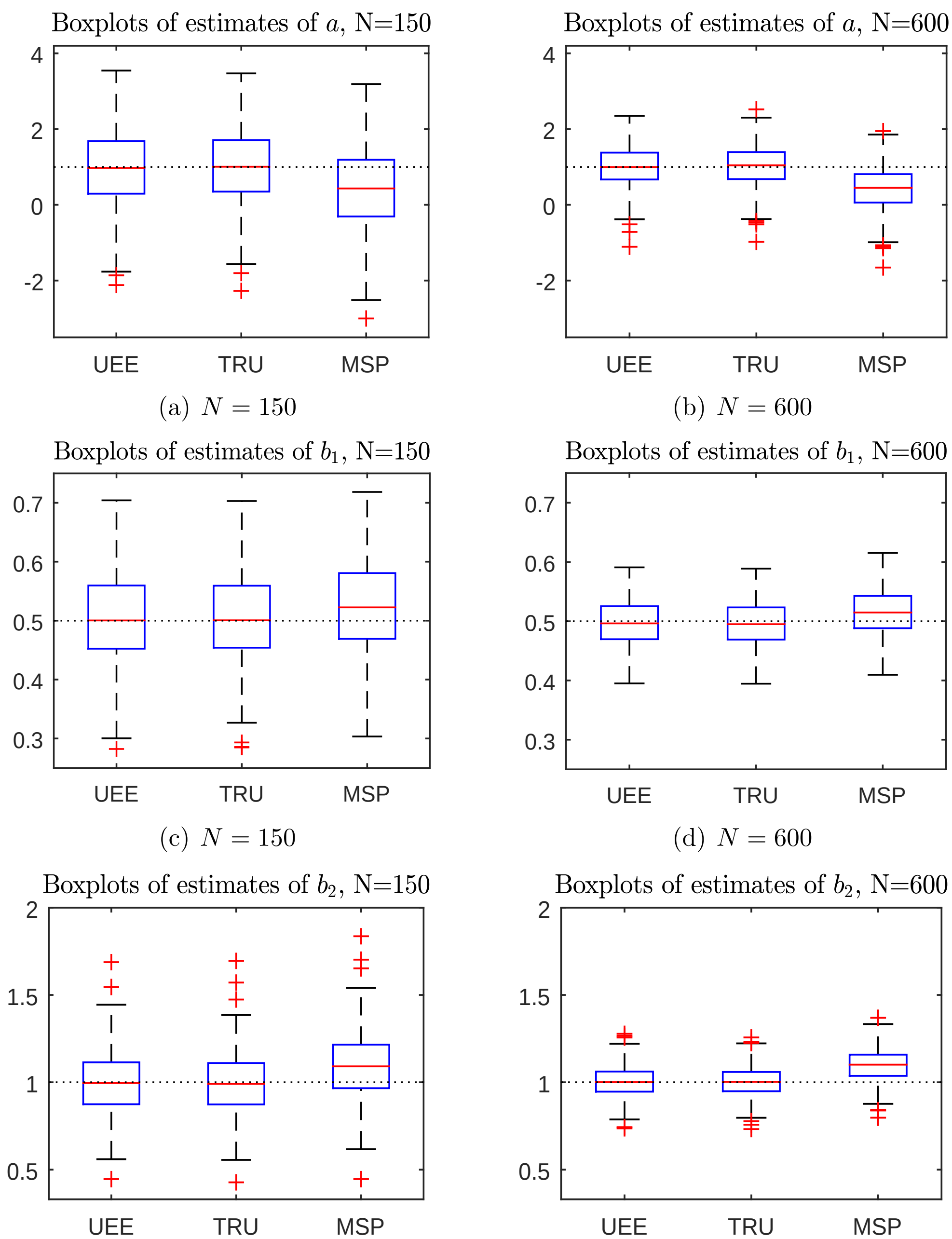
$N = 150$	$a : 1$	$b_1 : 0.5$	$b_2 : 1$	$\tau_y^2 : 2$
UEE	0.971 (0.046)	0.504 (0.003)	1.000 (0.008)	1.805 (0.043)
TRU	1.008 (0.045)	0.503 (0.003)	0.993 (0.007)	–
MSP	0.435 (0.048)	0.522 (0.003)	1.092 (0.008)	–
$N = 600$	$a : 1$	$b_1 : 0.5$	$b_2 : 1$	$\tau_y^2 : 2$
UEE	1.012 (0.024)	0.497 (0.002)	1.002 (0.004)	1.935 (0.024)
TRU	1.026 (0.024)	0.497 (0.002)	1.000 (0.004)	–
MSP	0.429 (0.025)	0.516 (0.002)	1.098 (0.004)	–

and the coefficients b_1 and b_2 have significant positive biases. Figure 1 shows boxplots of the regression-coefficient estimates for these three cases, which reinforce our conclusions from Table 1. In particular, under MSP, regression-coefficient estimates are clearly biased and so any validation based on them would be biased. Under UEE, regression-coefficient estimates show essentially no bias, and variances of the parameter estimates are comparable to those of TRU.

Table 2: Relative efficiencies (REs) of parameter estimates for the HI scenario. The results were obtained based on $L = 500$ simulated regression datasets.

$N = 150$	a	b_1	b_2
UEE	0.96	0.99	0.95
MSP	0.68	0.84	0.67
$N = 600$	a	b_1	b_2
UEE	0.93	0.98	0.95
MSP	0.44	0.82	0.40

Table 2 shows the Relative Efficiencies (RE) of parameter estimates, where an RE value is defined as the ratio of mean squared error of parameter estimates under TRU relative to that under UEE or that under MSP. Since TRU is the "gold standard," an RE value as large and as close to 1 as possible, is desirable. It is clear that MSP (i.e., the misspecified model



(a) $N = 150$ (b) $N = 600$

(c) $N = 150$ (d) $N = 600$

(e) $N = 150$ (f) $N = 600$

Figure 1: Boxplots of the regression-coefficient estimates from $L = 500$ simulated regression datasets under different specifications of τ_y^2, namely UEE, TRU, and MSP. Results are given for the HI scenario, namely $a = 1$, $b_1 = 0.5$, $b_2 = 1$, and $\tau_y^2 = 2$. The dotted lines indicate the true values of these regression coefficients.

with $\tau_y^2 = 0$) results in very poor RE values for the regression coefficients. In contrast, the RE values for UEE are very satisfactory, all greater than 0.9.

Table 3: Comparison of the empirical (from simulation) standard errors of parameter estimates (denoted by "EMP") with the corresponding asymptotic standard errors obtained from the Godambe information matrix (denoted by "ASY") for the case UEE and the HI scenario. The EMP results were obtained based on $L = 500$ simulated regression datasets.

$N = 150$	a	b_1	b_2	τ_y^2
EMP	1.0191	0.0738	0.1695	0.9647
ASY	1.0022	0.0724	0.1638	0.9981
$N = 600$	a	b_1	b_2	τ_y^2
EMP	0.5458	0.0384	0.0866	0.5376
ASY	0.5414	0.0365	0.0859	0.5277

Last, Table 3 shows the empirical standard errors of parameter estimates for UEE, in comparison with the asymptotic standard errors obtained based on the Godambe information matrix given by (17); the asymptotic standard errors of parameter estimates were obtained by substituting in true values of model parameters. Table 3 shows that when sample size $N = 150$, the asymptotic standard errors are comparable to the corresponding empirical ones for the regression coefficients, but the asymptotic standard error of τ_y^2 is a bit (about 4%) larger than its empirical counterpart. For the larger sample size, $N = 600$, the asymptotic standard errors match the empirical results quite well for all regression parameters, including τ_y^2.

Table 4: Parameter-estimation results for simulations with $p = 2$, where the true values of the parameters are $a = 1/3$, $b_1 = 1/6$, $b_2 = 1/3$, and $\tau_y^2 = 2$; that is, the LO signal-to-noise-ratio scenario. The means of parameter estimates are reported, with standard errors in parentheses. Results were obtained based on $L = 500$ simulated regression datasets.

$N = 150$	$a : 0.333$	$b_1 : 0.167$	$b_2 : 0.333$	$\tau_y^2 : 2$
UEE	0.302 (0.039)	0.167 (0.003)	0.341 (0.006)	1.806 (0.043)
TRU	0.314 (0.039)	0.166 (0.003)	0.339 (0.006)	–
MSP	0.116 (0.042)	0.172 (0.003)	0.373 (0.007)	–
$N = 600$	$a : 0.333$	$b_1 : 0.167$	$b_2 : 0.333$	$\tau_y^2 : 2$
UEE	0.315 (0.020)	0.166 (0.001)	0.337 (0.003)	2.012 (0.021)
TRU	0.316 (0.020)	0.166 (0.001)	0.337 (0.003)	–
MSP	0.075 (0.023)	0.173 (0.002)	0.380 (0.004)	–

For the LO scenario, Table 4 shows the parameter-estimation results under the three different specifications of τ_y^2 (namely, UEE, TRU, and MSP), and similar conclusions to those from Table 1 hold. Under both sample sizes, ignoring the systematic error in $\{Y_i\}$ (MSP) results in biased estimates of a and $(b_1, b_2)^T$, especially for the intercept a and the coefficient b_2. As before, negative biases are observed for the intercept a. Further, UEE still results in estimates of a and $(b_1, b_2)^T$ comparable with TRU, which are very close to the true parameter values of $a = 1/3$, $b_1 = 1/6$, and $b_2 = 1/3$.

5. APPLICATION TO VALIDATION OF THE OCO-2 DATA

In this section, we first apply the proposed model with $p = 1$ (Section 2.3) to validate OCO-2 satellite data (Mandrake et al., 2015; Crisp et al., 2017) using TCCON ground-based data (Wunch et al., 2011, 2017), where OCO-2 target-mode observations (Y) are regressed against coincident TCCON observations (X). The final step to obtain Version 7 of the OCO-2 retrieved data products involved fitting a straight line to $N = 66$ pairs $\{(X_i, Y_i) : i = 1, \ldots, 66\}$ (Mandrake et al., 2015). The OCO-2 regression datum Y_i for the i-th TCCON station/OCO-2 orbit combination was the sample median of a set of individual target-mode observations (typically centered on various TCCON stations), while the corresponding TCCON regression datum X_i was the sample mean of a set of individual TCCON observations in a 2-hour time window centered at the mean target time of the OCO-2 individual observations. The Version 7 validation involved fitting an EIV regression model (York et al., 2004), through the origin (i.e., $a = 0$), to $N = 66$ regression data, from which a calibration line, $Y = \hat{b}X$, was obtained. Then any retrieved OCO-2 datum, Y^0, was transformed to the Version 7 value, $Y^{V7} \equiv Y^0/\hat{b}$, and this was done for all past and new OCO-2 observations.

In what follows, we use our proposed method on the same regression data $\{(X_i, Y_i) : i = 1, \ldots, N\}$ and individual OCO-2 and TCCON observations that were used to obtain Version 7's regression line. That is, we fit a regression line to the $N = 66$ points $\{(X_i, Y_i)\}$ using an EIV model that includes systematic errors (Section 2), and the model parameters are estimated (asymptotically) unbiasedly (Section 3). The two-stage parameter-estimation procedure results in estimates of random-error variances, $\{\hat{\sigma}^2_{\epsilon,x,i} : i = 1, \ldots, N\}$ and $\{\hat{\sigma}^2_{\epsilon,y,i} : i = 1, \ldots, N\}$, at Stage 1. This was achieved in Zhang et al. (2017) by applying a temporal-process model to fit the individual TCCON observations and a spatial-process model to fit the individual OCO-2 observations. We found that the temporal correlations between individual TCCON observations were generally very weak, which resulted in an effective sample size, $\tilde{n}_{x,i}$, very close to the actual sample size, $n_{x,i}$; in contrast, the individual OCO-2 observations displayed strong spatial correlations, which resulted in an effective sample size, $\tilde{n}_{y,i}$, much smaller than the actual sample size, $n_{y,i}$. The estimated marginal variances of the individual observations divided by the effective sample sizes yield the estimated random-error variances.

Consider the i-th individual TCCON observations $\{\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,n_{x,i}}\}$ that were used to compute the i-th regression datum X_i. Such individual TCCON observations have already been tied to the WMO scale using aircraft profile data (Wunch et al., 2010; Messerschmidt et al., 2011). The aircraft profile data are collected using precise and accurate *in situ* instrumentation flown on aircraft over the TCCON stations, which can be treated to have zero systematic errors. By comparing 31 independent (aggregated) aircraft profiles of CO_2 with coincident TCCON measurements, a Method-of-Moments estimate of τ_x^2 can be obtained:

$$\hat{\tau}_x^2 = \max\left\{\sum_{i=1}^{31}(X_i - Z_i)^2 - \hat{\sigma}_{\epsilon,x,i}^2 - \hat{\sigma}_{\epsilon,z,i}^2,\ 0\right\},$$

where $\{(X_i, Z_i)\}$ are pairs of the TCCON and aircraft observations, and $\hat{\sigma}_{\epsilon,x,i}^2$ and $\hat{\sigma}_{\epsilon,z,i}^2$ are estimated random-error variances for the TCCON and aircraft data, respectively. Note that this validation procedure is independent of the OCO-2 validation, and the TCCON data used here are different from those used in the OCO-2 validation. The estimated systematic-error variance in TCCON is $\hat{\tau}_x^2 = 0.258$, which will be substituted into Stage 2 of our estimation procedure.

Recall that at Stage 2, we substituted the parameter estimates from the previous stage into the unbiased estimating equations given in Section 3.3 and estimated the remaining model parameters. Since Version 7 of the OCO-2 data fixed the intercept a to be zero, we considered three estimation scenarios: We estimated a, b, and τ_y^2 all together; we estimated b and τ_y^2 with a fixed at zero; and we estimated slope b with intercept a and the systematic-error variances τ_x^2 and τ_y^2 fixed at zero. These results were compared with those of Version 7, which did not model systematic errors in $\{(X_i, Y_i)\}$ and fixed a at zero. Note that the last scenario is closest to Version 7, only differing in how the regression weights in (5) are specified (see below). The parameter-estimation results are given in Table 5.

Table 5: Parameter-estimation results for the regression analysis, with estimated asymptotic standard errors of parameter estimates given in parentheses. The slope b is associated with the TCCON covariate. The sum of squared residuals (SSR) is also reported, where $\mathrm{SSR} = \sum_{i=1}^{N}(Y_i - \hat{a} - \hat{b}X_i)^2$.

Model parameters	a	b	τ_y^2	SSR
a, b, τ_y^2	-6.462 (17.937)	1.013331 ($4.51 \cdot 10^{-2}$)	0.514 (0.159)	66.34
$a = 0$, b, τ_y^2	0 (fixed)	0.997100 ($3.00 \cdot 10^{-4}$)	0.516 (0.158)	65.96
$a = 0$, b, $\tau_x^2 = \tau_y^2 = 0$	0 (fixed)	0.996601 ($5.26 \cdot 10^{-5}$)	0 (fixed)	69.22
Version 7	0 (fixed)	0.996941 ($1.15 \cdot 10^{-3}$)	–	66.43

For the proposed EIV model that includes the systematic-error variances τ_x^2 and τ_y^2, both the zero-intercept case (a is fixed at 0 and estimate b) and the nonzero-intercept case (estimate both a and b) result in a comparable and statistically significant nonzero estimate of τ_y^2. The approximate 95% confidence interval for τ_y^2 is $(0.206, 0.826)$ for the zero-intercept case and $(0.202, 0.826)$ for the nonzero-intercept case. Thus, we have statistical confidence that the systematic-error variance τ_y^2 is not negligible for OCO-2 observations and should be accounted for. For the zero-intercept case, we can compare the models with systematic error (τ_x^2 is nonzero and estimate τ_y^2) and without systematic error ($\tau_x^2 = \tau_y^2 = 0$). When τ_y^2 is estimated, a slightly larger estimate of b with a much more conservative standard error is obtained than those obtained from the model ignoring the systematic errors in X_i and Y_i. We also see a smaller SSR value for the model with nonzero systematic errors, as expected. Since, from Table 5, intercept a is not significantly different from zero, we follow the same physical reasoning given in the Version 7 validation that led to fixing $a = 0$ (i.e., zero XCO_2 should be recognized by both instruments). That is, we focus on the model that estimates b and τ_y^2 with $a = 0$ and nonzero τ_x^2. Then we compare our results with those of Version 7.

The Version 7 validation used the algorithm in York et al. (2004) to estimate b (with intercept a fixed at zero) by minimizing the objective function in equation (5). It attempted to account for systematic errors by inflating the uncertainties using the sample variances, $S_{x,i}^2$ and $S_{y,i}^2$, of the individual observations to estimate $\mathrm{var}(X_i) \equiv \sigma_{x,i}^2$ and $\mathrm{var}(Y_i) \equiv \sigma_{y,i}^2$.

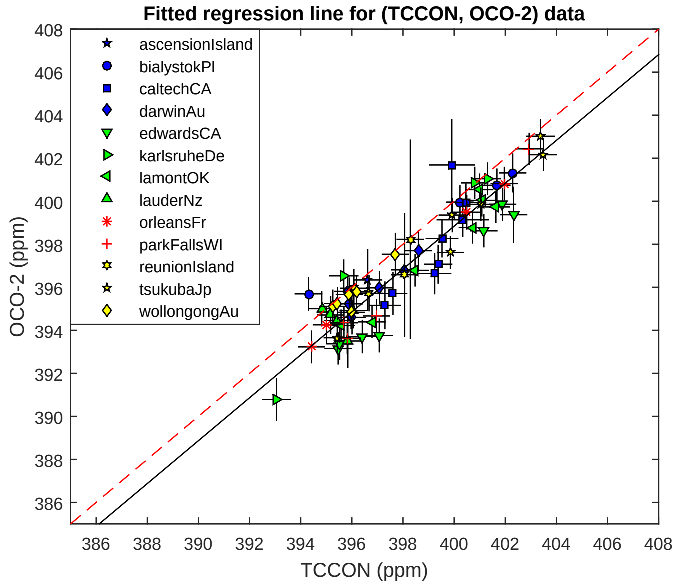


Figure 2: The solid line is the fitted straight line for the proposed EIV model that estimates b and τ_y^2 with $a = 0$ and nonzero τ_x^2. The dashed line is the reference line with slope one and intercept zero. The error bars of plus or minus one standard error were also plotted for each regression point (X_i, Y_i).

Then the regression weights in equation (5) were specified as $w_{x,i} = 1/S_{x,i}^2$ and $w_{y,i} = 1/S_{y,i}^2$, for $i = 1, \ldots, N$. However, there was no accounting for the spatial/temporal correlations between the individual observations for the Version 7 validation. After obtaining estimates of a and b, their standard errors were calculated following the procedure in York et al. (2004) that is based on a method of partial differentiation. However, from the statistical methodology presented in this paper, there is doubt that the Version 7 estimates are unbiased and that the standard errors calculated yield valid confidence intervals.

The Version 7 results are given in the last row of Table 5. We can see that the Version 7 estimated regression slope is slightly smaller than that given by our EIV analysis in the presence of systematic-error variances τ_x^2 and τ_y^2. However, the Version 7 standard error of $\hat{b}$ is much larger (more than three times larger) than its counterpart in our proposed model. That is, the standard error from the Version 7 validation appears to be conservative by more than a factor of three, which leads to prediction intervals that are wider than they would be using our proposed methodology.

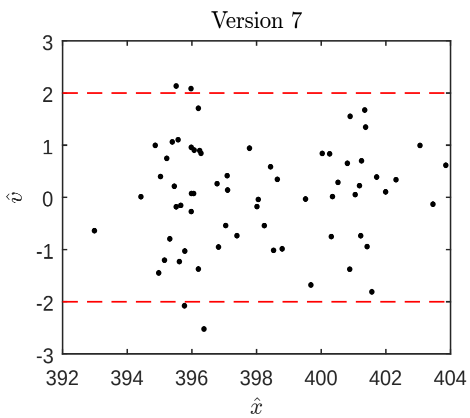
Then we performed model diagnosis for both the Version 7 validation and our proposed unbiased EIV analysis based on the standardized residuals:

$$\hat{v}_i = \frac{(Y_i - \hat{b}X_i)}{\sqrt{\hat{b}^2(\hat{\tau}_x^2 + \hat{\sigma}_{\epsilon,x,i}^2) + \hat{\tau}_y^2 + \hat{\sigma}_{\epsilon,y,i}^2}}. \tag{18}$$

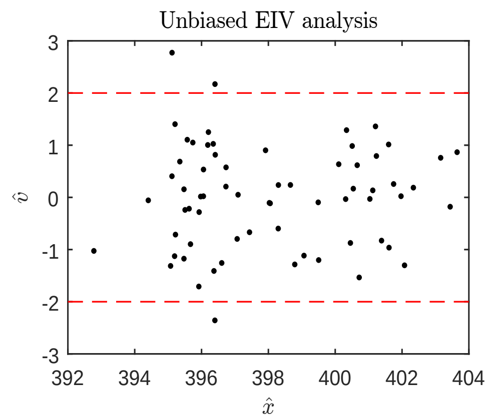
The residuals should satisfy $\hat{v}_i \overset{\text{ind.}}{\sim} \mathcal{N}(0,1)$ approximately, and we plotted them against $\hat{x}_i$, where $\hat{x}_i$ is obtained via maximum likelihood as follows:

$$\hat{x}_i = \frac{(\hat{\sigma}_{\epsilon,y,i}^2 + \hat{\tau}_y^2)X_i + (\hat{\sigma}_{\epsilon,x,i}^2 + \hat{\tau}_x^2)\hat{b}Y_i}{(\hat{\sigma}_{\epsilon,x,i}^2 + \hat{\tau}_x^2)\hat{b}^2 + \hat{\sigma}_{\epsilon,y,i}^2 + \hat{\tau}_y^2}\ . \tag{19}$$

For the Version 7 validation, we remark that in both equations (18) and (19), $\{\hat{\sigma}_{\epsilon,x,i}^2\}$ and $\{\hat{\sigma}_{\epsilon,y,i}^2\}$ were the sample variances calculated from the corresponding individual observations, and $\hat{\tau}_x^2 = \hat{\tau}_y^2 = 0$.



(a) $\hat{v}_i$ versus $\hat{x}_i$.



(b) $\hat{v}_i$ versus $\hat{x}_i$.

Figure 3: Diagnostic plots based on the standardized residuals, $\{\hat{v}_i\}$. Left panel: Version 7; right panel: our proposed EIV analysis with zero intercept and nonzero systematic-error variances τ_x^2 and τ_y^2.

Figure 3 shows the plots of $\{\hat{v}_i\}$ versus $\{\hat{x}_i\}$, where the Version 7 results are shown in the left panel and our EIV results are given in the right panel. For both the Version 7 results and our EIV results, there is no obvious trend in the standardized residuals, and it is clear that most of $\hat{v}_i$ are within two standard deviations from zero. Considering that approximately 95% of these 66 normal random numbers are expected to be within two standard deviations from zero, the 3 or 4 exceptions shown in Figure 3 seem reasonable. We also looked at Q-Q plots for $\{\hat{v}_i\}$ and a standard Gaussian distribution, and we performed the Kolmogorov-Smirnov test (Massey Jr, 1951) and the Shapiro-Wilk test (Shapiro and Wilk, 1965) on $\{\hat{v}_i\}$ for both validation approaches. No departures from Gaussianity were detected in any of these follow-up diagnostics.

Therefore, the Version 7 validation and our proposed validation each produce reasonable results with an estimate of the regression slope b that is close to the other. For a nominal OCO-2 value of 396 ppm, the calibration difference between Version 7 and the proposed method is $396 * (1/0.996941 - 1/0.997100) = 0.063$ ppm. Although the difference is small in this case, our proposed EIV method may be preferred, since it is more statistically justifiable in estimating systematic-error and random-error variances, and it results in an (asymptotically) unbiased estimate of b. The Version 7 results do not enjoy these statistical properties, and our simulations given in Section 4 show that the validation may be biased.

Last, we implemented our proposed EIV analysis with $p = 2$, which includes one more covariate in addition to TCCON. The additional covariate considered here is one of: Latitude (denoted by "Lat") of the OCO-2 observations; an indicator variable that equals one if the OCO-2 observation location is in the Northern Hemisphere and equals zero otherwise (denoted by "Hem"); and the solar zenith angle of the OCO-2 observations (denoted by "Sza") during a target-mode maneuver. Notice that these covariates were assumed to be observed without measurement error, which is a physically realistic assumption. The parameter-estimation results for these cases are given in Table 6. We fixed the intercept a at zero for all cases, as did the Version 7 validation, and because a was not significantly different from zero for the EIV model that estimated both a and $\mathbf{b}$.

Table 6: Parameter-estimation results for the proposed EIV model with $p = 2$ and intercept $a = 0$, where the asymptotic standard errors of parameter estimates are given in parentheses. SSR is defined in the Table 5 caption.

Covariates	b_1 (TCCON)	b_2	τ_y^2	SSR
TCCON	0.997100 ($3.00 \cdot 10^{-4}$)	–	0.516 (0.158)	65.96
TCCON& Lat	0.997473 ($3.41 \cdot 10^{-4}$)	−0.0074 (0.0035)	0.461 (0.148)	62.52
TCCON& Hem	0.998346 ($5.16 \cdot 10^{-4}$)	−0.7087 (0.2459)	0.416 (0.139)	59.83
TCCON& Sza	0.997497 ($1.27 \cdot 10^{-3}$)	−0.0035 (0.0110)	0.513 (0.157)	66.12

We can see from Table 6 that for significance level $\alpha = 5\%$, the regression coefficient of

Sza is not significantly different from zero, and the regression coefficients of Lat and Hem are significant; the coefficient of Hem is also significant for $\alpha = 1\%$. For the EIV model with TCCON and Hem as covariates and a nominal OCO-2 value of 396 ppm, the difference between the validation correction at a Northern Hemisphere location and at a Southern Hemisphere location is

$$(396 + 0.7087)/0.998346 - 396/0.998346 = 0.7087/0.998346 = 0.710 \text{ ppm},$$

indicating a large hemisphere effect on the validation process.

It is likely that Hem is a proxy for a more appropriate second covariate, as it is not descriptive of a physical characteristic of the measurements. The bias with respect to Hem may be caused by aerosol-loading differences between hemispheres or another physical property to which the radiance measurements or retrievals are sensitive. We also performed a simulation study for the misspecified EIV model that misses one covariate in the regression relation, and the results show that the resulting estimates of the remaining regression coefficients can be biased (see the Supplementary Material).

Including more covariates in the proposed EIV model and testing their significance assures an appropriate validation. If some of those covariates are highly related, their cross-dependencies may need to be captured in our EIV model through the off-diagonal entries in the random-error covariance matrices, $\{\Sigma_{\epsilon,x,i} : i = 1, \ldots, N\}$. Let $\mathcal{D}_{x,i} \equiv \cup_{j=1}^{p} \mathcal{D}_{x,i,j}$ be the set of all individual covariate measurements for the i-th regression datum $\mathbf{X}_i$. Then for $1 \leq j < k \leq p$, to estimate $\text{cov}(X_{i,j}, X_{i,k})$, REML estimation for a p-variate Gaussian process with a parametric cross-covariance function (Genton and Kleiber, 2015) on $\mathcal{D}_{x,i}$ can be applied. According to Mandrake et al. (2015), before using the regression relation between TCCON and OCO-2 to conduct the global bias correction, the effects of environmental variables such as the surface pressure and the retrieved abundance of coarse aerosol have been mostly removed from the OCO-2 data as a form of pre-processing. Nevertheless, treating environmental variables as covariates in our multivariate EIV model would both serve as a diagnostic of the pre-processing and capture any remaining effects.

6. CONCLUDING REMARKS

In this paper, we have proposed new methodology for inference in a multiple-regression model with errors in variables. An important application is to the validation of satellite remote sensing data (Y) from a concomitant data source ($\mathbf{X}$), where for the i-th regression datum $(\mathbf{X}_i, Y_i)$, the measurement-error term in the covariate vector $\mathbf{X}_i$ and the response Y_i potentially comprises a random-error component and a systematic-error component. The systematic-error component is used to account for possible biases in the regression data, which can lead to a biased estimate of the regression relation between the means of $\{\mathbf{X}_i : i = 1\ldots, N\}$ and the means of $\{Y_i : i = 1, \ldots, N\}$. Identifiability issues for parameters of the proposed model are resolved when individual observations become available that, when aggregated, result in $\mathbf{X}_i$ and Y_i. Then it is clear that the only extra knowledge needed is the systematic-error variances of the covariates $\{\mathbf{X}_i\}$. In this paper, a two-stage parameter-estimation procedure is defined that is practical, consistent, and efficient.

When applying our proposed EIV model (with $p = 1$) to validate the OCO-2 satellite data using the more accurate and precise ground-based TCCON data, we obtained a significant systematic-error variance estimate, $\hat{\tau}_y^2$, for modeling the error variance of $\{Y_i\}$. Through simulation, we saw that failure to account for τ_y^2 can lead to substantial biases in the regression-coefficient estimates.

We compared our results to those from OCO-2's Version 7 validation. There, an attempt was made to account for systematic errors by using the sample variances of the individual observations to estimate the regression-data variances in (5). The resulting estimate of the slope b is comparable to that from our unbiased EIV analysis, with confidence interval about 3 times larger than our result. However, since the Version 7 procedure ignored the spatial/temporal correlations between the individual satellite/ground-based observations, the estimates of $\mathrm{var}(X_i)$ and $\mathrm{var}(Y_i)$ are themselves biased in general. Hence, there is doubt that the Version 7 estimates of the regression coefficients enjoy the statistical property of unbiasedness. Our proposed methodology properly accounts for systematic errors and infers the regression coefficients consistently and efficiently.

ACKNOWLEDGEMENTS

The OCO-2 data used in this article were produced by the OCO-2 project at the Jet Propulsion Laboratory, California Institute of Technology, and they were obtained from the OCO-2 data archive maintained at the NASA Goddard Earth Science Data and Information Services Center (`http://disc.sci.gsfc.nasa.gov/OCO-2`). TCCON data were obtained from the TCCON Data Archive, hosted by the Carbon Dioxide Information Analysis Center (CDIAC) (`tccon.ornl.gov`). Zhang and Cressie's research was partially supported by a 2015-2017 Australian Research Council Discovery Grant, number DP150104576; Cressie's research was also partially supported by NASA grant NNH11-ZDA001N-OCO2. We thank Gregory Osterman for helpful discussions related to this research.

APPENDIX A. DERIVATION OF THE PROFILE LOG-LIKELIHOOD IN (8)

The fixed but unknown parameters are the means $\{\mathbf{x}_i : i = 1, \ldots, N\}$ and $\boldsymbol{\theta}$ whose log-likelihood is defined by (7). By using the score equation, $\partial \ell(\mathbf{x}_i, \boldsymbol{\theta})/\partial \mathbf{x}_i = \mathbf{0}$, one can obtain:

$$\hat{\mathbf{x}}_i(\boldsymbol{\theta}) = \left(\Sigma_{x,i}^{-1} + \frac{\mathbf{b}\mathbf{b}^T}{\sigma_{\epsilon,y,i}^2 + \tau_y^2} \right)^{-1} \left(\Sigma_{x,i}^{-1} \mathbf{X}_i + \frac{(Y_i - a)\mathbf{b}}{\sigma_{\epsilon,y,i}^2 + \tau_y^2} \right), \tag{A.1}$$

where recall that $\Sigma_{x,i} \equiv \mathrm{var}(\mathbf{X}_i) = \Sigma_{\epsilon,x,i} + T_x$, for the systematic-error covariance matrix T_x and the random-error covariance matrix $\Sigma_{\epsilon,x,i}$. By substituting (A.1) into the log-likelihood in (7) and after some simplifications, one can obtain the profile log-likelihood,

$$\begin{aligned} \ell(\boldsymbol{\theta}) \quad = \quad & -\frac{1}{2} \sum_{i=1}^{N} \mathbf{X}_i^T \left(\Sigma_{x,i}^{-1} - \Sigma_{x,i}^{-1} A_i^{-1} \Sigma_{x,i}^{-1} \right) \mathbf{X}_i + \sum_{i=1}^{N} \frac{(Y_i - a)\mathbf{X}_i^T \Sigma_{x,i}^{-1} A_i^{-1} \mathbf{b}}{\sigma_{\epsilon,y,i}^2 + \tau_y^2} \\ & -\frac{1}{2} \sum_{i=1}^{N} (Y_i - a)^2 (\mathbf{b}^T \Sigma_{x,i} \mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2)^{-1} \\ & -\frac{1}{2} \sum_{i=1}^{N} \log |\Sigma_{x,i}| - \frac{1}{2} \sum_{i=1}^{N} \log(\sigma_{\epsilon,y,i}^2 + \tau_y^2) + \text{constant}, \end{aligned} \tag{A.2}$$

where $A_i \equiv \Sigma_{x,i}^{-1} + \mathbf{b}\mathbf{b}^T(\sigma^2_{\epsilon,y,i} + \tau_y^2)^{-1}$. By the Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950; Henderson and Searle, 1981),

$$A_i^{-1} = \Sigma_{x,i} - \frac{\Sigma_{x,i}\mathbf{b}\mathbf{b}^T\Sigma_{x,i}}{\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2} \ .$$

Thus,

$$\Sigma_{x,i}^{-1} - \Sigma_{x,i}^{-1}A_i^{-1}\Sigma_{x,i}^{-1} = \mathbf{b}\mathbf{b}^T/(\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2),$$

and

$$\Sigma_{x,i}^{-1}A_i^{-1}\mathbf{b}/(\sigma^2_{\epsilon,y,i} + \tau_y^2) = \mathbf{b}/(\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2).$$

Therefore, the sum of the first three terms in (A.2) is $\{-\frac{1}{2}\sum_{i=1}^{N}(Y_i - a - \mathbf{b}^T\mathbf{X}_i)^2/(\mathbf{b}^T\Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2)\}$, resulting in the profile log-likelihood given by (8).

APPENDIX B. CALCULATIONS OF THE GODAMBE INFORMATION MATRIX

For notational simplicity, in this appendix we use $U_{\tau_y^2}(\boldsymbol{\theta})$ to denote the adjusted score function of τ_y^2 that is unbiased (and denoted by $\tilde{U}_{\tau_y^2}(\boldsymbol{\theta})$ in Section 3.3). Let $\mathbf{U}(\boldsymbol{\theta}) \equiv (U_a(\boldsymbol{\theta}), \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})^T, U_{\tau_y^2}(\boldsymbol{\theta}))^T$ be the column vector of unbiased estimating functions and $V(\boldsymbol{\theta}) \equiv \partial\mathbf{U}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ be the Hessian matrix. The Godambe information matrix for unbiased estimating functions (Section 3.3) is given as follows (Godambe, 1960):

$$E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})^T)E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)^{-1}E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})). \tag{A.3}$$

We now calculate $E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta}))$ and $E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)$ in (A.3). Recall that $\omega_i \equiv \mathbf{b}^T \Sigma_{x,i}\mathbf{b} + \sigma^2_{\epsilon,y,i} + \tau_y^2$; then

$$\begin{aligned}
U_a(\boldsymbol{\theta}) &= \sum_{i=1}^N \frac{1}{\omega_i}(Y_i - a - \mathbf{X}_i^T\mathbf{b}),\\
\mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta}) &= \sum_{i=1}^N \frac{1}{\omega_i}(Y_i - a - \mathbf{X}_i^T\mathbf{b})\mathbf{X}_i + \sum_{i=1}^N \frac{1}{\omega_i^2}(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 \Sigma_{x,i}\mathbf{b},\\
U_{\tau_y^2}(\boldsymbol{\theta}) &= \frac{1}{2}\sum_{i=1}^N \frac{1}{\omega_i^2}(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 - \frac{1}{2}\sum_{i=1}^N \frac{1}{\omega_i}.
\end{aligned}$$

The partial derivatives of these estimating functions are as follows:

$$\begin{aligned}
\frac{\partial U_a(\boldsymbol{\theta})}{\partial a} &= -\sum_{i=1}^N \frac{1}{\omega_i},\\
\frac{\partial \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})}{\partial \mathbf{b}^T} &= \sum_{i=1}^N \frac{1}{\omega_i^3}\{-\omega_i^2 \mathbf{X}_i\mathbf{X}_i^T - 2\omega_i(Y_i - a - \mathbf{X}_i^T\mathbf{b})(\mathbf{X}_i\mathbf{b}^T\Sigma_{x,i} + \Sigma_{x,i}\mathbf{b}\mathbf{X}_i^T)\\
&\quad +(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 \Sigma_{x,i}(\omega_i I_p - 4\mathbf{b}\mathbf{b}^T\Sigma_{x,i})\},\\
\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \tau_y^2} &= -\sum_{i=1}^N \frac{1}{\omega_i^3}(Y_i - a - \mathbf{X}_i^T\mathbf{b})^2 + \frac{1}{2}\sum_{i=1}^N \frac{1}{\omega_i^2},
\end{aligned}$$

where I_p is the $p \times p$ identity matrix. Similarly,

$$\begin{aligned}
\frac{\partial U_a(\boldsymbol{\theta})}{\partial \mathbf{b}} &= \frac{\partial \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})}{\partial a} = -\sum_{i=1}^N \frac{\mathbf{X}_i}{\omega_i} - 2\sum_{i=1}^N \frac{1}{\omega_i^2}(Y_i - a - \mathbf{X}_i^T\mathbf{b})\Sigma_{x,i}\mathbf{b},\\
\frac{\partial U_a(\boldsymbol{\theta})}{\partial \tau_y^2} &= \frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial a} = -\sum_{i=1}^N \frac{Y_i - a - \mathbf{X}_i^T\mathbf{b}}{\omega_i^2},\\
\frac{\partial \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})}{\partial \tau_y^2} &= -\sum_{i=1}^N \frac{Y_i - a - \mathbf{X}_i^T\mathbf{b}}{\omega_i^3}(\omega_i\mathbf{X}_i + 2(Y_i - a - \mathbf{X}_i^T\mathbf{b})\Sigma_{x,i}\mathbf{b}),\\
\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \mathbf{b}} &= \frac{\partial \mathbf{U}_{\mathbf{b}}(\boldsymbol{\theta})}{\partial \tau_y^2} + \sum_{i=1}^N \frac{\Sigma_{x,i}\mathbf{b}}{\omega_i^2}.
\end{aligned}$$

By taking expectations of the partial derivatives above, we obtain

$$
\begin{aligned}
&E\left(\frac{\partial U_a(\boldsymbol{\theta})}{\partial a}\right) = -\sum_{i=1}^{N} \frac{1}{\omega_i}, \quad E\left(\frac{\partial U_a(\boldsymbol{\theta})}{\partial \mathbf{b}}\right) = E\left(\frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial a}\right) = -\sum_{i=1}^{N} \frac{\mathbf{x}_i}{\omega_i}, \\
&E\left(\frac{\partial U_a(\boldsymbol{\theta})}{\partial \tau_y^2}\right) = E\left(\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial a}\right) = 0, \quad E\left(\frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial \mathbf{b}^T}\right) = -\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\omega_i}, \\
&E\left(\frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial \tau_y^2}\right) = -\sum_{i=1}^{N} \frac{\Sigma_{x,i}\mathbf{b}}{\omega_i^2}, \quad E\left(\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \mathbf{b}}\right) = 0, \\
&E\left(\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \tau_y^2}\right) = -\frac{1}{2}\sum_{i=1}^{N} \frac{1}{\omega_i^2}.
\end{aligned}
$$

Since

$$
V(\boldsymbol{\theta}) = \begin{pmatrix}
\frac{\partial U_a(\boldsymbol{\theta})}{\partial a} & \frac{\partial U_a(\boldsymbol{\theta})}{\partial \mathbf{b}^T} & \frac{\partial U_a(\boldsymbol{\theta})}{\partial \tau_y^2} \\
\frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial a} & \frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial \mathbf{b}^T} & \frac{\partial \mathbf{U_b}(\boldsymbol{\theta})}{\partial \tau_y^2} \\
\frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial a} & \frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \mathbf{b}^T} & \frac{\partial U_{\tau_y^2}(\boldsymbol{\theta})}{\partial \tau_y^2}
\end{pmatrix},
$$

we obtain the desired $E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta}))$.

Next, we evaluate $E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)$. By using the fact that $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$ are mutually independent and distributed according to a multivariate Gaussian distribution, some straightforward algebra yields

$$
\begin{aligned}
&E(U_a(\boldsymbol{\theta})^2) = \sum_{i=1}^{N} \frac{1}{\omega_i}, \quad E(U_a(\boldsymbol{\theta})\mathbf{U_b}(\boldsymbol{\theta})) = \sum_{i=1}^{N} \frac{\mathbf{x}_i}{\omega_i}, \\
&E(\mathbf{U_b}(\boldsymbol{\theta})\mathbf{U_b}(\boldsymbol{\theta})^T) = \sum_{i=1}^{N} \left(\frac{\Sigma_{x,i}}{\omega_i} + \frac{\mathbf{x}_i\mathbf{x}_i^T}{\omega_i} - \frac{\Sigma_{x,i}\mathbf{b}\mathbf{b}^T\Sigma_{x,i}}{\omega_i^2}\right), \\
&E(U_a(\boldsymbol{\theta})U_{\tau_y^2}(\boldsymbol{\theta})) = 0, \quad E(\mathbf{U_b}(\boldsymbol{\theta})U_{\tau_y^2}(\boldsymbol{\theta})) = \mathbf{0}, \\
&E(U_{\tau_y^2}(\boldsymbol{\theta})^2) = \frac{1}{2}\sum_{i=1}^{N} \frac{1}{\omega_i^2}.
\end{aligned}
$$

Note that $\mathbf{x}_i$ is estimated via maximum likelihood as $\hat{\mathbf{x}}_i(\boldsymbol{\theta}) = (\Sigma_{x,i}^{-1} + \mathbf{b}\mathbf{b}^T/(\sigma_{\epsilon,y,i}^2 + \tau_y^2))^{-1}(\Sigma_{x,i}^{-1}\mathbf{X}_i + (Y_i - a)\mathbf{b}/(\sigma_{\epsilon,y,i}^2 + \tau_y^2))$; see (A.1) in Appendix A. By substituting all parameter estimates obtained in Section 3 into $E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta}))$ and $E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)$, the estimated Godambe information matrix can be readily obtained and used for inference on the parameters and the associated validation.

SUPPLEMENTARY MATERIAL

The supplementary material contains simulation results for the misspecified EIV model that misses one covariate in the regression relation.

REFERENCES

Carroll, R. and D. Ruppert (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician 50*, 1–6.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: CRC Press.

Cheng, C.-L. and J. Riu (2006). On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics 48*, 511–519.

Cheng, C.-L. and J.-R. Tsai (2015). On interval estimation in linear relationships with heteroscedastic measurement errors in both axes. *Chemometrics and Intelligent Laboratory Systems 142*, 276–284.

Cheng, C.-L. and J. W. Van Ness (1999). *Statistical Regression with Measurement Error*. New York: Oxford University Press.

Cressie, N. and D. M. Hawkins (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology 12*, 115–125.

Crisp, D., R. Atlas, F.-M. Breon, L. Brown, J. Burrows, P. Ciais, B. Connor, S. Doney, I. Fung, D. Jacob, et al. (2004). The Orbiting Carbon Observatory (OCO) mission. *Advances in Space Research 34*, 700–709.

Crisp, D., H. R. Pollock, R. Rosenberg, L. Chapsky, R. A. M. Lee, F. A. Oyafuso, and et al. (2017). The on-orbit performance of the Orbiting Carbon Observatory (OCO-2) instrument and its radiometrically calibrated products. *Atmospheric Measurement Techniques 10*, 59–81.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, Hoboken, NJ.

Genton, M. G. and W. Kleiber (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science 30*, 147–163.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics 31*, 1208–1211.

Gong, G. and F. J. Samaniego (1981). Pseudo maximum likelihood estimation: Theory and applications. *Annals of Statistics 9*, 861–869.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–338.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Review 23*, 53–60.

Kulathinal, S., K. Kuulasmaa, and D. Gasbarra (2002). Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine 21*, 1089–1101.

Mandrake, L., C. O'Dell, D. Wunch, P. O. Wennberg, B. Fisher, G. B. Osterman, and A. Eldering (2015). Orbiting Carbon Observatory–2 (OCO-2) warn level, bias correction, and lite file product description. Technical report, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association 46*, 68–78.

Matérn, B. (2013). *Spatial Variation*, Volume 36. Springer Science & Business Media, New York, NY.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B 52*, 325–344.

Messerschmidt, J., M. C. Geibel, T. Blumenstock, H. Chen, N. M. Deutscher, A. Engel, D. G. Feist, C. Gerbig, M. Gisi, F. Hase, et al. (2011). Calibration of TCCON column-averaged CO2: The first aircraft campaign over European TCCON sites. *Atmospheric Chemistry and Physics 11*, 10765–10777.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Riu, J. and F. X. Rius (1995). Univariate regression models with errors in both axes. *Journal of Chemometrics 9*, 343–362.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*, 591–611.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics 21*, 124–127.

Sprent, P. (1966). A generalized least-squares approach to linear functional relationships. *Journal of the Royal Statistical Society, Series B 28*, 278–297.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer Science & Business Media.

Titterington, D. M. and A. N. Halliday (1979). On the fitting of parallel isochrons and the method of maximum likelihood. *Chemical Geology 26*, 183–195.

Viscarra Rossel, R., H. Taylor, and A. McBratney (2007). Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. *European Journal of Soil Science 58*, 343–353.

Wunch, D., G. C. Toon, J.-F. L. Blavier, R. A. Washenfelder, J. Notholt, B. J. Connor, D. W. Griffith, V. Sherlock, and P. O. Wennberg (2011). The Total Carbon Column Observing Network. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 369*, 2087–2112.

Wunch, D., G. C. Toon, V. Sherlock, N. M. Deutscher, C. Liu, D. G. Feist, and P. O. Wennberg (2015). The Total Carbon Column Observing Network's GGG2014 Data Version. Technical report, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

Wunch, D., G. C. Toon, P. O. Wennberg, S. C. Wofsy, B. B. Stephens, M. L. Fischer, O. Uchino, J. B. Abshire, P. Bernath, S. C. Biraud, et al. (2010). Calibration of the Total Carbon Column Observing Network using aircraft profile data. *Atmospheric Measurement Techniques 3*, 1351–1362.

Wunch, D., P. O. Wennberg, G. Osterman, B. Fisher, B. Naylor, C. M. Roehl, C. O'Dell, and et al. (2017). Comparisons of the Orbiting Carbon Observatory-2 (OCO-2) XCO2 measurements with TCCON. Atmospheric Measurement Techniques Discussions.

Yi, G. Y. and N. Reid (2010). A note on mis-specified estimating functions. *Statistica Sinica 20*, 1749–1769.

York, D. (1968). Least squares fitting of a straight line with correlated errors. *Earth and Planetary Science Letters 5*, 320–324.

York, D., N. M. Evensen, M. L. Martínez, and J. D. B. Delgado (2004). Unified equations for the slope, intercept, and standard errors of the best straight line. *American Journal of Physics 72*, 367–375.

Zhang, B., N. Cressie, and D. Wunch (2017). Statistical properties of atmospheric greenhouse gas measurements: Looking down from space and looking up from the ground. *Chemometrics and Intelligent Laboratory Systems 162*, 214–222.