

University of Wollongong

Research Online

Faculty of Engineering and Information
Sciences - Papers: Part B

Faculty of Engineering and Information
Sciences

2018

Inference for Errors-in-Variables Models in the Presence of Systematic Errors with an Application to a Satellite Remote Sensing Campaign

Bohai Zhang

University of Wollongong, bohaid@uow.edu.au

Noel A. Cressie

University of Wollongong, ncressie@uow.edu.au

Debra Wunch

University of Toronto, dwunch@caltech.edu

Follow this and additional works at: <https://ro.uow.edu.au/eispapers1>



Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

Recommended Citation

Zhang, Bohai; Cressie, Noel A.; and Wunch, Debra, "Inference for Errors-in-Variables Models in the Presence of Systematic Errors with an Application to a Satellite Remote Sensing Campaign" (2018).

Faculty of Engineering and Information Sciences - Papers: Part B. 1818.

<https://ro.uow.edu.au/eispapers1/1818>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Inference for Errors-in-Variables Models in the Presence of Systematic Errors with an Application to a Satellite Remote Sensing Campaign

Abstract

Motivated by a satellite remote sensing mission, this article proposes a multivariable errors-in-variables (EIV) 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 regression model is used to estimate the global divisor for the satellite data. The error structure of the proposed EIV model comprises two components: A random-error component whose variance is inversely proportional to sample size of underlying individual observations which are aggregated to obtain 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 obtain estimates from two-stage parameter estimation. We illustrate our proposed procedure through both simulation studies and an application to validating measurements of atmospheric column-averaged CO₂ from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. The validation uses coincident target-mode OCO-2 data that are temporally and spatially sparse and ground-based measurements from the Total Carbon Column Observing Network (TCCON) that are spatially sparse but more accurate. Supplementary materials for the article are available online.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Zhang, B., Cressie, N. & Wunch, D. (2019). Inference for Errors-in-Variables Models in the Presence of Systematic Errors with an Application to a Satellite Remote Sensing Campaign. *Technometrics*, 61 (2), 187-201.

Inference for Errors-in-Variables Models in the Presence
of Systematic Errors 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: bohaid@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 a multivariable errors-in-variables (EIV) 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 regression model is used to estimate the global divisor for the satellite data. The error structure of the proposed EIV model comprises two components: A random-error component whose variance is inversely proportional to sample size of underlying individual observations which are aggregated to obtain 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 obtain estimates from two-stage parameter estimation. We illustrate our proposed procedure through both simulation studies and an application to validating measurements of atmospheric column-averaged CO₂ from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. The validation uses coincident target-mode OCO-2 data that are temporally and spatially sparse and ground-based measurements from the Total Carbon Column Observing Network (TCCON) that are spatially sparse but more accurate.

KEYWORDS: Estimating equations, OCO-2, regression analysis, scaling bias, 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-quality measurements are matched to more-plentiful measurements that cover more factor combinations in the design space. For example, soil scientists measure the mineral content of soil over large tracts inexpensively using proximal γ -ray sensors, but they also take a small number of 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; see the review by Loew et al. (2017). 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, $\mathbf{X} = (X_1, X_2, \dots, X_p)^T \in \mathbb{R}^p$, 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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$ 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

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

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

The estimates \hat{a}_{ols} and $\hat{\mathbf{b}}_{ols} \equiv (\hat{b}_{ols,1}, \dots, \hat{b}_{ols,p})^T$ 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, X_1^0 , 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 $i = 1, \dots, N$, 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 $\{(\mathbf{x}_i, y_i) : i = 1, \dots, N\}$, 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 $i = 1, \dots, N$ (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). The EIV models have been applied to a diverse set of disciplines as a regression-based calibration tool, such as chemistry (Riu and Rius, 1995; Cheng and Riu, 2006; Cheng and Tsai, 2015), epidemiology (Kulathinal et al., 2002; de Castro et al., 2008; Cao et al., 2012), and climatology (Christiansen, 2014).

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

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

where $\{\boldsymbol{\mu}_{x,i}\}$ and $\{\mu_{y,i}\}$ are fixed and represent bias; they are sometimes called systematic errors because they cannot be annihilated through aggregation. The errors $\{\boldsymbol{\epsilon}_{x,i}\}$ and $\{\epsilon_{y,i}\}$ are Gaussian with mean zero, so that now $\boldsymbol{\delta}_{x,i} \sim \mathcal{N}(\boldsymbol{\mu}_{x,i}, \boldsymbol{\Sigma}_{x,i})$ and, independently, $\delta_{y,i} \sim \mathcal{N}(\mu_{y,i}, \sigma_{y,i}^2)$ 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 that model the biases with distributions depending on far fewer parameters. That is, for $i = 1, \dots, N$,

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

where $\{\boldsymbol{\eta}_{x,i} : i = 1, \dots, N\}$ are independent and identically distributed (iid) and follow the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \text{diag}(\tau_{x,1}^2, \dots, \tau_{x,p}^2))$; and, independently of $\boldsymbol{\eta}_{x,i}$, $\{\eta_{y,i} : i = 1, \dots, N\}$ are iid and follow the Gaussian distribution, $\mathcal{N}(0, \tau_y^2)$. Notice that the parameterization in (3) replaces $N(p+1)$ fixed-effects parameters $\{(\boldsymbol{\mu}_{x,i}, \mu_{y,i}) : i = 1, \dots, N\}$ with just $(p+1)$ random-effects parameters $\{\tau_{x,1}^2, \dots, \tau_{x,p}^2, \tau_y^2\}$. Here we capture the systematic error in $\{\mathbf{X}_i\}$ and $\{Y_i\}$ with the random effects $\{\boldsymbol{\eta}_{x,i}\}$ and $\{\eta_{y,i}\}$, respectively. In this article, we use the error structure in (3) to fully account for uncertainties present in $\{\mathbf{X}_i\}$ and $\{Y_i\}$; that is, the “errors” in the EIV model are modeled using (3).

A leading application of (2) and (3) is to the validation component of a remote sensing mission, where the η -terms are referred to as “systematic errors” and the ϵ -terms are re-

ferred to as “random errors.” 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 many other state variables affect the retrieval of the primary variable of interest (e.g., atmospheric carbon dioxide).

All retrieved variables possess uncertainty (Connor et al., 2008, 2016), but some (e.g., aerosols) are more uncertain than others, so that the resulting data products have variabilities that could potentially contain biases. In order to validate satellite remote sensing measurements of selected state variables, the measurements are matched with more precise and accurate (usually ground-based) datasets to identify biases. Validation usually proceeds with a step that fits a linear-regression relationship from a small subset of temporally and spatially sparse satellite data as the response (Y) and a coincident set of more accurate and precise data 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 article, our application focuses on the validation of the state variable, X_{CO_2} , which is the *column-averaged dry-air mole fraction of atmospheric carbon dioxide* (CO_2) concentrations in units of parts per million (ppm), collected by the Orbiting Carbon Observatory-2 (OCO-2) satellite (Crisp et al., 2017). The OCO-2 satellite provides an unprecedented opportunity for observing CO_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 be combined with the calibrated *in situ* surface CO_2 measurements to identify sources and sinks of CO_2 (Wunch et al., 2017). This is achieved through the ground-based Total Carbon Column Observing Network (TCCON) (e.g., Wunch et al., 2011, 2015). TCCON measurements of X_{CO_2} are obtained from around 25 ground-monitoring stations, which provide spatially sparse but temporally dense 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 (e.g., 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, spatially dependent OCO-2 measurements of X_{CO_2} are aggregated to form one value Y , and around 65 of their coincident individual, temporally dependent TCCON measurements of X_{CO_2} 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 at times that are far apart, resulting in regression data $\{(X_i, Y_i) : i = 1, \dots, N\}$; for example, N was equal to 66 for obtaining Version 7 of the OCO-2 data product (Section 5).

In this paper, we focus on the final step of the validation set out in Mandrake et al. (2015), which aims to estimate a global divisor for the OCO-2 data. It should be noted that adjustments of the OCO-2 data using state variables such as the albedo, surface pressure, and the abundance of aerosols (Mandrake et al., 2015) have already been accounted for in earlier validation steps. Then a regression relationship between the (aggregated) OCO-2 target-mode data (Y) and the (aggregated) TCCON data (X) is computed, where spatial and temporal dependencies in the individual measurements are local and are removed by aggregation. This provides a global divisor between the retrieved X_{CO_2} from OCO-2 and TCCON, which is used to do a final adjustment (e.g., the Version-7 OCO-2 data product; see Mandrake et al., 2015). In this article, we shall re-do the calculation of the global divisor based on the statistical methodology we develop in Sections 2 – 4.

Calibration approaches based on the Gaussian process (GP), as well as other related calibration methods (e.g., Kennedy and O’Hagan, 2001; Higdon et al., 2004, 2008; Xiong et al., 2013; Tuo and Wu, 2015, 2016; Konomi et al., 2017), have been widely applied for calibrating computer-code outputs. In this paper, the GP-based calibration in Kennedy and O’Hagan (2001) is not pursued, because the calibration design for OCO-2 results in

regression data that are both temporally and spatially very sparse. The sparsity of the data, inspection of the spatial semivariogram shown in Figure 2 (Section 5), and the conclusion in Section S5 of the supplementary material that the X_{CO_2} process has a spatial range of only 250 - 300 km, imply that a GP-based calibration would not be effective. Rather, we believe that our approach using the mean function and systematic errors is more appropriate for OCO-2 validation.

Furthermore, the Kennedy and O’Hagan (K&OH) calibration approach usually assumes that high-fidelity responses are related to low-fidelity data, but not to the error-free component behind the low-fidelity data. Since in the remote sensing validation problem presented here, both the low-fidelity and high-fidelity datasets can contain noise, the original model used in the K&OH calibration approach would need to be modified to include a noise term, and the relationship between the high-fidelity and the low-fidelity data would need to be established based on their error-free components, to ensure unbiased estimation of model parameters (e.g., Li et al., 2009; Huque et al., 2014).

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 multivariable EIV model is defined with $p \geq 1$ covariates, a special case of which is $p = 1$. In Section 3, we use the local spatial and temporal dependence of individual observations to estimate the variances of the aggregated regression data, and we develop a parameter-estimation procedure for the multivariable EIV model with a discussion of parameter identifiability and the estimators’ properties. In Section 4, our proposed methodology is illustrated through simulation studies. In Section 5, we re-visit the Version-7 calculation between the OCO-2 data and the TCCON data. We compare our proposed EIV inferences, in both univariable-regression and multivariable-regression settings, to the OCO-2 validation team’s inferences. 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 the motivation for the EIV regression models presented in this section. Here, the OCO-2 measurements of X_{CO_2} are regressed on the ground-based

measurements of X_{CO_2} from TCCON, to determine an overall slope between them. The X_{CO_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 $\{(X_i, Y_i) : i = 1, \dots, N\}$ 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 simply called York’s method (York, 1966). The model used for this calibration was:

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

where for $i = 1, \dots, N$, $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are mutually independent measurement errors with mean zero and variances $\sigma_{x,i}^2$ and $\sigma_{y,i}^2$, 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:

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

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 (i.e., $a = 0$), in a regression relating OCO-2 and TCCON X_{CO_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, Titterton 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 $w_{x,i}/w_{y,i} = \sigma_{y,i}^2/\sigma_{x,i}^2$, for $i = 1, \dots, N$ (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 $i = 1, \dots, N$. Since $\{X_i\}$ and $\{Y_i\}$ are typically aggregated data obtained from datasets of individual observations (in Section 5, they are from the TCCON time series and the OCO-2 spatial observations, respectively), Zhang et al. (2017) proposed to estimate error variances of the aggregated 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 article, 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 $i = 1, \dots, N$. 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 univariable-regression setting to the multivariable-regression setting.

2.2 The Proposed Multivariable Errors-in-Variables Model

Here we generalize the univariable EIV model to the multivariable setting where the calibration equation has $p \geq 1$ covariates. For the i -th regression point, let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^T$ 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 multivariable EIV model considered in this paper is: For $i = 1, \dots, N$,

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

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; $\boldsymbol{\eta}_{x,i} \equiv (\eta_{x,i,1}, \dots, \eta_{x,i,p})^T$ 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 $p \times p$ covariance matrix $\text{var}(\boldsymbol{\eta}_{x,i}) \equiv T_x = \text{diag}(\tau_{x,1}^2, \dots, \tau_{x,p}^2)$; $\eta_{y,i}$ is the systematic error (i.e., random effect) in Y_i with mean $E(\eta_{y,i}) = 0$ and variance τ_y^2 ; $\boldsymbol{\epsilon}_{x,i} = (\epsilon_{x,i,1}, \dots, \epsilon_{x,i,p})^T$ is the vector of random errors in \mathbf{X}_i with mean $E(\boldsymbol{\epsilon}_{x,i}) = \mathbf{0}_{p \times 1}$ and $p \times p$ covariance $\Sigma_{\boldsymbol{\epsilon}_{x,i}}$; and $\epsilon_{y,i}$ is the random error in Y_i with mean $E(\epsilon_{y,i}) = 0$ and variance $\sigma_{\epsilon_{y,i}}^2$. 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 the individual observations that are used to obtain the regression data are available in the analysis, and when there exist extra validation data for covariates \mathbf{X} from which we can estimate its systematic-error variances.

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. Such models can be found in a variety of disciplines (e.g., Kulathinal et al., 2002; Cheng and Riu, 2006; Christiansen, 2014). 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 would 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 would not be unbiased.

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

$$\boldsymbol{\theta} \equiv \{a, \mathbf{b}, \Sigma_{\epsilon,x,i}, \sigma_{\epsilon,y,i}^2, T_x, \tau_y^2\}.$$

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

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

where “constant” is a term that does not depend on the parameters. 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),

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

where the “constant” term may be different from that given in (7).

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, critically, the individual observations are available. Then, the square of the sample standard error may seem like a natural choice for estimating the random-error variances, but this assumes that the individual observations are iid. When correlation between individual observations is present, the square of the sample standard error is a biased estimator of the random-error variances. To account for spatial/temporal correlations between individual OCO-2/TCCON observations in this first stage of the calibration, we propose (asymptotically) unbiased estimates based on local spatial/temporal covariance functions; 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. 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 $i = 1, \dots, N$,

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

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 independent mean-zero random errors with variances $\sigma_{\epsilon_{x,i}}^2$ and $\sigma_{\epsilon_{y,i}}^2$, 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:

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

where the fixed but unknown mean parameters $\{x_i\}$ are “profiled out” of the log-likelihood (Titterton and Halliday, 1979). This single-covariate model will be applied to validation of OCO-2’s X_{CO_2} data in Section 5.

3. PARAMETER ESTIMATION

An important application of the multivariable EIV model in (6) is 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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$ and because of the additivity of random effects, the random-error variances, $\{\Sigma_{\epsilon,x,i}\}$ and $\{\sigma_{\epsilon,y,i}^2\}$, are not identifiable in the presence of the respective unknown systematic-error parameters, T_x and τ_y^2 . We are able to resolve this identifiability issue in situations where, behind each $X_{i,1}, \dots, X_{i,p}$ and Y_i , there are sets of individual observations, $\mathcal{D}_{x,i,j} \equiv \{\tilde{X}_{i,j,k} : k = 1, \dots, n_{x,i,j}\}$ for $j = 1, \dots, p$, and $\mathcal{D}_{y,i} \equiv \{\tilde{Y}_{i,\ell} : \ell = 1, \dots, n_{y,i}\}$, respectively, that are available. This is the case for the application to calibration of remote sensing data found in Section 5. Section 3.2 gives details about the estimation of $\{\Sigma_{\epsilon,x,i}\}$ and $\{\sigma_{\epsilon,y,i}^2\}$ from datasets $\{\mathcal{D}_{x,i,j}\}$ and $\{\mathcal{D}_{y,i}\}$ of individual observations. For the moment, we assume that $\{\Sigma_{\epsilon,x,i}\}$ and $\{\sigma_{\epsilon,y,i}^2\}$ 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 $\{\tau_{x,j}^2\}$. The score equation for τ_y^2 is:

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

where recall that for $i = 1, \dots, N$, $\Sigma_{x,i} = \Sigma_{\epsilon,x,i} + T_x$, and $T_x = \text{diag}(\tau_{x,1}^2, \dots, \tau_{x,p}^2)$. Also, for $j = 1, \dots, p$, the score equations for $\{\tau_{x,j}^2\}$ are

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

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:

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

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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$ and the individual observations $\{\mathcal{D}_{x,i,1}\}, \dots, \{\mathcal{D}_{x,i,p}\}$, 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 $\{\tau_{x,j}^2 : j = 1, \dots, p\}$ 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 CO₂ from aircraft in the atmosphere above TCCON sites. Specifically, let $\mathbf{Z}_i \equiv (Z_{i,1}, \dots, Z_{i,p})^T$ denote the (aggregated) validation data coincident with \mathbf{X}_i that do not contain systematic errors,

for $i = 1, \dots, M$. Then for the j -th covariate, we have

$$\begin{aligned} X_{i,j} &= x_{i,j} + \eta_{x,i,j} + \epsilon_{x,i,j}, \\ Z_{i,j} &= x_{i,j} + \epsilon_{z,i,j}, \end{aligned}$$

where recall that $x_{i,j}$ is the fixed but unknown mean parameter representing the true value of $X_{i,j}$, and $\epsilon_{x,i,j}$ and now $\epsilon_{z,i,j}$ are two mean-zero random errors with variances, $\sigma_{\epsilon,x,i,j}^2$ and $\sigma_{\epsilon,z,i,j}^2$, respectively. By taking differences of $(X_{i,j}, Z_{i,j})$ pairs, a method-of-moments estimator can be obtained as follows:

$$\hat{\tau}_j^2 = \max \left\{ \frac{1}{M} \sum_{i=1}^M ((X_{i,j} - Z_{i,j})^2 - \hat{\sigma}_{\epsilon,x,i,j}^2 - \hat{\sigma}_{\epsilon,z,i,j}^2), 0 \right\}, \quad (14)$$

where $\hat{\sigma}_{\epsilon,x,i,j}^2$ and $\hat{\sigma}_{\epsilon,z,i,j}^2$ are estimates of $\sigma_{\epsilon,x,i,j}^2$ and $\sigma_{\epsilon,z,i,j}^2$; the latter can be obtained using the underlying individual observations from which $X_{i,j}$ and $Z_{i,j}$ are obtained. Note that in practice, the values of the $\{\mathbf{X}_i\}$ that are used to estimate the systematic errors $\{\tau_{x,1}^2, \dots, \tau_{x,p}^2\}$ as discussed above, are usually different from those used in the validation of $\{Y_i\}$. Now that T_x has been estimated with $\hat{T}_x \equiv \text{diag}(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2)$, we focus on estimating the remaining model parameters, where we assume that T_x is known.

Our strategy is as follows. 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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$. 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 $\{1, \dots, N\}$.

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

$$\tilde{X}_{j,k} = x_j + \eta_{x,j} + \tilde{\epsilon}_{x,j,k}, \quad \tilde{Y}_\ell = y + \eta_y + \tilde{\epsilon}_{y,\ell}, \quad (15)$$

where $\tilde{\epsilon}_{x,j,k}$ and $\tilde{\epsilon}_{y,\ell}$ are mean-zero random errors with homogeneous variances $\tilde{\sigma}_{x,j}^2$ and $\tilde{\sigma}_y^2$, respectively, for $k = 1, \dots, n_{x,j}$ and $\ell = 1, \dots, n_y$; and $\eta_{x,j}$ and η_y are mean-zero systematic errors with variances $\tau_{x,j}^2$ and τ_y^2 , respectively. Recall that the means $\mathbf{x} \equiv (x_1, \dots, x_p)^T$ 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 $\tilde{\sigma}_{x,j}^2$ and $\tilde{\sigma}_y^2$. Furthermore, if the regression data, $\mathbf{X} = (X_1, \dots, X_p)^T$ and Y , are simple averages, $X_j = \sum_{k=1}^{n_{x,j}} \tilde{X}_{j,k} / n_{x,j}$, for $j = 1, \dots, p$, 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 $\sigma_{\epsilon,x,j}^2 = \tilde{\sigma}_{x,j}^2 / n_{x,j}$ and $\sigma_{\epsilon,y}^2 = \tilde{\sigma}_y^2 / n_y$, respectively. However, correlations between individual observations may be non-negligible due to local spatial (for OCO-2) and temporal (for TCCON) dependencies, which need to be accounted for when estimating $\tilde{\sigma}_{x,j}^2$ and $\tilde{\sigma}_y^2$. For 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 one to two hours), where dependence was modeled through spatial/temporal covariance functions. For example, the Matérn covariance function (e.g., Stein, 1999; Matérn, 2013) can be used in both space and time to model variances and correlations:

$$\mathcal{C}(h; \boldsymbol{\psi}) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (h/\phi)^\nu \mathcal{K}_\nu(h/\phi), \quad (16)$$

where $h \geq 0$ is a spatial/temporal distance between two individual observations. In (16),

$\boldsymbol{\psi} \equiv \{\sigma^2, \phi, \nu\}$, where $\sigma^2 > 0$ is the variance parameter, $\phi > 0$ is a scale 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 (16)). 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 methodology that downweights outliers (e.g., based on the Cressie-Hawkins variogram estimator; see 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 (15) 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, $\hat{\boldsymbol{\psi}}_{x,j} \equiv \{\hat{\sigma}_{x,j}^2, \hat{\phi}_{x,j}, \hat{\nu}_{x,j}\}$ and $\hat{\boldsymbol{\psi}}_y \equiv \{\hat{\sigma}_y^2, \hat{\phi}_y, \hat{\nu}_y\}$, random-error variances in $X_{i,j}$ and Y_i are estimated by

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

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, $\text{var}(X_j)$ and $\text{var}(Y)$ can be estimated under temporal dependence and spatial dependence, respectively, 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:

$$U_a(\boldsymbol{\theta}) \equiv \partial \ell(\boldsymbol{\theta}) / \partial a = 0, \quad \mathbf{U}_b(\boldsymbol{\theta}) \equiv \partial \ell(\boldsymbol{\theta}) / \partial \mathbf{b} = \mathbf{0}, \quad \text{and} \quad U_{\tau_y^2}(\boldsymbol{\theta}) \equiv \partial \ell(\boldsymbol{\theta}) / \partial \tau_y^2, \quad (17)$$

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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$. According to assumptions of the model in equation (6), we have $E(Y_i^2) = (a + \mathbf{b}^T \mathbf{x}_i)^2 + \sigma_{\epsilon, y, i}^2 + \tau_y^2$, $E(\mathbf{X}_i \mathbf{X}_i^T) = \mathbf{x}_i \mathbf{x}_i^T + \Sigma_{x, i}$, and $E(\mathbf{X}_i Y_i) = \mathbf{x}_i (a + \mathbf{x}_i^T \mathbf{b})$, for $i = 1, \dots, N$. Then from (8) and (17),

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

and

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

where $\omega_i \equiv \mathbf{b}^T \Sigma_{x, i} \mathbf{b} + \sigma_{\epsilon, y, i}^2 + \tau_y^2 = \text{var}(Y_i - a - \mathbf{X}_i^T \mathbf{b})$. 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 given in Yi and Reid (2010), where the roots of the biased estimating equations are transformed.

Let $\mathbf{U}(\boldsymbol{\theta}) \equiv (U_a(\boldsymbol{\theta}), \mathbf{U}_b(\boldsymbol{\theta})^T, \tilde{U}_{\tau_y^2}(\boldsymbol{\theta}))^T$ 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:

$$E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta}))^{-1} E_{\boldsymbol{\theta}}(\mathbf{U}(\boldsymbol{\theta}) \mathbf{U}(\boldsymbol{\theta})^T) E_{\boldsymbol{\theta}}(V(\boldsymbol{\theta})^T)^{-1}, \quad (18)$$

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,

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

where $\hat{\sigma}_{\epsilon,y,i}^2$ 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 multivariable regression with two covariates (i.e., $p = 2$) and show by simulation that ignoring the nonzero systematic errors in $\{Y_i\}$ results 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 $\{x_{i,1} : i = 1, \dots, N\}$ and $\{x_{i,2} : i = 1, \dots, N\}$, 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 increasing N , sample sizes $N = 70, 150, 600$ were chosen. In what follows, the simulation was repeated $L = 500$ times, so that bias and variance of parameter estimates could be ascertained using Monte Carlo estimates. For $i = 1, \dots, N$, the errors for the observed covariate vector, $\mathbf{X}_i = (X_{i,1}, X_{i,2})^T$, 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 $\{\Sigma_{\epsilon,x,i} : i = 1, \dots, N\}$ were specified as follows: $\sigma_{\epsilon,x,i,1}^2 \equiv \Sigma_{\epsilon,x,i}(1, 1) = (0.1x_{i,1})^2$, $\sigma_{\epsilon,x,i,2}^2 \equiv \Sigma_{\epsilon,x,i}(2, 2) = (0.1x_{i,2})^2$, and $\Sigma_{\epsilon,x,i}(1, 2) = \Sigma_{\epsilon,x,i}(2, 1) = \rho \sqrt{\sigma_{\epsilon,x,i,1}^2 \sigma_{\epsilon,x,i,2}^2}$, with $\rho = 0.5$. The systematic-error covariance matrix was specified as $T_x = \text{diag}(0.5, 0.5)$. The responses $\{Y_i : i = 1, \dots, N\}$ were then randomly generated from $\mathcal{N}(y_i, \sigma_{\epsilon,y,i}^2 + \tau_y^2)$, 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 $\sigma_{\epsilon,y,i}^2 = (0.25y_i)^2$ over $i = 1, \dots, N$, and $\tau_y^2 = 2$.

Since $\frac{1}{N} \sum_{i=1}^N (y_i^2 / (\sigma_{\epsilon,y,i}^2 + \tau_y^2))$ can be interpreted as a signal-to-noise ratio (SNR) that

can affect the estimation of regression parameters, we chose different levels of SNR in our simulation study. For the EIV-model parameters specified above, the SNR value is about 12, corresponding to a scenario with a relatively high SNR (denoted by “HI”). We also chose a low-SNR scenario (denoted by “LO”) with $a = 1/3$, $\mathbf{b} = (b_1, b_2)^T = (1/6, 1/3)^T$, $\sigma_{\epsilon,y,i}^2 = (0.75y_i)^2$, 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_y^2 , 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_y^2 . Estimates of a , \mathbf{b} , and τ_y^2 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 $\tau_y^2 = 2$ is used (denoted by “TRU”). The misspecified value of $\tau_y^2 = 0$ 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.

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 Monte Carlo (MC) means of parameter estimates are reported, with MC standard errors in parentheses. Results were obtained based on $L = 500$ simulated regression datasets.

$N = 70$	$a : 1$	$b_1 : 0.5$	$b_2 : 1$	$\tau_y^2 : 2$
UEE	0.952 (0.068)	0.500 (0.005)	1.011 (0.012)	1.477 (0.055)
TRU	1.059 (0.067)	0.498 (0.005)	0.990 (0.012)	—
MSP	0.502 (0.071)	0.512 (0.005)	1.095 (0.013)	—
$N = 150$	$a : 1$	$b_1 : 0.5$	$b_2 : 1$	$\tau_y^2 : 2$
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)	—

For the HI scenario, Table 1 gives the parameter-estimation results under different specifications of τ_y^2 . (This should be compared to the results in Table 4 below, for a low SNR scenario.) 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 $\tau_y^2 = 2$ (TRU); the estimate of τ_y^2 has significantly negative biases for small sample sizes $N = 70$ and 150 , and this remains the case although the estimate is much closer to $\tau_y^2 (= 2)$, for the larger sample size $N = 600$. In contrast, for the case where the systematic errors in $\{Y_i\}$ are ignored (i.e., MSP), estimates of a and \mathbf{b} are biased, especially for the intercept a (with significantly negative biases), and the coefficients b_1 and b_2 have significantly 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 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 relative to that under MSP. Since TRU is the “gold standard,” an RE value as large and as close to 1 as possible, is desirable. The results vary little with N : It is clear that MSP (i.e., the misspecified model 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 compares the empirical (or Monte Carlo) standard errors of parameter estimates for UEE, with the asymptotic standard errors obtained based on the Godambe information matrix given by (18); the asymptotic standard errors of parameter estimates were obtained by substituting in true values of model parameters. Table 3 shows that when sample size is relatively small ($N = 70$ or 150), the asymptotic standard errors of the parameter estimates are comparable to the corresponding empirical ones for \hat{a} and $\hat{\mathbf{b}}$, but the asymptotic standard error of $\hat{\tau}_y^2$ is larger than its empirical counterpart. For the larger sample size, $N = 600$, the

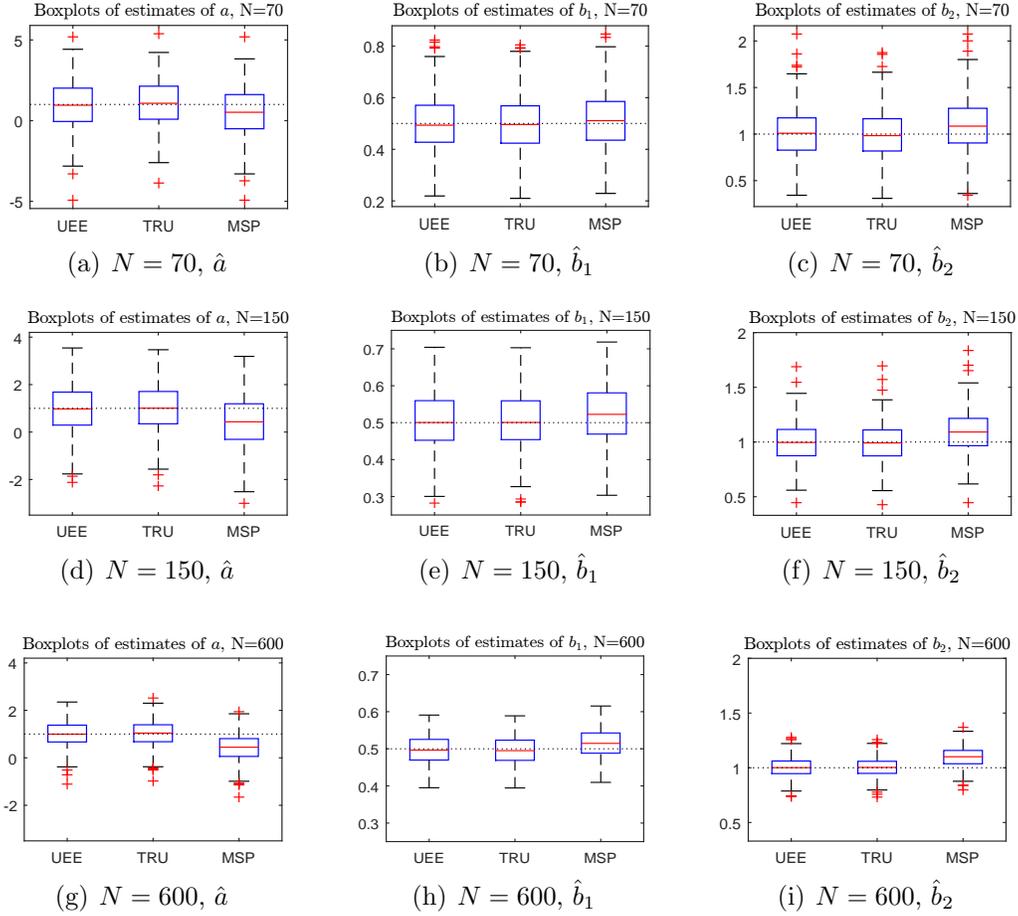


Figure 1: Boxplots of the regression-coefficient estimates from $L = 500$ simulated regression datasets under different specifications of τ_y^2 , namely UEE, TRU, and MSP. Results are given for the HI scenario, namely $a = 1$, $b_1 = 0.5$, $b_2 = 1$, and $\tau_y^2 = 2$. The dotted horizontal lines indicate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

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

$N = 70$	a	b_1	b_2
UEE	0.96	0.98	0.95
MSP	0.80	0.90	0.77
$N = 150$	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

asymptotic standard errors match the empirical results very well for \hat{a} , $\hat{\mathbf{b}}$, and $\hat{\tau}_y^2$.

Table 3: Comparison of the empirical (or Monte Carlo) 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.

$N = 70$	a	b_1	b_2	τ_y^2
EMP	1.4544	0.1042	0.2591	1.1794
ASY	1.5097	0.1033	0.2674	1.4902
$N = 150$	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

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. If one ignores the systematic error in $\{Y_i\}$ (denoted by MSP in the tables), biased estimates of a and $(b_1, b_2)^T$ are obtained, especially for the intercept a (negative bias). 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 $b_2 = 1/3$.

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 MC means of parameter estimates are reported, with MC standard errors in parentheses. Results were obtained based on $L = 500$ simulated regression datasets.

$N = 70$	$a : 0.333$	$b_1 : 0.167$	$b_2 : 0.333$	$\tau_y^2 : 2$
UEE	0.294 (0.059)	0.169 (0.005)	0.333 (0.012)	1.566 (0.053)
TRU	0.318 (0.058)	0.169 (0.005)	0.328 (0.012)	—
MSP	0.112 (0.063)	0.171 (0.005)	0.372 (0.013)	—
$N = 150$	$a : 0.333$	$b_1 : 0.167$	$b_2 : 0.333$	$\tau_y^2 : 2$
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)	—

In this simulation study, we only focused on Stage-2 estimation by using the true values of the remaining variance components, and hence the effects of Stage-1 estimation on Stage-2 estimation were not considered. In practice, the random-error variances and the systematic-error variances in \mathbf{X}_i are estimated from the individual pre-aggregation data, which can bring additional variability to the estimates resulting from Stage-2 estimation. Thus, in any application, sample sizes for the observations in Stage-1 estimation need be sufficiently large to avoid the impact of imprecise Stage-1 parameter estimation.

The parameter-estimation performance of Stage-2 estimation also depends on the signal-to-noise ratio of both the covariates and the response. The simulation studies in this section specify the standard deviations of the two covariates and the response to be 10%, 10%, and 25% of their true values, respectively, which are typically larger than those in our application to OCO-2 validation. When the standard deviations of the covariates and the response are smaller (corresponding to an even higher signal-to-noise ratio scenario), the parameter-estimation performance of τ_y^2 using a small sample size improves substantially (see the supplementary material, Section S1).

5. APPLICATION TO VALIDATION OF OCO-2 SATELLITE 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 on 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 $\{(X_i, Y_i) : i = 1, \dots, 66\}$ (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 $\{(X_i, Y_i) : i = 1, \dots, N\}$ and individual OCO-2 and TCCON observations that were used to obtain Version 7's regression line. That is, we fit a regression line with intercept $a = 0$ to the $N = 66$ points $\{(X_i, Y_i) : i = 1, \dots, 66\}$ using an EIV model that includes systematic errors (Section 2.3) and assumes statistical independence of the random errors across the 66 points. The independence assumption is appropriate since the spatial and temporal distances of the regression data are large and a semivariogram analysis of within-orbit OCO-2 data shows spatial dependence only up to 250 - 300 km (see Figure 2 and Section 5 of the supplementary material). Figure 2 is obtained from the median-based Cressie-Hawkins semivariogram estimator (Cressie, 1993, Section 2.4.3) applied to within-orbit X_{CO_2} data with latitudinal trend removed; the semivariogram levels out at spatial lags on the order of 250 - 300 km. Heterogeneous means of Version-7 data influence the empirical semivariogram at larger spatial and temporal lags, but that behavior is resolved by detrending. Additionally, the validation data are temporally sparse. Hence, the spatio-temporal correlation of random errors in the X_{CO_2} validation data can be considered negligible for our problem.

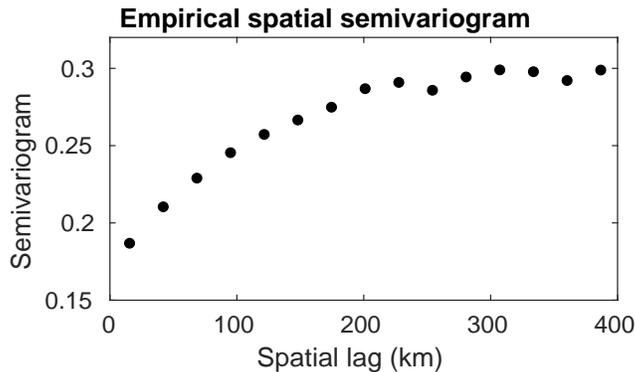


Figure 2: Empirical semivariogram for within-orbit Version-7 OCO-2 data on January 1, 2017.

The two-stage parameter-estimation procedure results in estimates of random-error variances, $\{\hat{\sigma}_{\epsilon,x,i}^2 : i = 1, \dots, N\}$ and $\{\hat{\sigma}_{\epsilon,y,i}^2 : i = 1, \dots, N\}$, at Stage 1. This was achieved in Zhang et al. (2017) by applying a time series model fitted to the individual TCCON observations and a spatial-statistical model fitted to the individual OCO-2 observations. We found

that the local temporal correlations between individual TCCON observations were generally very weak, which resulted in an effective sample size, $\tilde{n}_{x,i}$, that was very close to the actual sample size, $n_{x,i}$; in contrast, the individual OCO-2 observations displayed strong local spatial correlations, which resulted in an effective sample size, $\tilde{n}_{y,i}$, that was much smaller than the actual sample size, $n_{y,i}$. The estimated marginal variances of the individual observations divided by the corresponding effective sample sizes yield the estimated random-error variances.

Consider the i -th individual TCCON observations $\{\tilde{X}_{i,1}, \dots, \tilde{X}_{i,n_{x,i}}\}$ 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 selected TCCON stations, which can be treated as having zero systematic errors. Let Z_i denote the aircraft profile of CO₂ coincident with the TCCON data X_i such that $Z_i = x_i + \epsilon_{z,i}$, where x_i is the true value of X_i and $\epsilon_{z,i}$ is a mean-zero random error with variance $\sigma_{\epsilon,z,i}^2$ (Section 3.1). By comparing 31 independent (aggregated) aircraft profiles of CO₂ with coincident TCCON measurements, a method-of-moments estimate of τ_x^2 was obtained using equation (14):

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

where $\hat{\sigma}_{\epsilon,x,i}^2$ 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 described at the beginning of this section. 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. Also recall that in obtaining the regression line, Version 7 of the OCO-2 data fixed the intercept a to be zero. Of the three estimation scenarios we considered, one

of them allows a to be estimated: Estimate a , b , and τ_y^2 all together; estimate slope b and τ_y^2 with a fixed at zero; and estimate slope b with a fixed at zero and both 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 $\{(X_i, Y_i) : i = 1, \dots, 66\}$ and a was fixed at zero. Note that the last of the three scenarios 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 last column gives the sum of squared residuals (SSR), where $\text{SSR} = \sum_{i=1}^N (Y_i - \hat{a} - \hat{b}X_i)^2$ and $N = 66$.

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
$a = 0, b, \tau_y^2$	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 (τ_x^2 is fixed at 0.258 and τ_y^2 is estimated), both the zero-intercept case (a is fixed at 0 and b is estimated) and the nonzero-intercept case (both a and b are estimated) 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. Diagnostics that check the fit of these models are given in Section S2 of the Supplementary Material.

For the zero-intercept case, we can compare the models *with* systematic error (τ_x^2 is fixed at 0.258 and τ_y^2 is estimated) and *without* systematic error (τ_x^2 and τ_y^2 are both fixed at 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, the 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 X_{CO_2} should be recognized by both instruments). That is, we focus on the model where b and τ_y^2 are estimated with a fixed at 0 and τ_x^2 fixed at 0.258. Then we compare our results with those of Version 7.

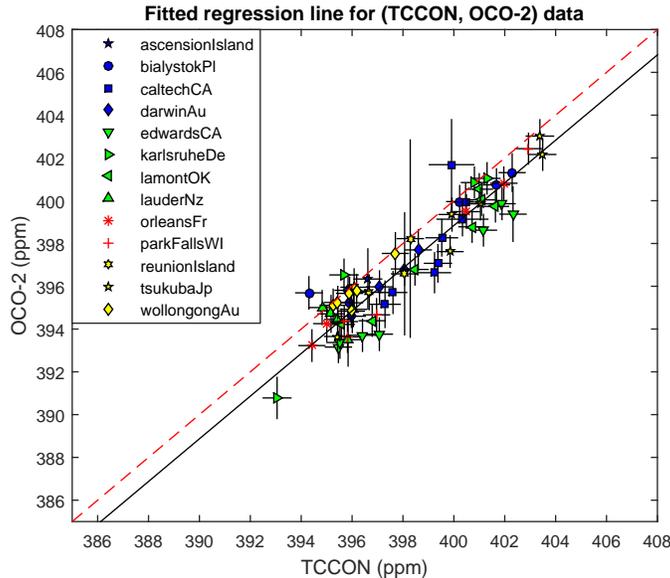


Figure 3: The solid line is the fitted straight line for the proposed EIV model that estimates b and τ_y^2 with $a = 0$ and nonzero τ_x^2 . The dashed line is the reference line with slope 1 and intercept 0. The error bars of plus or minus one standard error were also plotted for each regression point (X_i, Y_i) , for $i = 1, \dots, 66$.

The Version-7 validation used the algorithm in York et al. (2004) to estimate b (with intercept a fixed at 0) by minimizing the objective function in equation (5), and it attempted to account for systematic errors by inflating the uncertainties. Specifically, it used the sample variances, $S_{x,i}^2$ and $S_{y,i}^2$, of the individual observations to estimate $\text{var}(X_i) \equiv \sigma_{x,i}^2$ and $\text{var}(Y_i) \equiv \sigma_{y,i}^2$; 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 $i = 1, \dots, N$. 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. Based on the statistical methodology presented in this paper, there is doubt that the Version-7 estimate of b is unbiased and doubt that the standard error released with the estimate yields valid confidence

intervals, as we now discuss.

The Version-7 results are given in the last row of Table 5, where it can be seen that the 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 Version-7 validation appears to be less efficient by more than a factor of three, which leads to a larger mean squared prediction error of the adjusted OCO-2 retrieval (see Section S3 in the Supplementary Material).

Version-7 validation and the validation we are proposing produce estimates of the regression slope b that are 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 is based on statistical methodology that defines and estimates systematic-error and random-error variances, it results in an (asymptotically) unbiased estimate of b , and it has a much tighter confidence interval for b . The Version-7 results do not enjoy these statistical properties, and the simulations given in Section 4 show it 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. For reasons given above, we fixed the intercept a at 0 for all cases, as did the Version-7 validation.

It can be seen 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

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

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 quite a large “hemisphere effect” within 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 Section S4 of the Supplementary Material). That is, the random and systematic errors in (3) cannot be relied on to compensate for a missing covariate.

Including more covariates in the proposed EIV model for the final step of the validation, followed by significance testing of their regression coefficients, also provides a back-up check to see whether the effects of these covariates have been sufficiently removed in the initial validation steps given in Mandrake et al. (2015). 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, $\{\Sigma_{\epsilon,x,i} : i = 1, \dots, N\}$. 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 to estimate $\text{cov}(X_{i,j}, X_{i,k})$ for $1 \leq j < k \leq p$, REML estimation can be applied to a p -variate Gaussian process with a parametric cross-covariance function (e.g., Genton and Kleiber, 2015) on $\mathcal{D}_{x,i}$. 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 aerosols have been mostly removed from the OCO-2 data in the initial validation steps. Nevertheless, treating environmental variables as covariates in our multivariable EIV model could 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 multivariable errors-in-variables model. 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 $\{\mathbf{X}_i : i = 1 \dots, N\}$ and the means of $\{Y_i : i = 1, \dots, N\}$. Identifiability issues for parameters of the proposed model are resolved when individual observations for the i -th regression point become available and, 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 obtained from OCO-2's Version-7 validation. There, an attempt was made to account for the systematic errors, namely the sample variances of the individual observations were used to estimate the regression-data variances in (5). Both our proposed method and the Version-7 validation produced estimates of the regression slope b that are close to each other, but the Version-7 validation appears to be less efficient by

more than a factor of three. Since the Version-7 procedure did not account for the local spatial/temporal correlations between the individual satellite/ground-based observations, the estimates of $\text{var}(X_i)$ and $\text{var}(Y_i)$ 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

We thank the Associate Editor and all reviewers for their constructive comments on this work, and Gregory Osterman for helpful discussions related to this research. 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 CaltechDATA (<http://tccodata.org>). Zhang and Cressie's research was partially supported by an Australian Research Council Discovery Project, number DP150104576; Cressie's research was also partially supported by NASA grant NNH11-ZDA001N-OCO2.

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

The fixed but unknown parameters are the means $\{\mathbf{x}_i : i = 1, \dots, N\}$ 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:

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

where recall that $\Sigma_{x,i} \equiv \text{var}(\mathbf{X}_i) = \Sigma_{\epsilon,x,i} + T_x$, for the systematic-error covariance matrix T_x and the random-error covariance matrix $\Sigma_{\epsilon,x,i}$. By substituting (A.1) into the joint log-

likelihood (7), and after some simplifications, one obtains the profile log-likelihood,

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

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

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

Thus,

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

and

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

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

APPENDIX B. CALCULATION 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. (It is denoted by $\tilde{U}_{\tau_y^2}(\boldsymbol{\theta})$ in Section 3.3.) Let $\mathbf{U}(\boldsymbol{\theta}) \equiv (U_a(\boldsymbol{\theta}), \mathbf{U}_b(\boldsymbol{\theta})^T, U_{\tau_y^2}(\boldsymbol{\theta}))^T$ 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 estimat-

ing functions (Section 3.3) is given as follows (Godambe, 1960):

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

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 $\omega_i \equiv \mathbf{b}^T \Sigma_{x,i} \mathbf{b} + \sigma_{\epsilon,y,i}^2 + \tau_y^2$; then

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

The partial derivatives of these estimating functions are as follows:

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

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

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

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 $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, N\}$ are mutually independent and distributed according to a multivariate Gaussian distribution, some straightforward algebra yields

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

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 a simulation study on Stage-2 estimation with a very high signal-to-noise ratio, model diagnostics for the EIV regression analysis of the OCO-2/TCCON data in Section 5, results for the prediction variance and bias of the adjusted OCO-2 data, simulation results for the misspecified EIV model that misses one covariate in the regression relation, and some discussion on the range of spatial dependence in the X_{CO_2} validation data.

REFERENCES

- Cao, C.-Z., J.-G. Lin, and X.-X. Zhu (2012). On estimation of a heteroscedastic measurement error model under heavy-tailed distributions. *Computational Statistics & Data Analysis* 56, 438–448.
- 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*. Oxford University Press, New York, NY.
- Christiansen, B. (2014). Straight line fitting and predictions: On a marginal likelihood approach to linear regression and errors-in-variables models. *Journal of Climate* 27, 2014–2031.
- Connor, B., H. Bösch, J. McDuffie, T. Taylor, D. Fu, C. Frankenberg, C. O’Dell, V. H. Payne, M. Gunson, R. Pollock, et al. (2016). Quantification of uncertainties in OCO-2 measurements of X_{CO_2} : Simulations and linear error analysis. *Atmospheric Measurement Techniques* 9, 5227–5238.
- Connor, B. J., H. Boesch, G. Toon, B. Sen, C. Miller, and D. Crisp (2008). Orbiting Carbon Observatory: Inverse method and prospective error analysis. *Journal of Geophysical Research: Atmospheres* 113, D05305. doi:10.1029/2006JD008336.
- Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. John Wiley & Sons, Hoboken, NJ.
- 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.
- de Castro, M., M. Galea, and H. Bolfarine (2008). Hypothesis testing in an errors-in-variables model with heteroscedastic measurement errors. *Statistics in Medicine* 27, 5217–5234.
- 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.
- Higdon, D., J. Gattiker, B. Williams, and M. Rightley (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* 103, 570–583.
- Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafo, and R. D. Ryne (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing* 26, 448–466.
- Huque, M. H., H. D. Bondell, and L. Ryan (2014). On the impact of covariate measurement error on spatial regression modelling. *Environmetrics* 25, 560–570.
- Kennedy, M. and A. O’Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(3), 425–464.

- Konomi, B. A., G. Karagiannis, K. Lai, and G. Lin (2017). Bayesian treed calibration: An application to carbon capture with AX sorbent. *Journal of the American Statistical Association* 112, 37–53.
- 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.
- Li, Y., H. Tang, and X. Lin (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica* 19, 1077.
- Loew, A., W. Bell, L. Brocca, C. E. Bulgin, J. Burdanowitz, X. Calbet, R. V. Donner, D. Ghent, A. Gruber, T. Kaminski, et al. (2017). Validation practices for satellite based earth observation data across communities. *Reviews of Geophysics* 55, 779–817.
- 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.
- 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.
- 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.
- Sprenst, 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.
- Titterton, D. M. and A. N. Halliday (1979). On the fitting of parallel isochrons and the method of maximum likelihood. *Chemical Geology* 26, 183–195.
- Tuo, R. and C. J. Wu (2015). Efficient calibration for imperfect computer models. *Annals of Statistics* 43, 2331–2352.
- Tuo, R. and C. J. Wu (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification* 4, 767–795.
- 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, CaltechDATA, California Institute of Technology, doi: 10.14291/TCCON.GGG2014.DOCUMENTATION.R0/1221662.

- 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* 10, 2209–2238.
- Xiong, S., P. Z. Qian, and C. J. Wu (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics* 55, 37–46.
- Yi, G. Y. and N. Reid (2010). A note on mis-specified estimating functions. *Statistica Sinica* 20, 1749–1769.
- York, D. (1966). Least-squares fitting of a straight line. *Canadian Journal of Physics* 44, 1079–1086.
- 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.