SIAM/ASA J. Uncertainty Quantification
Vol. 10, No. 3, pp. 827–859

Objective Frequentist Uncertainty Quantification for Atmospheric CO_2 Retrievals[*]

Pratik Patil[†], Mikael Kuusela[‡], and Jonathan Hobbs[§]

Abstract. The steadily increasing amount of atmospheric carbon dioxide (CO_2) is affecting the global climate system and threatening the long-term sustainability of Earth's ecosystem. In order to better understand the sources and sinks of CO_2, NASA operates the Orbiting Carbon Observatory-2 and -3 satellites to monitor CO_2 from space. These satellites make passive radiance measurements of the sunlight reflected off the Earth's surface in different spectral bands, which are then inverted in an ill-posed inverse problem to obtain estimates of the atmospheric CO_2 concentration. In this work, we propose a new CO_2 retrieval method that uses known physical constraints on the state variables and direct inversion of the target functional of interest to construct well-calibrated frequentist confidence intervals based on convex programming. We compare the method with the current operational retrieval procedure, which uses prior knowledge in the form of probability distributions to regularize the problem. We demonstrate that the proposed intervals consistently achieve the desired frequentist coverage, while the operational uncertainties are poorly calibrated in a frequentist sense both at individual locations and over a spatial region in a realistic simulation experiment. We also study the influence of specific nuisance state variables on the length of the proposed intervals and identify certain key variables that can greatly reduce the final uncertainty given additional deterministic or probabilistic constraints. We then develop a principled framework to incorporate such additional information into our method.

Key words. Orbiting Carbon Observatory-2 and -3, remote sensing, constrained inverse problem, frequentist coverage, variable importance, convex programming

MSC codes. 62P12, 15A29, 62F30, 90C90

DOI. 10.1137/20M1356403

1. Introduction.

Global measurements of atmospheric carbon dioxide (CO_2) concentration are essential for understanding Earth's carbon cycle, a key component of our planet's climate system. Space-borne observing systems provide the primary way of obtaining atmospheric CO_2 measurements globally at spatial and temporal resolutions useful for investigating central questions in carbon cycle science [40]. A series of satellites named Orbiting Carbon

[*] Received by the editors July 29, 2020; accepted for publication (in revised form) February 9, 2022; published electronically July 29, 2022.
https://doi.org/10.1137/20M1356403
Funding: The work of the first and second authors was supported by JPL subcontract 1629749. The work of the third author was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

[†] Department of Statistics and Data Science and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (pratik@cmu.edu).

[‡] Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (mkuusela@andrew.cmu.edu).

[§] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109 USA (Jonathan.M.Hobbs@jpl.nasa.gov).

Observatory-2 and -3 (OCO-2 and OCO-3) [13, 14], launched by NASA in July 2014 and May 2019, respectively, constitute the current state of the art in space-based CO_2 observing systems. These instruments use the sunlight reflected off the Earth's surface to infer the CO_2 concentration in the atmosphere below. Since the observations are indirect measurements of the quantity of interest, the task of estimating the atmospheric state, known as *retrieval* in remote sensing [41], is an ill-posed inverse problem [23, 16, 53]. The ultimate goal of these missions is to estimate the vertically averaged atmospheric CO_2 concentration at high precision in order to better understand the sources and sinks of CO_2 in the Earth system [10].

Estimating atmospheric CO_2 concentrations from space is a highly nontrivial task. Designing and building the required remote sensing instrument and developing the mathematical forward model for relating the scientifically relevant quantities to the actual satellite observations are both extremely challenging tasks [37]. However, statistically, the main complication arises from the fact that in order to convert the raw satellite observations into CO_2 concentrations, one needs to solve the associated ill-posed inverse problem [8]. A satellite on low-Earth orbit is only able to measure CO_2 indirectly through its effect on the sunlight passing through the atmosphere. Information about CO_2 at different altitudes will therefore inevitably be confounded in the raw observations. Inverting the forward model to obtain a reconstruction of the atmospheric CO_2 profile at different altitudes will hence result in highly oscillatory and uncertain solutions which, at first glance, may seem to have little scientific value.

The OCO-2 and OCO-3 science teams are well aware of these challenges, and the operational missions essentially employ two strategies to circumvent the forward model ill-posedness [4]. First, the missions acknowledge that it is not feasible to retrieve the full vertical CO_2 profile from space. Instead, the missions have identified the vertically averaged CO_2 concentration, denoted by X_{CO2}, as their primary quantity of interest, and the retrieval and validation efforts are focused on the accuracy and precision of this scalar quantity. Second, in order to estimate X_{CO2}, the missions employ a strategy where first a regularized CO_2 profile is reconstructed (or, more precisely, a regularized state vector containing the CO_2 profile and other retrieved atmospheric quantities), which is then used to calculate the corresponding X_{CO2} value. The regularization is achieved using a Bayesian approach where a prior distribution on the underlying state variables is used to promote physically plausible CO_2 profiles [4, 6, 7, 9]. The prior mean of the CO_2 profile is carefully designed to incorporate major large-scale variations in CO_2 over both space (latitude) and time (seasonality, long-term trends) [4, 37]. Even so, regional biases are found in the retrieved X_{CO2} when compared to ground-based validation sources [59, 26, 25, 58].

In this paper, we focus on rigorous uncertainty quantification for the retrieved X_{CO2}. In contrast to most existing works in remote sensing, we approach the problem from the perspective of frequentist statistics. We demonstrate that the existing retrieval procedure, if evaluated using frequentist performance measures, may lead to miscalibrated uncertainties for X_{CO2} due to the intermediate regularization step. We then show that it is possible to obtain better-calibrated uncertainties by adopting an approach that avoids explicit regularization and instead directly forms an implicitly regularized confidence interval for X_{CO2}. The proposed method is developed for linear or linearized forward operators, but extensions to nonlinear cases are possible.

To introduce some of the key ideas, it is worth considering a simplified version of the CO_2 retrieval problem. The problem is typically formulated in terms of an unknown state vector $\boldsymbol{x}$ that includes both the vertical CO_2 profile of interest and other geophysical nuisance variables that affect the satellite observations. Assume that the state vector $\boldsymbol{x}$ is related to the observations $\boldsymbol{y}$ by the linear model $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{K}$ is a known forward operator dictated by the physics of the problem and $\boldsymbol{\varepsilon}$ represents stochastic noise in the measurement device with mean zero and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. The fundamental challenge here is that $\boldsymbol{K}$ is an ill-conditioned matrix so that its singular values decay rapidly. Assume, for now, that $\boldsymbol{K}$ has full column rank, and therefore the least-squares estimator of $\boldsymbol{x}$ is given by $\hat{\boldsymbol{x}} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}$. The covariance matrix of this estimator is $\mathrm{cov}(\hat{\boldsymbol{x}}) = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{K}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}$. Due to the ill-posedness of $\boldsymbol{K}$, the fluctuations in the noise $\boldsymbol{\varepsilon}$ get amplified in the inversion, and the estimator $\hat{\boldsymbol{x}}$ exhibits large oscillations within the CO_2 profile that tend to be anticorrelated from one altitude to the next. This is also reflected in the covariance $\mathrm{cov}(\hat{\boldsymbol{x}})$, and any confidence intervals derived for the individual CO_2 elements in $\boldsymbol{x}$ based on $\mathrm{cov}(\hat{\boldsymbol{x}})$ would be extremely wide, indicating, as they should, that the observations $\boldsymbol{y}$ do not contain enough information to effectively constrain CO_2 at a given altitude. However, this should not deter us from trying to constrain *other functionals of* $\boldsymbol{x}$ based on $\hat{\boldsymbol{x}}$. Of particular interest, in our case, is the vertically averaged CO_2 concentration given by the functional $X_{\mathrm{CO2}} = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$, where $\boldsymbol{h}$ is a known vector of weights. The plug-in estimator of X_{CO2} is $\hat{X}_{\mathrm{CO2}} = \boldsymbol{h}^{\mathrm{T}}\hat{\boldsymbol{x}}$ with variance $\mathrm{var}(\hat{X}_{\mathrm{CO2}}) = \boldsymbol{h}^{\mathrm{T}}\mathrm{cov}(\hat{\boldsymbol{x}})\boldsymbol{h}$. Since the mapping from $\boldsymbol{x}$ to X_{CO2} is an averaging operation, one would expect that the anticorrelated fluctuations in the unregularized $\hat{\boldsymbol{x}}$ largely cancel out as it is mapped into $\hat{X}_{\mathrm{CO2}}$, resulting in a well-behaved estimator of X_{CO2}, as also suggested by the results in [39]. When the noise $\boldsymbol{\varepsilon}$ is Gaussian, which is a good approximation here, one can then use the variance $\mathrm{var}(\hat{X}_{\mathrm{CO2}})$ to construct a frequentist confidence interval around $\hat{X}_{\mathrm{CO2}}$. Assuming that the forward model is correctly specified, these intervals have guaranteed frequentist coverage for X_{CO2}, *without requiring any additional information about* $\boldsymbol{x}$ (e.g., information about smoothness or specification of a prior distribution). Arguably, these intervals provide an objective measure of uncertainty of X_{CO2} in the absence of specific prior information about $\boldsymbol{x}$.

The actual retrieval problem is more complex than the simplified situation described above. First, the forward operator relating the state vector $\boldsymbol{x}$ to the observations $\boldsymbol{y}$ is a nonlinear function of $\boldsymbol{x}$ [4]. Second, there are known physical constraints on the state vector $\boldsymbol{x}$ that should ideally be taken into account in the retrieval. For example, those elements of $\boldsymbol{x}$ that correspond to CO_2 concentrations should be constrained to be nonnegative. Third, the forward mapping need not be injective. This means, for example, that the matrices corresponding to a linearization of the forward mapping may be rank deficient. In this paper, we address these last two complications in the case of a linearized approximation to the nonlinear forward operator. In other words, we seek to rigorously quantify the uncertainty of $X_{\mathrm{CO2}} = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$ under the model $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{K}$ need not have full column rank, $\boldsymbol{x} \in C$, where C is a set of known physical constraints (i.e., constraints that hold with probability 1), and $\boldsymbol{\varepsilon}$ is noise with a known Gaussian distribution. We focus on the case of affine constraints for the elements of the state vector $\boldsymbol{x}$ and, in particular, on nonnegativity constraints for certain elements of the state vector. Under this setup, we seek to construct $(1 - \alpha)$ frequentist confidence intervals for X_{CO2} without imposing any other regularization on $\boldsymbol{x}$. We propose a procedure that is demonstrated to consistently provide nearly nominal $(1-\alpha)$ frequentist coverage, including in situations where the existing retrieval procedure can be severely miscalibrated. Even though our procedure relies on much weaker assumptions, the new intervals are not excessively wide as the problem is implicitly regularized by the choice of the functional $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$ and the constraints $\boldsymbol{x} \in C$.

We also investigate the potential implications of these results on CO_2 flux estimates [15] by studying the behavior of the different methods over a small spatial domain. We find that in the existing operational retrievals, the interaction between the regularizing prior and the spatially dependent true state vectors can lead, at least in the specific example studied, to a situation where the miscalibration of the X_{CO2} intervals varies in a spatially coherent fashion. As a result, the reported uncertainties can be systematically too small or too large over a given spatial region. It is possible that retrievals with such uncertainties could lead to spurious CO_2 flux estimates in downstream analyses. On the other hand, the sampling properties of our proposed intervals do not vary spatially, which makes them potentially more suitable for downstream scientific use.

In addition, we study the contributions of individual state vector elements to the X_{CO2} uncertainty, identifying surface pressure and a certain aerosol variable as the key parameters that contribute most to the final uncertainty. This means that the X_{CO2} uncertainty could potentially be further reduced if additional external information were available to constrain these two variables. We provide a principled framework for incorporating such information in either deterministic or probabilistic forms into our method and investigate the extent to which such additional information on surface pressure reduces the X_{CO2} uncertainty.

This work relates to a wider discussion on uncertainty quantification in ill-posed inverse problems (see, e.g., [49, 50, 54, 52]). In applied situations, uncertainty quantification in inverse problems tends to be dominated by Bayesian approaches that regularize the problem using a prior distribution. This is certainly the case in atmospheric sounding [41], but also in other domain sciences (e.g., [3, 23, 22, 32, 56, 57]). Penalized frequentist techniques, such as penalized maximum likelihood or Tikhonov regularization (also known as ridge regression [21]), are closely related to Bayesian approaches since one can usually interpret the penalty term as a Bayesian log-prior (see, e.g., sections 7.5 and 7.6 in [33]). These techniques, in which the problem is explicitly regularized, are challenging from the perspective of frequentist uncertainty quantification since intervals centered around a regularized point estimator tend to be systematically offset from the true value of the unknown quantity due to the bias in the point estimator that regularizes the problem. This bias has been investigated in multiple remote sensing retrieval settings [41, 34, 31] and has been shown to lead to drastic undercoverage for the intervals in other applied situations [28, 29, 27]. There exists, however, a lesser-known line of work (see [47, 45, 44] and the references therein) that attempts to construct truly frequentist confidence intervals in ill-posed problems without relying on explicitly regularized point estimators. One of the key ideas is to use physically known objective constraints to regularize the problem instead of a subjective prior distribution or a penalty term. This enables deriving intervals with *guaranteed frequentist coverage* [47, 48]. This paper builds upon these ideas, but adds to the discussion by highlighting the important role of the functional of interest in implicitly regularizing the problem. We also focus on intervals which are designed to constrain one functional at a time, in contrast to some previous techniques [47, 29] that provide simultaneous intervals for *all* functionals at once, which leads to substantial overcoverage if only one or a small subset of functionals is needed.

The rest of this paper is organized as follows. To set up the problem, we briefly describe the physics of observing CO_2 from space and the corresponding statistical model in section 2. We then describe the proposed frequentist uncertainty quantification method and discuss its properties in section 3. Next, we outline the existing operational retrieval procedure and analyze its properties in section 4. Section 5 compares the coverage performance of the operational and proposed procedures both for an individual sounding location and over a small spatial region using simulated data from a realistic generative model. In section 6, we further investigate the proposed method to better understand the impact of the individual state vector elements on the final interval length, and we provide a framework in which additional deterministic or probabilistic constraints can be incorporated into our method. Finally, section 7 offers concluding remarks and directions for future work. The appendices and the supplementary material [38] contain derivations and other supplementary results.

2. Problem background and setup.

2.1. Remote sensing of carbon dioxide. Remote sensing of atmospheric CO_2 is feasible due to the absorption of solar radiation by CO_2 molecules at specific wavelengths, particularly in the infrared (IR) portion of the electromagnetic spectrum. In this part of the spectrum, variations in the observed top-of-the-atmosphere radiation can also be induced by other surface and atmospheric properties, including albedo (surface reflectivity), absorption by other atmospheric trace gases, and absorption and scattering in the presence of clouds and aerosol particles. These processes are illustrated schematically in Figure 1. These additional effects explain most of the variation in the radiance (intensity of the observed radiation) that is seen by a downward looking satellite at the top of the atmosphere. Radiance changes due to variation in CO_2 are more subtle. CO_2-focused remote sensing instruments, such as OCO-2 OCO-3, therefore require high-precision radiance observations at fine spectral resolution. The OCO-2 and OCO-3 instruments are duplicates of the same design. Each instrument includes three imaging grating spectrometers that each correspond to a narrow IR band. These are the O_2 A-band centered around 0.765 μm, the weak CO_2 band centered around 1.61 μm, and the strong CO_2 band centered near 2.06 μm. The O_2 A-band includes numerous absorption lines for atmospheric O_2, and the two CO_2 bands include absorption lines for CO_2 [4].

A collection of observed radiances at a particular time and location is known as a *sounding*. For OCO-2 and OCO-3, a sounding includes 1016 radiances in each of the three spectral bands. Figure S1 in the supplementary material [38] depicts an example sounding for OCO-2. The fine wavelength spacing within each band ensures the ability to resolve individual absorption features. Since atmospheric O_2 has a nearly constant fractional abundance of 0.209, the absorption in the O_2 A-band can be used to estimate the total amount of dry air in the atmospheric column, which is sometimes termed the radiative path length. In the retrieval, this is formally represented by retrieving the atmospheric surface pressure. This can be combined with the absolute absorption in the CO_2 bands to estimate the relative abundance, or dry-air mole fraction, of CO_2. In addition, the A-band in particular has sensitivity to cloud and aerosol scattering, which are also estimated in the retrieval process.

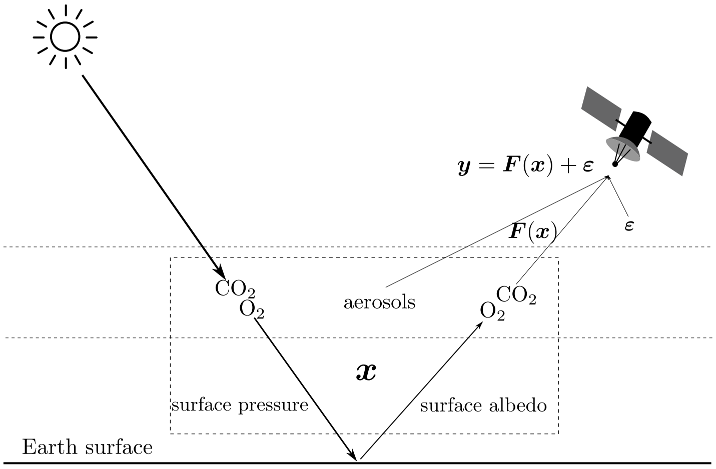


Figure 1. *Schematic of space-based CO_2 sensing in the OCO-2 mission.*

While the instruments themselves are nearly identical, OCO-2 and OCO-3 have different observing patterns. OCO-2 is in a polar orbit as part of a satellite constellation known as the A-train with observations collected exclusively in the early afternoon local time [4]. OCO-3 has recently been installed on the International Space Station (ISS) and is collecting observations in tropical and mid-latitude regions at varying times of the day following the ISS precessing orbit [14]. In the rest of this paper, we primarily focus on OCO-2, but we expect our conclusions also apply to OCO-3 due to the similarity of the two instruments.

2.2. Mathematical model. The physical model for CO_2 remote sensing as illustrated in Figure 1 can be mathematically written as

$$\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is an unknown state vector, $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of observed radiances, $\boldsymbol{F} : \mathbb{R}^p \to \mathbb{R}^n$ models the physical processes described in section 2.1 that relate the state vector to the expected radiances, and $\boldsymbol{\varepsilon}$ represents zero-mean instrument noise. The noise is assumed to have a Gaussian distribution, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$, with a known diagonal covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$. For the OCO-2 mission, we have $n \gg p$. Despite this, the problem of retrieving $\boldsymbol{x}$ based on $\boldsymbol{y}$ is badly ill-posed due to poor conditioning of $\boldsymbol{F}$.

The ultimate goal of the retrieval is to estimate a certain functional of the state vector $\theta(\boldsymbol{x}) \in \mathbb{R}$ using the observations $\boldsymbol{y}$. We assume that the functional of interest is linear so it can be written in the form $\theta = \boldsymbol{h}^T\boldsymbol{x}$, where the weights $\boldsymbol{h}$ are assumed to be known. We specifically focus on θ corresponding to X_{CO2}, the column-averaged CO_2 concentration at the sounding location, which, to a good approximation, is of this form.

The state vector $\boldsymbol{x}$ contains all the physical quantities that are thought to affect the radiance measurement $\boldsymbol{y}$. This includes the vertical CO_2 concentrations, but also other geophysical quantities, as outlined in section 2.1 and described in more detail in section 5.1.1. Statistically, these other quantities can be understood as nuisance variables since the functional of interest does not directly depend on them ($h_i = 0$ for these variables).

The actual full-physics forward operator $\boldsymbol{F}$ is a nonlinear map from $\mathbb{R}^p$ to $\mathbb{R}^n$ [4]. This complication can detract from the fundamental challenges involved in quantifying the uncertainty of the retrievals. In order to be able to focus on the key statistical issues, we linearize in this work the forward operator $\boldsymbol{F}(\boldsymbol{x})$ at a particular $\boldsymbol{x} = \boldsymbol{x}'$ such that $\boldsymbol{F}(\boldsymbol{x}) \approx \boldsymbol{K}\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x}') - \boldsymbol{K}\boldsymbol{x}'$, where $\boldsymbol{K} = \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial \boldsymbol{x}}\big|_{\boldsymbol{x}=\boldsymbol{x}'}$ is the Jacobian of $\boldsymbol{F}$ evaluated at $\boldsymbol{x}'$. This differs from the OCO-2 operational retrieval method which takes the nonlinearity of the forward operator into account. Even so, the operational uncertainty estimate uses a linearization about the final solution [4].

Putting these elements together, we have the following Gaussian linear model:

$$\boldsymbol{y}' = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}), \tag{2.2}$$

where $\boldsymbol{y}' = \boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x}') + \boldsymbol{K}\boldsymbol{x}'$. To simplify the notation, we denote $\boldsymbol{y}'$ as $\boldsymbol{y}$ in the rest of this paper. Under this model, our goal is to obtain a $(1-\alpha)$ confidence interval of the form $[\underline{\theta}, \overline{\theta}]$ for the functional $\theta = \boldsymbol{h}^T\boldsymbol{x}$ with the frequentist coverage guarantee $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}]) \approx 1 - \alpha$ for any $\boldsymbol{x}$, where $(1-\alpha)$ is the desired confidence level and the probability statement is with respect to the distribution of the noise $\boldsymbol{\varepsilon}$.

3. Proposed frequentist retrieval procedure.

3.1. Motivation. The key idea of our proposed method is to let known physical constraints and the functional of interest regularize the problem without imposing other external a priori beliefs about the state elements. We demonstrate using simulations that this suffices for obtaining well-calibrated and reasonably sized confidence intervals, as long as the constraints hold with probability 1 and the functional is an operation, such as averaging or smoothing, that tends to reduce noise. The procedure is formulated in terms of convex optimization problems that find the upper and lower endpoints of the confidence interval [44, 45, 47]. Below, we first describe the procedure, followed by a brief analysis of its properties. We provide two complementary perspectives on the method, one from the point of view of optimization in the state space $\mathbb{R}^p$ and another from the dual perspective of optimization in the radiance space $\mathbb{R}^n$.

3.2. Method outline. Unlike the operational procedure described in detail in section 4, our proposed method directly constructs confidence intervals for the functional of interest $\theta = \boldsymbol{h}^T\boldsymbol{x}$. More specifically, the goal is to construct a $(1-\alpha)$ confidence interval $[\underline{\theta}, \overline{\theta}]$ for θ under the model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}, \boldsymbol{I})$ subject to external information on $\boldsymbol{x}$ in the form of the affine constraint $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ and without requiring $\boldsymbol{K}$ to have full column rank. The linear forward model in (2.2) can always be transformed into this form by taking the Cholesky factorization $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{L}\boldsymbol{L}^T$, calculating $\tilde{\boldsymbol{y}} = \boldsymbol{L}^{-1}\boldsymbol{y}$ and $\tilde{\boldsymbol{K}} = \boldsymbol{L}^{-1}\boldsymbol{K}$, and then redefining $\boldsymbol{y} \leftarrow \tilde{\boldsymbol{y}}$ and $\boldsymbol{K} \leftarrow \tilde{\boldsymbol{K}}$. We assume throughout the remainder of this section that this transformation has been applied to the model. The matrix $\boldsymbol{A}$ and the vector $\boldsymbol{b}$ can encode various types of affine constraints on the state vector elements: for example, nonnegativity constraints for individual elements of $\boldsymbol{x}$, two-sided bounds for individual elements of $\boldsymbol{x}$, or affine constraints involving multiple elements of $\boldsymbol{x}$ at once. The endpoints of the interval $[\underline{\theta}, \overline{\theta}]$ are obtained as the objective function values of two convex optimization problems. The convex programs are chosen so that the coverage $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}])$ is as close as possible to the nominal value $(1-\alpha)$ for all $\boldsymbol{x}$ satisfying the constraint $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$.

3.2.1. Primal point of view.

The lower endpoint $\underline{\theta}$ is the optimal objective function value of the following minimization problem [44, 45]:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^T \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + s^2, \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{array} \tag{3.1}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ standard normal quantile and the slack factor s^2 is defined as the objective function value of the following program:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{array} \tag{3.2}$$

The upper endpoint $\overline{\theta}$ is the optimal value of a similar maximization problem:

$$\begin{array}{ll} \text{maximize} & \boldsymbol{h}^T \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + s^2, \\ & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{array} \tag{3.3}$$

where s^2 is again given by program (3.2).

To explain the intuition behind this construction, we start with the approach described in [47]. Consider the two sets $D = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq \chi_{n,1-\alpha}^2\}$, where $\chi_{n,1-\alpha}^2$ is the $(1-\alpha)$ quantile of the χ^2 distribution with n degrees of freedom, and $C = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$. Here D is a $(1-\alpha)$ confidence set for the entire state vector $\boldsymbol{x}$ and the set C encodes the feasible set of $\boldsymbol{x}$ given the constraints. Therefore, the set $C \cap D$ is also a $(1-\alpha)$ confidence set for $\boldsymbol{x}$. We can then use this confidence set to obtain a $(1-\alpha)$ confidence interval for the functional $\theta = \boldsymbol{h}^T\boldsymbol{x}$ by simply finding the extremal values of the functional over $C \cap D$ [47], which corresponds to problems (3.1) and (3.3) with $z_{1-\alpha/2}^2 + s^2$ replaced by $\chi_{n,1-\alpha}^2$. However, since this choice of D guarantees coverage for the entire vector $\boldsymbol{x}$, this construction produces simultaneously valid confidence intervals for any arbitrarily large collection of functionals of $\boldsymbol{x}$. Thus, for the one particular functional we primarily care about, it produces valid but typically excessively wide intervals that are likely to have substantial overcoverage. The idea of the method above therefore is to shrink the set D by calibrating the radius appropriately. It is suggested in [44, 45] that the appropriate radius for one-at-a-time coverage, i.e., for obtaining coverage for a single target functional, is $z_{1-\alpha/2}^2 + s^2$, where s^2 is the objective function value of program (3.2). We will use this radius throughout the rest of this paper. One of our goals will be to study the validity of this choice and in particular to illustrate that the intervals defined by (3.1) and (3.3) are indeed well calibrated in the X_{CO2} retrieval problem.

When we calculate the intervals in practice, we improve the computing time by using a simplification of (3.1)–(3.3) that allows us to replace these programs by equivalent optimization problems involving p-variate norms instead of n-variate norms; see Appendix A for details. These simplified problems are then solved using the interior-point solvers in MATLAB 2019a.

3.2.2. Dual point of view.

To gain more insight into this construction, we next look at the Lagrangian dual [5] of problems (3.1) and (3.3). When the optimal objective function value of the dual program equals that of the primal program, the problem is said to satisfy strong duality. Since programs (3.1) and (3.3) are convex, strong duality is guaranteed if the norm constraint in (3.1) and (3.3) is strictly feasible (equation (5.27) in [5]). This is true if we assume that the minimizer of the slack problem (3.2) is attained, as any such minimizer is strictly feasible for the norm constraint.

Then the lower endpoint $\underline{\theta}$ can also be obtained as the objective function value of the following program, which is derived starting from (3.1) in Appendix B:

$$\begin{aligned} &\text{maximize} && \boldsymbol{w}^T\boldsymbol{y} - \sqrt{z_{1-\alpha/2}^2 + s^2}\|\boldsymbol{w}\| - \boldsymbol{b}^T\boldsymbol{c} \\ &\text{subject to} && \boldsymbol{h} + \boldsymbol{A}^T\boldsymbol{c} - \boldsymbol{K}^T\boldsymbol{w} = \boldsymbol{0}, \\ &&& \boldsymbol{c} \geq \boldsymbol{0}, \end{aligned} \tag{3.4}$$

where the optimization is over the variables $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{c} \in \mathbb{R}^q$, with q the number of affine constraints on $\boldsymbol{x}$, and s^2 is as defined above. The upper endpoint $\overline{\theta}$ is given by an analogous program which is dual to (3.3):

$$\begin{aligned} &\text{minimize} && \boldsymbol{w}^T\boldsymbol{y} + \sqrt{z_{1-\alpha/2}^2 + s^2}\|\boldsymbol{w}\| + \boldsymbol{b}^T\boldsymbol{c} \\ &\text{subject to} && \boldsymbol{h} - \boldsymbol{A}^T\boldsymbol{c} - \boldsymbol{K}^T\boldsymbol{w} = \boldsymbol{0}, \\ &&& \boldsymbol{c} \geq \boldsymbol{0}. \end{aligned} \tag{3.5}$$

The dual perspective provides us more insight into the proposed interval $\left[\underline{\theta}, \overline{\theta}\right]$. To see this, consider the interval

$$\left[\underline{\boldsymbol{w}}^T\boldsymbol{y} - z_{1-\alpha/2}\|\underline{\boldsymbol{w}}\| - \boldsymbol{b}^T\underline{\boldsymbol{c}},\ \overline{\boldsymbol{w}}^T\boldsymbol{y} + z_{1-\alpha/2}\|\overline{\boldsymbol{w}}\| + \boldsymbol{b}^T\overline{\boldsymbol{c}}\right]. \tag{3.6}$$

As we show below and in Appendix C, if $(\underline{\boldsymbol{w}}, \underline{\boldsymbol{c}})$ and $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{c}})$ are any fixed elements of $\mathbb{R}^n \times \mathbb{R}^q$ satisfying the constraints in programs (3.4) and (3.5), respectively, then the above interval has correct coverage $(1-\alpha)$. This is true even when $\boldsymbol{K}$ is rank deficient and under the constraint $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ for $\boldsymbol{x}$. Therefore, it makes sense to find $(\underline{\boldsymbol{w}}, \underline{\boldsymbol{c}})$ and $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{c}})$ within the appropriate constraint sets such that the lower endpoint is maximized and the upper endpoint is minimized so that the overall interval is as short as possible. This optimized interval would have correct coverage if the optimized variables did not depend on $\boldsymbol{y}$, but unfortunately that is not the case here. In order to account for this optimism, it is necessary to inflate the interval to preserve coverage. The method proposed in [44, 45], and further considered here, does this by replacing $z_{1-\alpha/2}$ with $\sqrt{z_{1-\alpha/2}^2 + s^2}$, where s^2 is the slack defined above.

3.3. Method properties.

We can show the following properties for the proposed method:

- *Coverage*: The dual formulation enables us to gain some understanding of the coverage of the proposed interval. Consider a lower endpoint of the form $\underline{\theta} = \boldsymbol{w}^T\boldsymbol{y} - z_{1-\alpha/2}\|\boldsymbol{w}\| - \boldsymbol{b}^T\boldsymbol{c}$ for some fixed $\boldsymbol{w}$ and $\boldsymbol{c}$ satisfying the constraints in program (3.4). As shown in Appendix C, we can bound the miscoverage probability to obtain $\mathbb{P}_{\boldsymbol{\varepsilon}}(\underline{\theta} \geq \theta) \leq \alpha/2$. Similarly, for an upper endpoint of the form $\overline{\theta} = \boldsymbol{w}^T\boldsymbol{y} + z_{1-\alpha/2}\|\boldsymbol{w}\| + \boldsymbol{b}^T\boldsymbol{c}$, where $\boldsymbol{w}$ and $\boldsymbol{c}$ are fixed and satisfy the constraints in program (3.5), we have $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \geq \overline{\theta}) \leq \alpha/2$. Combining the two, we have $\mathbb{P}_{\boldsymbol{\varepsilon}}(\underline{\theta} \leq \theta \leq \overline{\theta}) \geq 1-\alpha$, giving the desired coverage probability. Notice, however, that when we optimize over $\boldsymbol{w}$ and $\boldsymbol{c}$, the optimized variables will depend on the observations $\boldsymbol{y}$ and the proof in Appendix C no longer holds. To account for this, the method introduces the slack factor s^2 to inflate the interval. Proving that the inflated interval has correct coverage is nontrivial since the slack s^2 itself is also a function of $\boldsymbol{y}$, but we demonstrate empirically in section 5 that the coverage is consistently very close to the desired value $(1-\alpha)$.

- *Length*: Since the optimization problems defining the interval depend on the observed data $\boldsymbol{y}$, these intervals can have variable length. Our experiments in section 5 confirm that the interval lengths indeed do vary across $\boldsymbol{y}$ realizations, but, in our experimental setup at least, the average length does not appear to change much across different $\boldsymbol{x}$.
- *Connection with classical intervals:* In the special case where $\boldsymbol{K}$ has full column rank, i.e., $\operatorname{rank}(\boldsymbol{K}) = p$, and there are no constraints on $\boldsymbol{x}$, the proposed interval reduces to the usual Gaussian standard error interval induced by the unregularized least-squares estimator of $\boldsymbol{x}$. That is, in this special case, the solutions of problems (3.1) and (3.3) yield the interval $[\hat{\theta}_{\mathrm{LS}} - z_{1-\alpha/2}\operatorname{se}(\hat{\theta}_{\mathrm{LS}}), \hat{\theta}_{\mathrm{LS}} + z_{1-\alpha/2}\operatorname{se}(\hat{\theta}_{\mathrm{LS}})]$, where $\hat{\theta}_{\mathrm{LS}} = \boldsymbol{h}^T\hat{\boldsymbol{x}}_{\mathrm{LS}}$ is the induced estimator of θ, $\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}$ is the unregularized least-squares estimator of $\boldsymbol{x}$, and $\operatorname{se}(\hat{\theta}_{\mathrm{LS}}) = \sqrt{\boldsymbol{h}^T(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{h}}$ is the standard error of $\hat{\theta}_{\mathrm{LS}}$. The proof is given in Appendix D. By standard arguments, this interval has exact $(1-\alpha)$ coverage and will have reasonable length when the mapping $\boldsymbol{x} \mapsto \theta$ acts as an implicit regularizer. In this special case, the interval has fixed length. When $\boldsymbol{K}$ is rank deficient and/or there are constraints on $\boldsymbol{x}$, the classical interval no longer applies, but the proposed interval does. The proposed interval can therefore be seen as an extension of the classical unregularized interval to these more complex situations.

3.4. Commentary.

The proposed method takes advantage of the fact that certain functionals themselves provide enough regularity so that we can retrieve them with reasonably sized confidence intervals given only objectively known physical constraints and without having to use additional subjective knowledge. This way the method avoids dependence on subjective external beliefs for the coverage guarantees. In practice, these intervals tend to be better calibrated but longer than the operational intervals which rely on such subjective knowledge. The interval length can be improved if additional objective information about the state variables is available to shrink the constraint set. This information could come either in the form of additional hard constraints or in the form of soft constraints of coverage statements for some of the unknown variables. Since this method is designed to satisfy a frequentist coverage statement, it is possible to combine these different uncertainties to obtain a valid, shorter interval in the end. These extensions are explored in section 6.

4. Existing operational retrieval procedure.

4.1. Motivation.

The existing OCO-2 operational retrieval procedure is based on a Bayesian maximum a posteriori estimator [4, 41], where the key idea is to let a prior distribution on the state vector $\boldsymbol{x}$ regularize the problem. In remote sensing literature, this approach is called “optimal estimation” [41], although optimality here depends on the choice of a cost function and typically assumes that the prior is correctly specified. We describe below the operational retrieval for our simplified setup with a linearized forward model and analyze its frequentist properties. In the actual full-physics operational retrievals with a nonlinear forward operator, finding the maximum of the posterior is a nonlinear optimization problem which is solved using the iterative Levenberg–Marquardt algorithm [4].

4.2. Method outline. The existing operational method for estimation and uncertainty quantification assumes a Gaussian prior distribution on the state vector, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, where $\boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma}_a$ are the prior mean and covariance, respectively. The posterior under this assumption and the linear forward model (2.2) is also Gaussian and is given by

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}((\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a), (\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}). \tag{4.1}$$

The point estimator $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ is chosen to be the maximizer of the posterior distribution,

$$\hat{\boldsymbol{x}} = (\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a),$$

which in our simplified setup is also the posterior mean. Recalling that $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}$, this estimator can be written as a sum of three terms, $\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\mu}_a + \boldsymbol{G}\boldsymbol{\varepsilon}$, where $\boldsymbol{G} = (\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}$ and $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{K}$ are called the retrieval gain matrix and the averaging kernel matrix, respectively [41]. The estimator for $\theta = \boldsymbol{h}^T\boldsymbol{x}$ is chosen to be the plug-in estimator $\hat{\theta} = \boldsymbol{h}^T\hat{\boldsymbol{x}}$.

To quantify the uncertainty of θ, we note that the posterior distribution on $\boldsymbol{x}$ induces a Gaussian posterior distribution on θ given by

$$\theta|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{h}^T(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a), \boldsymbol{h}^T(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1}\boldsymbol{h}). \tag{4.2}$$

A $(1-\alpha)$ central credible interval for θ is then given by

$$[\underline{\theta}, \overline{\theta}] = [\hat{\theta} - z_{1-\alpha/2}\sigma, \hat{\theta} + z_{1-\alpha/2}\sigma], \tag{4.3}$$

where $\sigma^2 = \boldsymbol{h}^T(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K}+\boldsymbol{\Sigma}_a^{-1})^{-1}\boldsymbol{h}$ is the posterior variance of θ and $\hat{\theta}$ the plug-in estimator of θ, or equivalently the maximizer/mean of $p(\theta|\boldsymbol{y})$. The credible interval (4.3) is used to quantify the uncertainty of X_{CO2} in the operational OCO-2 retrievals [4].

4.3. Frequentist properties. We describe in this section selected frequentist properties of the linearized operational retrieval method in order to compare its properties with those of our proposed method. It is straightforward to derive the following properties for the point estimator $\hat{\theta}$ and the credible interval $[\underline{\theta}, \overline{\theta}]$ given in (4.3):

- *Bias:* The bias of the estimator $\hat{\theta}$, denoted by $\text{bias}(\hat{\theta})$, can be calculated as

$$\begin{aligned}\text{bias}(\hat{\theta}) &= \mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{\theta}] - \theta = \boldsymbol{h}^T(\mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{\boldsymbol{x}}] - \boldsymbol{x}) \\ &= \boldsymbol{h}^T(\boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\mu}_a - \boldsymbol{x}) = \boldsymbol{h}^T(\boldsymbol{A} - \boldsymbol{I})(\boldsymbol{x} - \boldsymbol{\mu}_a) = \boldsymbol{m}^T(\boldsymbol{x} - \boldsymbol{\mu}_a),\end{aligned} \tag{4.4}$$

  where $\boldsymbol{m} = (\boldsymbol{A}^T - \boldsymbol{I})\boldsymbol{h} = \left(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1} - \boldsymbol{I}\right)\boldsymbol{h}$ is a vector of bias multipliers. The bias depends on $\boldsymbol{x} - \boldsymbol{\mu}_a$, i.e., the difference between the true state $\boldsymbol{x}$ and the prior mean $\boldsymbol{\mu}_a$. Notice that the bias is 0 if and only if $\boldsymbol{x} = \boldsymbol{\mu}_a$ or $\boldsymbol{m} = \boldsymbol{0}$ or if the vector $\boldsymbol{x} - \boldsymbol{\mu}_a$ is orthogonal to $\boldsymbol{m}$. In other cases, depending on $\boldsymbol{x} - \boldsymbol{\mu}_a$ and how it interacts with the forward operator $\boldsymbol{K}$, the prior covariance $\boldsymbol{\Sigma}_a$, the noise covariance $\boldsymbol{\Sigma}_\varepsilon$, and the functional $\boldsymbol{h}$, there might be a positive or a negative bias.

- *Coverage:* As shown in Appendix E, the frequentist coverage of the interval (4.3) can be written down in closed form and is given by

$$\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}]) = \Phi\left(\frac{\operatorname{bias}(\hat{\theta})}{\operatorname{se}(\hat{\theta})} + z_{1-\alpha/2}\frac{\sigma}{\operatorname{se}(\hat{\theta})}\right) - \Phi\left(\frac{\operatorname{bias}(\hat{\theta})}{\operatorname{se}(\hat{\theta})} - z_{1-\alpha/2}\frac{\sigma}{\operatorname{se}(\hat{\theta})}\right), \tag{4.5}$$

  where $\operatorname{se}(\hat{\theta}) = \sqrt{\operatorname{var}_{\boldsymbol{\varepsilon}}(\hat{\theta})}$ is the standard error of $\hat{\theta}$ and

$$\begin{aligned} \operatorname{var}_{\boldsymbol{\varepsilon}}(\hat{\theta}) &= \operatorname{var}_{\boldsymbol{\varepsilon}}(\boldsymbol{h}^T \boldsymbol{G}\boldsymbol{\varepsilon}) = \boldsymbol{h}^T \boldsymbol{G}\boldsymbol{\Sigma}_\varepsilon \boldsymbol{G}^T \boldsymbol{h} \\ &= \boldsymbol{h}^T (\boldsymbol{K}^T \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1} \boldsymbol{K}^T \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{K} (\boldsymbol{K}^T \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{K} + \boldsymbol{\Sigma}_a^{-1})^{-1} \boldsymbol{h} \end{aligned} \tag{4.6}$$

  is the variance of $\hat{\theta}$ computed with respect to the distribution of the noise $\boldsymbol{\varepsilon}$. The coverage depends on $\boldsymbol{x}$ only through $\operatorname{bias}(\hat{\theta})$. It is an even function of $\operatorname{bias}(\hat{\theta})$ and the maximum is obtained with $\operatorname{bias}(\hat{\theta}) = 0$. In that case,

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}]) &= \Phi\left(z_{1-\alpha/2}\frac{\sigma}{\operatorname{se}(\hat{\theta})}\right) - \Phi\left(-z_{1-\alpha/2}\frac{\sigma}{\operatorname{se}(\hat{\theta})}\right) \\ &> \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(-z_{1-\alpha/2}\right) = 1 - \alpha, \end{aligned}$$

  since $\sigma/\operatorname{se}(\hat{\theta}) > 1$. In other words, the interval $[\underline{\theta}, \overline{\theta}]$ has overcoverage for $\operatorname{bias}(\hat{\theta}) = 0$. It is also easy to see that the coverage is a strictly decreasing function of $|\operatorname{bias}(\hat{\theta})|$. As $|\operatorname{bias}(\hat{\theta})|$ increases, the coverage eventually crosses the nominal value $(1-\alpha)$, followed by undercoverage. In the limit $|\operatorname{bias}(\hat{\theta})| \to \infty$, the coverage becomes zero.
- *Length:* The interval $[\underline{\theta}, \overline{\theta}]$ has constant length given by $2z_{1-\alpha/2}\sigma$.
- *Comparison with standard error intervals:* A potential alternative for the credible interval (4.3) is the frequentist standard error interval

$$[\underline{\theta}, \overline{\theta}] = [\hat{\theta} - z_{1-\alpha/2}\operatorname{se}(\hat{\theta}), \hat{\theta} + z_{1-\alpha/2}\operatorname{se}(\hat{\theta})]. \tag{4.7}$$

  It is easy to show that the credible interval (4.3) is always longer than the standard error interval (4.7). This extra length can be understood as an attempt to inflate the uncertainties to account for the bias (4.4); see section 6.4 in [43] and the references therein. It follows that the coverage of the credible interval (4.3) is greater than that of the standard error interval (4.7), which undercovers whenever $\operatorname{bias}(\hat{\theta}) \neq 0$ [27].

4.4. Commentary.

The operational retrieval method is based on the well-established Bayesian framework where the observed data are combined with the prior distribution to obtain inferences in the form of the posterior distribution. The operational inferences should therefore be interpreted as representing a Bayesian degree of belief about θ. However, a user of the retrieval method may also be interested in frequentist inference of θ and the above analysis shows that the operational method can be miscalibrated if used for frequentist inference. As is well known, the performance of Bayesian methods can depend critically on the choice of the prior distribution, and the same is true for the frequentist properties of the operational method. For example, the point estimator $\hat{\theta}$ would be unbiased if the prior mean was chosen to be equal to the true state, i.e., $\boldsymbol{\mu}_a = \boldsymbol{x}$, but this is unlikely in practice as it would require knowing beforehand what the value of $\boldsymbol{x}$ is. (The bias is also small if $\boldsymbol{x} - \boldsymbol{\mu}_a$ is nearly orthogonal to $\boldsymbol{m}$, but this is equally unlikely to hold true.) At least some amount of frequentist bias is therefore always present, with the potential for arbitrarily large biases depending on how much the prior mean deviates from the true state. Since the frequentist coverage of the intervals depends on the bias, this can result in wildly varying coverage performance. For small biases, the intervals overcover, i.e., $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}]) > 1 - \alpha$, while for large biases the intervals undercover, i.e., $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}]) < 1 - \alpha$. Irrespective of which of these two cases dominates, the intervals are bound to have some degree of frequentist miscalibration since it is unlikely that there would always be just the right amount of bias for nominal coverage. Since it is impossible to judge the coverage of the intervals without knowing the true $\boldsymbol{x}$, it is not possible to tell for real soundings if a given interval is well calibrated or not. Ideally, roughly $100 \times (1 - \alpha)\%$ of soundings from a given OCO-2 orbit would cover their true X_{CO2} values. However, this discussion shows that, for the current retrieval method, this fraction can be much smaller or much larger. Furthermore, in the real atmosphere, the nearby states $\boldsymbol{x}$ are spatially and temporally correlated. Since the bias depends on $\boldsymbol{x} - \boldsymbol{\mu}_a$, this means that the biases, and therefore also the coverage values, are spatially and temporally correlated, which may lead to misleading frequentist inferences over extended spatial regions or temporal periods. These effects are analyzed in greater detail using a simulated example scenario in section 5. It is also worth noting that these suboptimal frequentist properties of the operational method are not unexpected as a Bayesian method is not necessarily designed to have good frequentist properties. Indeed, the above issues are not necessarily problematic when seen from the Bayesian perspective. It is also possible to modify a Bayesian procedure to improve its frequentist properties [1, 2, 24]; however, in this work we focus on the operational retrieval method as it is currently implemented in OCO-2.

5. Numerical results.

5.1. Experiment setup.

5.1.1. Forward model and weight vector specifics. The starting point for our forward model is the OCO-2 surrogate model developed by Hobbs et al. [19]. The surrogate model is a computationally efficient approximation to the OCO-2 full-physics forward model [4]. Similar to the full model, it involves a nonlinear mapping from the state vector $\boldsymbol{x}$ to the radiances $\boldsymbol{y}$, but is much faster to evaluate. The surrogate model also makes certain simplifications to the full OCO-2 state vector. As described in section 2.2, we make a further approximation by linearizing the surrogate model, which leads to the linear model in (2.2). The linearization is done around the generative process mean $\boldsymbol{\mu}_{\boldsymbol{x}}$; see section 5.1.2.

The state vector $\boldsymbol{x}$ in the surrogate model has 39 elements ($p = 39$), of which the first 20 correspond to the vertical CO$_2$ profile and the remaining 19 are nuisance variables related to surface pressure (x_{21}), surface albedo ($x_{22}, \ldots, x_{27}$), and atmospheric aerosol concentrations ($x_{28}, \ldots, x_{39}$). A detailed description of these variables is given in the supplementary material [38]; see also [19]. These variables suffice in order to capture, to a good approximation, the relation between the atmospheric CO$_2$ profile $x_1, \ldots, x_{20}$ and the observed radiances $\boldsymbol{y}$ [19].

In addition to the state vector $\boldsymbol{x}$, the forward operator depends on additional parameters, most notably on the solar and satellite viewing geometries, which are assumed to be known during the retrieval. In our case, the forward model is evaluated for an OCO-2 orbit that took place in October 2015 near the Total Carbon Column Observing Network (TCCON) site (36.604°N, 97.486°W) in Lamont, OK. The satellite is in the nadir observing mode, i.e., pointed toward the ground directly underneath its orbit.

We can investigate the ill-posedness of the CO_2 retrieval problem by studying the singular values of the linearized forward operator represented by the 3048×39 matrix $\boldsymbol{K}$. The singular values, shown in Figure S2 in the supplementary material [38], decay exponentially, indicating that the retrieval problem is severely ill-posed [18]. The smallest singular value deviates from the exponential decay, which we take to indicate that $\boldsymbol{K}$ is rank deficient with rank 38. Hence, there is a one-dimensional null space. The condition number (the ratio of the largest to the smallest (numerically) nonzero singular value) is 3.62×10^{12}, consistent with a severely ill-posed problem.

The ultimate quantity of interest in the retrieval problem is the column-averaged CO_2 dry-air mole fraction $X_{\mathrm{CO2}} = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$, where $\boldsymbol{h}$ is a weight vector derived in [36]; see also [4]. Since X_{CO2} only involves the CO_2 profile, the nuisance variables get weight zero, i.e., $h_{21} = \cdots = h_{39} = 0$. The remaining weights are strictly positive and sum to one, $\sum_{i=1}^{20} h_i = 1$, so statistically X_{CO2} is a weighted average of the CO_2 concentrations $x_1, \dots, x_{20}$. In the full-physics retrievals, the weights h_i have a slight dependence on the nuisance variables, but in the surrogate model the weights do not depend on the state vector. In practice, the surrogate model weights are almost constant for the intermediate pressure levels, while the weights for the boundary levels are approximately half of that value.

5.1.2. Data generation. Our investigations require a realistic generative model from which synthetic states and observations can be simulated. A suitable multivariate distribution for the state vector $\boldsymbol{x}$, as well as a model for the spatial dependence among state vectors in a small spatial region, was developed in [20]. Briefly, the approach uses actual retrieved state vectors near Lamont, OK, during the month of October 2015. This collection is part of the OCO-2 Level 2 diagnostic data products, available at the NASA Goddard Earth Science Data and Information Services Center (GES DISC, https://disc.gsfc.nasa.gov/OCO-2). These are combined with a simulation-based assessment of the retrieval error properties to estimate the state vector mean $\boldsymbol{\mu_x}$ and the single-sounding covariance $\boldsymbol{\Sigma_x}$ for this location and time.

Synthetic data are then generated through the following steps:

1. *State vector generation for a single sounding:* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})$, where the parameters of the multivariate normal distribution were estimated from OCO-2 data as noted above.
2. *State vector generation for grid sounding:* We also simulate states $\boldsymbol{x}(\boldsymbol{s}_i)$ on a grid of $i = 1, \dots, 64$ locations within an OCO-2 orbit. Following [20], we assume that this spatial process $\boldsymbol{x}(\cdot) \sim \mathrm{GP}(\boldsymbol{\mu}(\cdot), \boldsymbol{C}(\cdot,\cdot))$ is a multivariate Gaussian process with a spatially constant mean function $\boldsymbol{\mu}(\cdot) = \boldsymbol{\mu_x}$ and cross-covariance function $\boldsymbol{C}(\cdot,\cdot)$ defined as $C_{kl}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{cov}(x_k(\boldsymbol{s}_i), x_l(\boldsymbol{s}_j)) = \Sigma_{\boldsymbol{x},kl}\, \mathcal{M}_{kl}\left(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right)$, where $\mathcal{M}_{kl}$ is a Matérn-type correlation function [51]. The parameters of the correlation function vary across k and l in a way that guarantees positive definiteness and were estimated from the above collection of OCO-2 retrieved state vectors [20].

3. *Noise generation:* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. The OCO-2 radiances are fundamentally photon counts in the detectors so these measurements have Poisson-like behavior. The noise can nevertheless be approximated well using an additive Gaussian noise term with zero mean and variance proportional to the mean signal. Following [19] and [30], we let $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ be diagonal with elements $\mathrm{var}(\varepsilon_j) = c_{b(j)} F_j(\boldsymbol{\mu}_{\boldsymbol{x}})$, where $b : \{1, \ldots, 3048\} \to \{1, 2, 3\}$, $j \mapsto b(j)$ indicates the spectral band (O$_2$, weak CO$_2$, strong CO$_2$) of j, c_i are band-specific constants, and $F_j(\cdot)$ is the jth element of the forward operator output. In the actual satellite, the noise model is somewhat more complicated, but its properties are nevertheless well understood [11]. The $\boldsymbol{\varepsilon}$ realizations are independent and identically distributed both across repetitions of the experiment for a fixed state $\boldsymbol{x}$ and over the different spatial sounding locations.
4. *Radiance observation:* $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ are given by the previous steps and the matrix $\boldsymbol{K}$ results from linearizing the forward operator $\boldsymbol{F}$ about the true mean $\boldsymbol{\mu}_{\boldsymbol{x}}$.

In addition, the operational procedure posits a prior distribution on the state $\boldsymbol{x}$, which is given by $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$. We use the prior mean $\boldsymbol{\mu}_a$ and prior covariance $\boldsymbol{\Sigma}_a$ derived from the OCO-2 operational prior near Lamont, OK, in October 2015. For OCO-2, the prior mean $\boldsymbol{\mu}_a$ varies in space and time but is dependent in part on climatology and expert knowledge, while $\boldsymbol{\Sigma}_a$ is the same for all retrievals.

An important point to highlight is that $\boldsymbol{\mu}_a \neq \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_a \neq \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Therefore, the true conditions, represented through $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ in our simulations, will be different from the prior mean and covariance. This misspecification is a real challenge for the operational retrievals and a source of bias [34]. The prior model and the generative model are visualized and compared in detail in the supplementary material [38], which also contains a visualization of the spatial dependence structure of the state vectors.

5.1.3. Constraints.

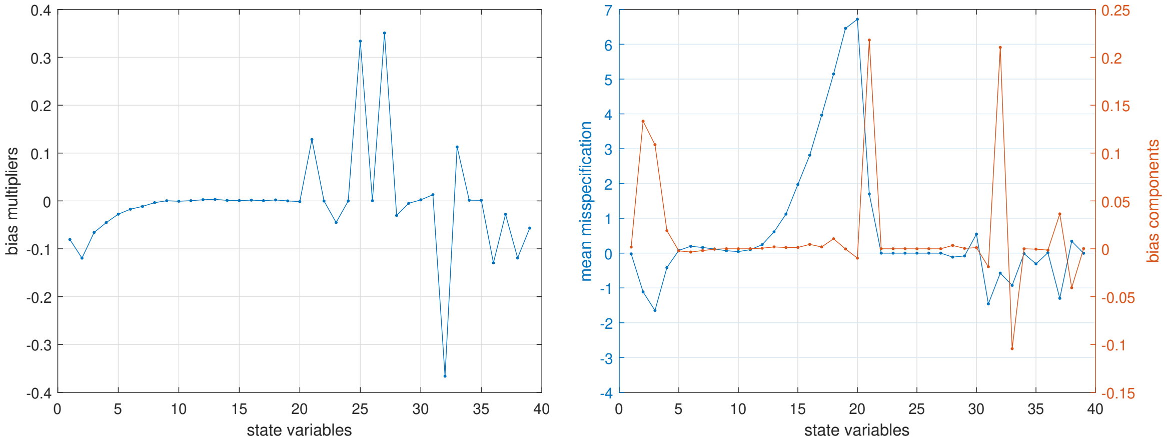
In the proposed frequentist procedure, we impose nonnegativity constraints on certain elements of the state vector $\boldsymbol{x}$. Since elements $x_1, \ldots, x_{20}$ are CO$_2$ concentrations, they need to be nonnegative by definition. Thus, we impose the constraint $x_i \geq 0$ for $i = 1, \ldots, 20$. The same argument applies to surface pressure, so we also include the constraint $x_{21} \geq 0$. The rest of the state vector elements are left unconstrained.

Since albedo is a fraction between 0 and 1, this implies in principle linear inequality constraints for the albedo variables $x_{22}, \ldots, x_{27}$. We experimented with adding these constraints but found that that made little difference in the results while causing some extra computational overhead. We therefore decided to leave these variables unconstrained. The aerosol variables $x_{28}, \ldots, x_{39}$ are parameterized in the surrogate model in such a way that there are no trivial constraints that could be imposed on those variables.

5.2. Single sounding results.

5.2.1. Distribution of bias of the operational method.

Since the linearized operational method is based on a linear estimator $\hat{\theta}$, we can write down $\mathrm{bias}(\hat{\theta}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{\theta}] - \theta$ in closed form for a given $\boldsymbol{x}$. This is done in (4.4), which shows that the bias is given by the inner product of the bias multiplier vector $\boldsymbol{m}$ and the prior mean misspecification $\boldsymbol{x} - \boldsymbol{\mu}_a$. For a given $\boldsymbol{x}$ sampled from the generative model, there will therefore always be a nonzero bias whose size



(a) Bias multipliers

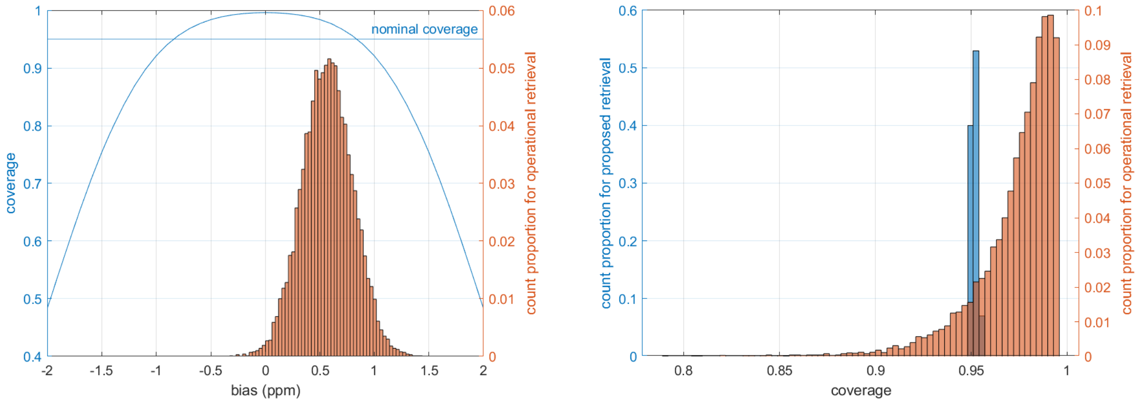
(b) Bias components

Figure 2. *Figure* (a) *illustrates the bias multiplier vector* $\boldsymbol{m}$, *while* (b) *shows the corresponding bias components* $m_i(\mu_{\boldsymbol{x},i} - \mu_{a,i})$ *for the misspecified means.*

depends on the details of the prior misspecification for that particular $\boldsymbol{x}$. To understand this interaction better, we show the bias multiplier vector $\boldsymbol{m}$ in Figure 2(a) for our particular retrieval setup. This highlights the role of the nuisance variables $x_{21}, \ldots, x_{39}$ in dictating the size of the bias. Notice that the bias multiplier $\boldsymbol{m}$ depends on the prior covariance $\boldsymbol{\Sigma}_a$ but not on the prior mean $\boldsymbol{\mu}_a$. Hence this can be seen as a way of decoupling the contribution of the prior mean on the bias from that of the prior covariance. It is also worth noting that here the bias is entirely caused by the regularization in the prior since we generate the data using the same linear forward model $\boldsymbol{K}$ that we use in the inversion; in real-life retrievals, there might be an additional component in the bias from the nonlinearity of the forward operator.

Assuming that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ gives a realistic distribution of $\boldsymbol{x}$'s for repeated satellite overpasses, we can also derive the distribution of the bias over repeated $\boldsymbol{x}$ realizations. In our particular case, we have $\text{bias}(\hat{\theta}) \sim \mathcal{N}(0.5714, 0.0533)$. This distribution is illustrated in Figure 3(a) showing the bias for 10 000 instances of $\boldsymbol{x}$ from the generative model. This shows that the biases are typically positive with a fair amount of spread around the central value. Negative biases and biases larger than 1.2 ppm are rare, at least in this particular setup for the retrieval problem.

We have that $\mathbb{E}_{\boldsymbol{x}}[\text{bias}(\hat{\theta})] = \boldsymbol{m}^T(\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\mu}_a)$, which corresponds to the bias expressions given in [34]. Hence, the distribution of $\text{bias}(\hat{\theta})$ has mean zero if and only if $\boldsymbol{\mu}_a = \boldsymbol{\mu}_{\boldsymbol{x}}$ or if $\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\mu}_a$ is orthogonal to $\boldsymbol{m}$. Even in those cases, $\text{bias}(\hat{\theta})$ would still have a spread around zero, so individual retrievals may be positively or negatively biased. In the more realistic case where $\boldsymbol{m}^T(\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\mu}_a) \neq 0$, the biases are either predominantly positive or negative depending on the details of the prior misspecification. Figure 2(b) shows a breakdown of the contribution of each state variable to the mean bias of 0.5714 in our particular setup. The figure visualizes the mean misspecification $\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{\mu}_a$ and the individual terms $m_i(\mu_{\boldsymbol{x},i} - \mu_{a,i})$ contributing to the mean bias. It enables us to conclude that the positive biases are primarily caused by the misspecification of the surface pressure variable x_{21}, the aerosol variable x_{32}, and the upper



(a) Bias distribution

(b) Coverage distribution

Figure 3. *Figure* (a) *shows the coverage as a function of bias (blue line) for the operational procedure and the corresponding histogram of operational retrieval bias. Figure* (b) *shows a histogram (in orange) of the operational retrieval coverage for* 95% *intervals. Also shown is a histogram (in blue) of empirical coverage for the proposed frequentist uncertainty quantification method.*

portion of the CO_2 profile, all of which contribute positively to the mean bias. The large misspecification of the lower portion of the CO_2 profile, on the other hand, makes a negligible contribution to the bias due to the small bias multipliers of those variables.

5.2.2. Coverage and length of the operational and proposed intervals.

The frequentist coverage of the operational method for a particular $\boldsymbol{x}$ defined as $\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta \in [\underline{\theta}, \overline{\theta}])$ can be calculated using (4.5). The coverage depends on $\boldsymbol{x}$ only through $\text{bias}(\hat{\theta})$. To understand the nature of this dependence, we plot in Figure 3(a) the coverage of 95% intervals as a function of the bias for our particular retrieval setup. We observe that for $|\text{bias}(\hat{\theta})| < 0.84$ ppm the intervals overcover, while for $|\text{bias}(\hat{\theta})| > 0.84$ ppm the intervals undercover, with the coverage dropping sharply for biases larger than 1 ppm in absolute value. Since our biases are predominantly positive (Figure 3(a)), we are mostly going to observe the subrange of coverages corresponding to $\text{bias}(\hat{\theta}) \in [-0.3 \text{ ppm}, 1.4 \text{ ppm}]$.

It is difficult to explicitly write down the distribution of the coverage corresponding to the assumed distribution of $\boldsymbol{x}$, but we evaluate the coverage distribution numerically in Figure 3(b) for 10 000 state vectors and 95% intervals. We see that the operational intervals are poorly calibrated in terms of their frequentist coverage. For most $\boldsymbol{x}$ realizations, the intervals have overcoverage. However, we also note that the coverage distribution is heavily left-skewed toward values below the nominal 95% coverage. In particular, for 12.03% of the $\boldsymbol{x}$ realizations, the intervals have undercoverage. The smallest coverage is 79.0%, and this could drop even lower depending on the $\boldsymbol{x}$ realization.

Such coverage behavior is inherent to the operational retrieval method because of the bias induced by the regularizing prior. This leads to the somewhat paradoxical conclusion that, if interpreted as frequentist confidence intervals, the operational intervals are too long for most $\boldsymbol{x}$'s, while for roughly 12% of the $\boldsymbol{x}$'s, the intervals are too short. Unfortunately, there is

Table 1

Comparison of coverage and interval length (in ppm) between the operational and proposed uncertainty quantification methods for 10 *state vector* $\boldsymbol{x}$ *realizations chosen uniformly between the minimum and maximum coverage for the operational method in Figure* 3(b)*. The target coverage in each case is* 95%*. Also shown are the bias of the operational point estimates and the standard deviation of the length of the proposed intervals (both in ppm). The proposed method is not based on a point estimator, so we do not report a bias value for it.*

$\boldsymbol{x}$ realization	Operational bias	Operational coverage	Operational length	Proposed coverage	Proposed avg. length	Proposed length s.d.
1	1.4173	0.7899	3.94	0.9515	11.20	0.29
2	1.3707	0.8090	3.94	0.9511	11.20	0.28
3	1.2986	0.8363	3.94	0.9510	11.20	0.29
4	1.2357	0.8579	3.94	0.9515	11.20	0.28
5	1.1590	0.8816	3.94	0.9513	11.20	0.28
6	1.0747	0.9042	3.94	0.9512	11.21	0.27
7	0.9721	0.9272	3.94	0.9515	11.20	0.29
8	0.8420	0.9500	3.94	0.9513	11.19	0.31
9	0.6477	0.9730	3.94	0.9508	11.19	0.32
10	0.0001	0.9959	3.94	0.9502	11.18	0.35

no easy way of telling when the intervals are too long or too short, so it is not possible to adaptively recalibrate their length.

The proposed frequentist direct retrieval method, on the other hand, has fundamentally different behavior. For this method, it is not straightforward to write down a closed-form expression for the coverage, but we can nevertheless evaluate it empirically. Here we evaluate the empirical coverage of 95% intervals using 10 000 realizations of the noise $\boldsymbol{\varepsilon}$. This is repeated for 100 realizations of $\boldsymbol{x}$ from the generative model to study the distribution of the coverage values. The results are shown in Figure 3(b). We find that the proposed method is well-calibrated across all considered $\boldsymbol{x}$ instances. The coverage peaks at slightly above 95%, with very little spread around that value. For some $\boldsymbol{x}$, the intervals have a small amount of overcoverage, but this is very minor in comparison to the operational method.

To further compare the two methods, we pick 10 instances of $\boldsymbol{x}$ corresponding to 10 different coverage values for the operational method ranging from the minimum operational coverage to the maximum in Figure 3(b). Table 1 compares the 95% intervals for the two methods for each of these 10 state vectors. We observe that while the coverage of the operational method can vary between substantial undercoverage and major overcoverage, the proposed method consistently achieves nearly nominal coverage irrespective of the $\boldsymbol{x}$ realization.

The two approaches also have different behaviors in terms of their interval lengths. The operational intervals have constant length $2z_{1-\alpha/2}\sigma$, where σ is the posterior standard deviation of θ that does not depend on the data $\boldsymbol{y}$. In our case, $\sigma = 1.0051$ ppm, so the operational intervals have constant length of 3.94 ppm at 95% confidence level. It is worth noting that $\mathrm{se}(\hat{\theta}) = 0.6856$ ppm. Hence, the operational intervals derived from the posterior of θ are almost 50% longer than what standard error intervals would be. This extra length gives the operational intervals some, but not enough, protection against undercoverage.

The proposed intervals, on the other hand, have data-dependent length. We report in Table 1 the average lengths and length standard deviations for these intervals across different

$\boldsymbol{\varepsilon}$ realizations for each fixed state vector $\boldsymbol{x}$. We observe that the interval lengths indeed vary across noise realizations with a coefficient of variation (ratio of standard deviation to average length) of about 3%. However, the average lengths are almost constant across different $\boldsymbol{x}$'s. We therefore conclude that while the proposed intervals have variable length, their average length does not seem to depend much on the true state $\boldsymbol{x}$.

The proposed intervals are longer than the operational ones, but in exchange seem to provide the desired coverage irrespective of the $\boldsymbol{x}$ realization. We can gain further insight into the behavior of the two constructions by visualizing some illustrative realizations of the state vectors and the associated intervals produced by the two methods. This is done in section S3 in the supplementary material [38].

5.3. Coverage over a spatial region.

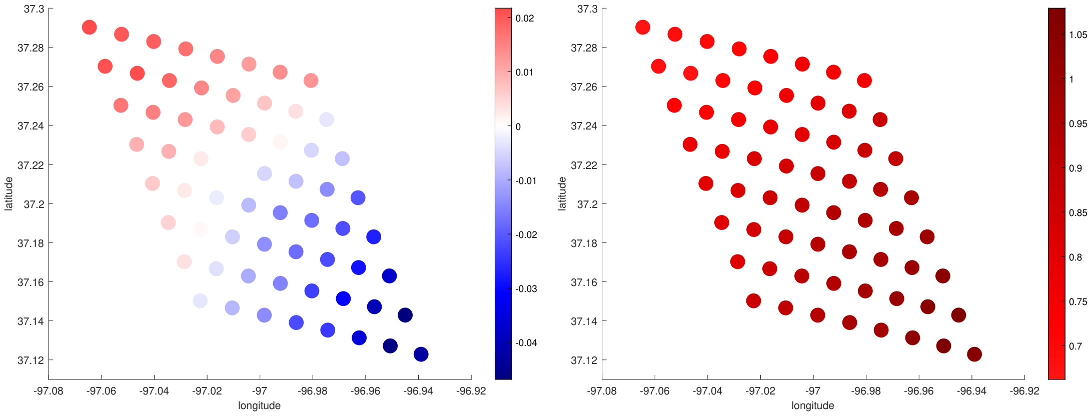
In this section, we investigate the performance of the operational method over a spatial grid of 8×8 soundings near Lamont, OK. The size of the region is approximately 8 km in the cross-track direction and 16 km in the along-track direction. We generate spatially correlated state vectors $\boldsymbol{x}(\boldsymbol{s}_i)$ and expected radiances $\boldsymbol{K}\boldsymbol{x}(\boldsymbol{s}_i)$ over the grid as described in section 5.1.2. We then investigate the bias of the operational point estimates and the pointwise coverage of the operational 95% intervals over the grid. To do this, we can simply use the closed-form expressions (4.4) and (4.5) at each sounding location. Since both the bias and the coverage of the operational method are functions of the state vectors $\boldsymbol{x}(\boldsymbol{s}_i)$, these properties inherit the spatial dependence between the state vectors and will hence exhibit spatially correlated patterns.

The observed patterns depend on the specific $\{\boldsymbol{x}(\boldsymbol{s}_i) : i = 1, \dots, 64\}$ realization over the grid. Figure 4(a) shows the coverage pattern for a case in which the coverage systematically changes from overcoverage to undercoverage when moving from the northwest corner of the grid to the southeast corner. The reason for this can be seen from the bias pattern in Figure 4(b), which shows that the overall positive bias has a systematic gradient across the region so that the bias is larger in the southeast corner and smaller in the northwest corner, which then affects the coverage as described in (4.5). It is also possible to observe undercoverage over the entire grid. Figure S9 in the supplementary material [38] shows the coverage and bias patterns for a case where the state vector realizations are such that all 64 intervals across the region have coverage below the nominal value due to a systematic large positive bias throughout the region.

These results illustrate one of the challenges of the operational retrieval method in that there can be entire regions with undercoverage or overcoverage. For example, in the case of Figure S9, all intervals across the region are systematically offset toward too large X_{CO2} values, which causes their lower bounds to miss the corresponding true X_{CO2} values more often than they should. There is a risk that such patterns could be mistaken as CO$_2$ flux signals. As such, these observations may have important implications for carbon flux estimates; see section 7. The coverage patterns shown here for the operational method are in contrast with the behavior of the proposed frequentist method, which does not exhibit systematic spatially correlated miscalibration.

6. Variable importance and effect of additional constraints on interval length.

As demonstrated in the previous section, the proposed frequentist method has good coverage performance, but the intervals are longer than in the operational retrieval method. In this



(a) Operational coverage (b) Operational bias

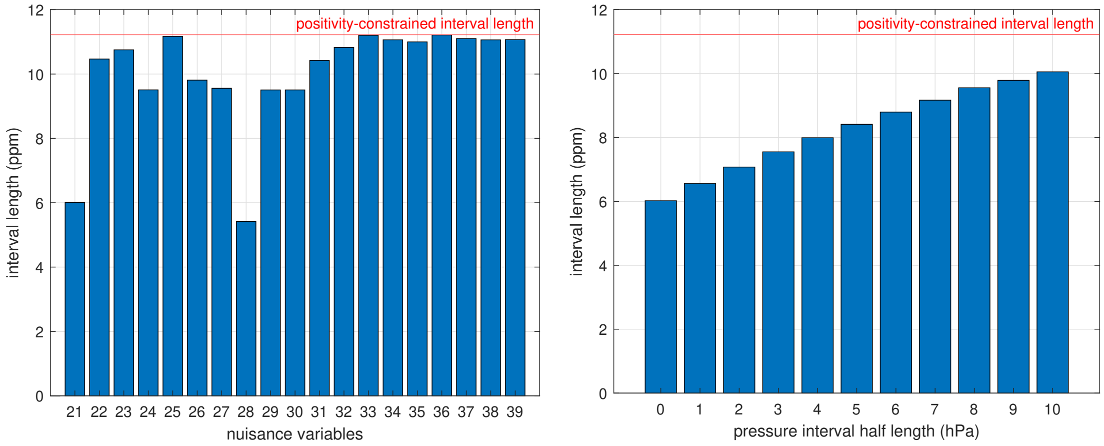
Figure 4. *Operational retrieval over a grid of* 8×8 *soundings for an instance where both undercoverage and overcoverage are present. Figure* (a) *shows the spatial coverage pattern relative to the nominal* 95% *in units of probability (i.e.,* -0.03, *for example, corresponds to coverage* 0.92, *instead of the nominal* 0.95*). The fraction of soundings below nominal coverage is* 0.55. *Figure* (b) *shows the corresponding bias pattern in ppm.*

section, we consider different variants of the proposed method to improve the interval length.

It is worth noting that the underlying inference problem is defined by (i) the forward operator $\boldsymbol{K}$, (ii) the functional of interest parameterized by the weight vector $\boldsymbol{h}$, (iii) the amount of noise in the problem controlled by the covariance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and (iv) the constraint set C. For a given sounding and quantity of interest, we cannot change $\boldsymbol{K}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, or $\boldsymbol{h}$, but we can potentially alter C. We therefore investigate how changes in C in the form of additional constraints affect the length of the proposed intervals.

6.1. Effect of individual nuisance variables. We have so far only used trivial positivity constraints on certain state vector elements; see section 5.1.3. However, if additional information were available to further constrain the state vector—for example, one could imagine that observational data from other sources tell us that, for some i, $x_i \in [\underline{x}_i, \overline{x}_i]$ with high probability—then including those constraints should result in shorter intervals for X_{CO2}.

To investigate what impact this would have, we start by considering how each nuisance variable $x_{21}, \ldots, x_{39}$ affects the final X_{CO2} interval length. Figure 5(a) shows the average interval lengths for the proposed method when one of the nuisance variables is assumed to be known. We can incorporate this assumption by using constraints of the form $x_{i,\mathrm{true}} \leq x_i \leq x_{i,\mathrm{true}}$ for one x_i at a time, in addition to the previously used positivity constraints. As expected, the interval lengths are smaller than the interval length without any additional information. In particular, variables x_{21} (surface pressure) and x_{28} (log aerosol optical depth for the first composite aerosol type; see the supplementary material [38]) have the greatest impact on the interval length. Therefore additional constraints on these two variables could be particularly helpful in reducing the interval length. Since it is not immediately clear what observational constraints might be available for x_{28}, we will in the following focus on constraints for the pressure variable x_{21}.



(a) Nuisance variable importance

(b) Pressure importance

Figure 5. *Figure* (a) *shows average* X_{CO2} *interval lengths at* 95% *confidence level when constraining one nuisance variable at a time to its true value. Figure* (b) *shows average* X_{CO2} *interval lengths at* 95% *confidence level for varying degrees of deterministic constraints on the surface pressure variable* x_{21}. *In both figures, the horizontal red line shows the interval length when only trivial positivity constraints are used.*

6.2. Deterministic pressure constraints.

We analyze the effect of various degrees of deterministic constraints on the pressure variable in Figure 5(b). Instead of assuming that the pressure is known exactly, as was done in Figure 5(a), we consider symmetric constraints about the true pressure value, i.e., constraints of the form $x_{21,\mathrm{true}} - \delta \leq x_{21} \leq x_{21,\mathrm{true}} + \delta$ for various δ. As expected, we observe that tighter constraints on pressure translate into shorter intervals for X_{CO2}. For example, knowing the pressure to within ±3 hPa lets us decrease the average X_{CO2} interval length from 11.19 ppm to 7.55 ppm. Knowing the surface pressure to within such, or even higher, accuracy is not implausible as there are other, complementary observing systems, such as ground-based weather stations, that are capable of providing pressure information within such limits.

We remark that the interval lengths in Figures 5(a) and 5(b) are averages over 100 noise realizations for the $\boldsymbol{x}$ realization corresponding to nominal operational coverage in Table 1; see also Figure S8 in the supplementary material [38]. We have also studied other $\boldsymbol{x}$'s from the generative model and found qualitatively similar results.

6.3. Probabilistic constraints and interval length optimization.

We analyzed above the effect of additional deterministic constraints on pressure. However, such constraints might not always be available with full certainty; instead, we might know that they hold with high probability. This is the case, for example, when a frequentist confidence interval is available from another observing system. In this section, we show how to incorporate such probabilistic constraints within the proposed method while still maintaining finite-sample coverage.

6.3.1. Coverage calibration.

To explain the key idea, imagine that instead of having deterministic constraints such as $x_{i,\mathrm{true}} - \delta \leq x_i \leq x_{i,\mathrm{true}} + \delta$, we have confidence intervals for one or more of the x_i's such that $\underline{x}_i(\alpha_i) \leq x_i \leq \overline{x}_i(\alpha_i)$ with frequentist coverage at least $(1-\alpha_i)$, i.e., $\mathbb{P}(x_i \in [\underline{x}_i(\alpha_i), \overline{x}_i(\alpha_i)]) \geq 1 - \alpha_i$. We can then construct a $(1-\alpha)$ confidence interval for the quantity of interest θ by running the proposed retrieval procedure with these probabilistic constraints at an internal confidence level $(1-\gamma)$ chosen so that, accounting for the α_i's, we can still maintain the required nominal coverage. As shown in Appendix F, we can bound the miscoverage probability for the quantity of interest as follows:

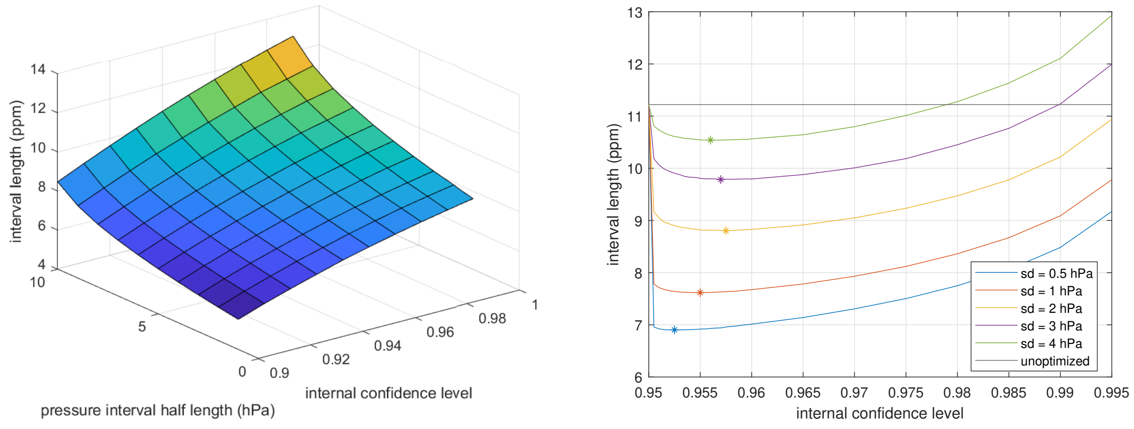
$$\mathbb{P}(\theta \notin [\underline{\theta}, \overline{\theta}]) \leq \gamma + \sum\nolimits_i \alpha_i, \tag{6.1}$$

where i ranges over those variables that have probabilistic constraints. Thus, if we choose γ and the α_i's in such a way that $\gamma + \sum_i \alpha_i = \alpha$, then we can keep the desired $(1-\alpha)$ coverage for the interval $[\underline{\theta}, \overline{\theta}]$.

6.3.2. Demonstration with pressure intervals. Since deterministic constraints on the pressure variable x_{21} provided a gain in the interval length, we now analyze the effect of probabilistic constraints on that variable. This demonstrates the simplest application of (6.1) in a case where there is a probabilistic constraint on a single variable only. We therefore need to choose γ and α_{21} such that $\gamma + \alpha_{21} = \alpha$, where we set $\alpha = 0.05$ to obtain a 95% final interval for X_{CO2}. By (6.1), any positive γ and α_{21} summing to 0.05 will give a valid final interval, but an optimal choice is such that it minimizes the final interval length. To start investigating the dependence of the X_{CO2} interval length on these choices, Figure 6(a) shows how the length of the pressure interval and the confidence level $(1-\gamma)$ of the X_{CO2} interval jointly affect the average X_{CO2} interval length. Using Figure 6(a), we can set the internal confidence level $(1-\gamma)$ to a value larger than 95% to account for the coverage probability $(1-\alpha_{21})$ of the pressure interval. To optimize this choice, we need to relate α_{21} to the length of the pressure interval. In this study, we assume that there is a pressure sensor that provides pressure observations $\hat{x}_{21}$ following the Gaussian distribution $\mathcal{N}(x_{21}, \sigma_{21}^2)$. We then assume that the pressure intervals are $(1-\alpha_{21})$ standard error intervals of length $2z_{1-\alpha_{21}/2}\sigma_{21}$, where σ_{21} is the pressure standard error.

For a given σ_{21}, we can then trace through Figure 6(a) for various γ and the corresponding α_{21} and record the final X_{CO2} interval length. Figure 6(b) shows examples of this for pressure standard errors σ_{21} ranging from 0.5 hPa to 4 hPa. Each curve represents the average interval length for the proposed method as a function of γ and can be used to choose γ such that the final interval length is optimized. Along each curve, we have indicated this optimal internal confidence level. We observe that for moderate values of the pressure standard error, the optimal internal confidence level is greater than 95%, but when the pressure standard error is either very small or very large, the optimal internal confidence level approaches 95%. This happens because when the pressure standard error is very small, it is almost as good as using the exact pressure value, while when the standard error is very large, it is almost as good as not using any additional constraints on pressure besides the nonnegativity constraint.

Since the proposed interval has variable, data-dependent length, there is an important subtlety in that the above interval length optimization must be done without using the observed data $\boldsymbol{y}$ so as to guarantee the coverage in (6.1). In addition, we would ideally like to optimize the average interval length, which cannot be done based on a single $\boldsymbol{y}$. We therefore need a candidate state vector $\boldsymbol{x}$ that can be used to calculate average interval lengths which are then used as the basis for the length optimization. Since we found in section 5 that the average



(a) Joint effect of pressure and internal confidence level

(b) Interval length optimization

Figure 6. *Figure* (a) *shows the joint effect of pressure interval length and internal confidence level on the average* X_{CO2} *interval length. Figure* (b) *shows the optimization of the average* X_{CO2} *interval length for various pressure standard errors by trading off the internal confidence level* $(1-\gamma)$ *for the confidence level of the pressure interval* $(1-\alpha_{21})$*. The values that optimize the interval length are marked with asterisks.*

interval lengths are not very sensitive to the choice of $\boldsymbol{x}$, it suffices to have a reasonable ansatz for $\boldsymbol{x}$. Luckily, we already have that in the prior mean $\boldsymbol{\mu}_a$ of the operational method. In this study, we therefore set the state vector $\boldsymbol{x}$ equal to $\boldsymbol{\mu}_a$ for the interval length optimization. The interval lengths shown in Figures 6(a) and 6(b) were obtained as averages over 100 noise realizations for this choice of $\boldsymbol{x}$ and for pressure intervals centered at $x_{21} = \mu_{a,21}$.

We now proceed to empirically verify the coverage of the final intervals constructed as described above. The length optimization phase can be run based on $\boldsymbol{\mu}_a$ and σ_{21} before seeing $\boldsymbol{y}$. This leads to an optimal choice of γ and α_{21} irrespective of $\boldsymbol{y}$, and, fixing these values, one can then check the coverage and length of the intervals for multiple $\boldsymbol{y}$ realizations corresponding to a fixed $\boldsymbol{x}$. We use the $\boldsymbol{x}$ that provides nominal coverage for the operational retrieval method in Table 1 as the state vector for this evaluation. While the length optimization was done without fluctuating the pressure intervals, the coverage study also accounts for the variation of the pressure intervals by simulating intervals of the form $[\hat{x}_{21} - z_{1-\alpha_{21}/2}\sigma_{21}, \hat{x}_{21} + z_{1-\alpha_{21}/2}\sigma_{21}]$ for the optimized α_{21} and for $\hat{x}_{21} \sim \mathcal{N}(x_{21}, \sigma_{21}^2)$ independently of $\boldsymbol{\varepsilon}$. The results are given in Table 2, which shows the optimized confidence levels $(1-\gamma)$ and $(1-\alpha_{21})$ as well as the empirical coverage and average length of the final X_{CO2} intervals based on 10 000 realizations. We observe that the intervals maintain the 95% coverage guarantee while significantly reducing the final interval length. The amount of gain provided by the pressure information depends on the level of uncertainty in the pressure intervals. In particular, for pressure standard error of 0.5 hPa, we are able to reduce the average interval length to 6.89 ppm from the original 11.19 ppm. The final intervals are somewhat conservative due to the slack in the inequality in (6.1). Notice also that the interval lengths predicted by the prior-based optimization in Figure 6(b) match well with the final values in Table 2 even though this evaluation is for a different $\boldsymbol{x}$.

Table 2

Optimized confidence levels, final empirical coverage and average length for the X_{CO2} intervals incorporating probabilistic pressure constraints at various levels of pressure standard error. The internal and pressure confidence levels are chosen so that the final interval has at least 95% coverage.

Pressure std. err. (hPa)	Internal conf. level $(1-\gamma)$	Pressure conf. level $(1-\alpha_{21})$	Final empirical coverage	Interval length (ppm)
0.5	0.9525	0.9975	0.9741	6.89
1	0.9550	0.9950	0.9782	7.61
2	0.9575	0.9925	0.9743	8.80
3	0.9570	0.9930	0.9684	9.71
4	0.9560	0.9940	0.9629	10.32

7. Conclusions and outlook.

Our focus on the frequentist properties of the uncertainty estimates is one of the main differences between this work and much of the other related work on uncertainty quantification in remote sensing, which tends to predominantly focus on Bayesian construction and evaluation of uncertainties. The frequentist and Bayesian paradigms answer fundamentally different questions about the unknown parameter θ, and as is well known from the extensive discussion in the literature (see [54, 3, 50, 46, 17] for references specific to inverse problems), both approaches are valuable in their own right. The question we set out to answer is the following: Given a fixed state of the atmosphere corresponding to a given satellite overpass, what are the repeated sampling properties of the uncertainty intervals when the repetitions are over the instrument noise $\boldsymbol{\varepsilon}$? Hence, most of our probabilities and expectations are taken with respect to the noise $\boldsymbol{\varepsilon}$, while some previous works take expectations over both $\boldsymbol{\varepsilon}$ and $\boldsymbol{x}$ [19, 34]. In the case of the operational retrieval, our studies constitute a frequentist evaluation of the underlying Bayesian procedure [1]. Arguably, properties calculated with respect to $\boldsymbol{\varepsilon}$ are potentially more relevant for downstream scientific use, where, for example, carbon flux estimates use OCO-2 data to gain information about the instantaneous state of the atmosphere corresponding to a particular $\boldsymbol{x}$ instead of an average $\boldsymbol{x}$.

It is important to clarify that there is a difference between frequentist and Bayesian criteria for evaluating uncertainties and frequentist and Bayesian constructions of uncertainties. Indeed, some Bayesian constructs can have desirable frequentist properties, while some frequentist constructs may have unexpectedly poor frequentist properties. The results in this paper show that the standard operational retrieval procedure does not fall in the former category, but alternative Bayesian constructs might have improved frequentist properties [2, 24]. Similarly, standard frequentist approaches to ill-posed inverse problems may have poor frequentist coverage performance [28, 27, 29]. For example, a variant of penalized maximum likelihood (or, equivalently, Tikhonov regularization or ridge regression) would have exactly the same point estimator $\hat{\theta}$ as the operational retrieval but with uncertainty quantified using the standard error interval (4.7) instead of the credible interval (4.3). The resulting interval has coverage always less than $(1-\alpha)$ [27, section 6.4.2]. In this sense, the operational Bayesian retrieval has better frequentist performance than this alternative frequentist construct (see [35, 42] for a similar observation in spline smoothing). The difference in the frequentist performance of the two methods considered in this paper is therefore less about the difference between frequentist and Bayesian constructs and more about the difference between explicit and implicit regularization. The proposed method achieves good frequentist calibration because it is implicitly regularized by the functional and the constraints, while the operational retrieval has poor calibration because of the explicit regularization from the prior, and the same conclusion would be true for other explicitly regularized methods. Similarly, it might be possible to obtain an implicitly regularized Bayesian construction by considering a uniform or nearly uniform vague prior consistent with the available physical constraints.

To interpret the frequentist X_{CO2} intervals, it is crucial to understand that the $(1-\alpha)$ coverage property holds not only for a collection of intervals from a given sounding location, but also for a collection of intervals arising from soundings at different locations, since the noise $\boldsymbol{\varepsilon}$ is independent across soundings. Imagine a collection of, say, 10 000 sounding locations within an OCO-2 orbit, each with a realization of a 95% frequentist confidence interval for X_{CO2}. Then we know that roughly 9 500 of these intervals cover their true X_{CO2} values, and the coverage/noncoverage pattern should not have any apparent spatial structure. It is foreseeable that such intervals could be used to produce rigorous uncertainties in downstream scientific tasks by, for example, using techniques similar to those described here for X_{CO2}. Our grid sounding experiments show that the same conclusion does not necessarily hold for the operational retrievals. Let I_k be the indicator random variable indicating whether the kth interval covers its true X_{CO2} value, where k ranges over the spatial sounding locations within the orbit. For well-calibrated frequentist intervals, the I_k's are independent and identically distributed across the sounding locations, while in the case of the operational retrievals, the I_k's are independent across the sounding locations, but no longer identically distributed. Instead, the coverage probability $\mathbb{P}_{\boldsymbol{\varepsilon}}(I_k = 1)$ varies throughout the orbit in a systematic, spatially coherent way, so that in some parts of the orbit perhaps 85% of the intervals are expected to cover, while in other parts maybe 99% of the intervals cover. Since, in the absence of oracle information, there is no straightforward way of telling which of these situations applies in a given region, it is not immediately clear how to properly use such uncertainties in downstream scientific tasks.

An important question for future work is to understand what implications these conclusions have on CO_2 flux estimates. A key question concerns the spatial length scales at which the biases occur in operational X_{CO2} point estimates. Our results indicate that there are spatially correlated biases at least at scales of 8×8 soundings (roughly 8 km × 16 km), which is likely to have implications for regional carbon flux estimates, for example, over urban areas. This conclusion is further corroborated by OCO-2's target mode observations taken on orbits near TCCON sites to assess the empirical behavior of OCO-2 retrievals for individual overpasses [59]. Indeed, the retrieval errors for a single target overpass have been found to exhibit substantial spatial correlation [60]. However, as of now, we do not know whether these bias patterns persist at the scale of a single pixel in global flux inversion models, where the grid resolution is typically of the order of a few hundred kilometers. If they do, then it would be useful to understand how to incorporate our proposed intervals, which do not exhibit spatially correlated offsets, into these models.

An important insight provided by this work is the identification of the surface pressure (x_{21}) and the aerosol optical depth of the first composite aerosol type (x_{28}) as key variables affecting the length of the proposed intervals (Figure 5(a)). This raises the interesting possibility of obtaining more precise X_{CO2} estimates by developing Level 2 retrieval methods that combine pressure or aerosol information from other satellites or observing systems with OCO-2 data. The surface pressure also plays an important role in explaining the performance of the operational retrievals (see Figures S6–S8 in the supplementary material [38]), which has also been noted in previous studies; see [25] and the references therein. A more comprehensive analysis of the effects of the different variables, including at different spatial regions, seasons, or observing modes, is left as a subject for future work.

In this paper, we have considered a linearized approximation of the nonlinear OCO-2 forward operator. A major topic for future work would be to extend this work to nonlinear forward operators. The basic primal approach from [47], outlined in section 3.2.1, still applies in that the extremal values of $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}$ over $\boldsymbol{x} \in C \cap D$ would still define valid $(1-\alpha)$ simultaneous confidence intervals. What is not immediately clear, however, is whether the approach from [44, 45] for turning these into one-at-a-time intervals still applies. Another major challenge concerns the computation of the intervals since now D can no longer be described by a quadratic inequality and might even be nonconvex, depending on the properties of the forward operator. Constructing and characterizing the dual problems would also be substantially more difficult. Nevertheless, since here $\boldsymbol{x}$ has a moderate dimension, it is plausible that methods can be developed for solving the primal optimization problems within reasonable time constraints. One potential approach would be to successively linearize the nonlinear part of the programs within an iterative quadratic programming algorithm.

We have shown empirically that the proposed intervals consistently have frequentist coverage very close to the nominal value. In future work, we hope to be able to show what conditions are needed to rigorously guarantee this. As has been pointed out in [55], the previous proof in [45] appears to be incorrect. The authors in [55] even provide a counterexample showing that the intervals can undercover for $\boldsymbol{h}$ containing both positive and negative elements. This leaves open the question of whether it is possible to guarantee the coverage when all elements of $\boldsymbol{h}$ have the same sign, as is the case here with the X_{CO2} functional. If it turns out to be difficult to provide such guarantees, it might be possible to consider alternative definitions of the slack factor s^2 so that coverage and other theoretical properties can be proved more easily.

While they have much better frequentist calibration, the proposed intervals are almost three times as long as the current operational intervals, when only trivial constraints are applied. Therefore, an important challenge for future work would be to understand what can be said about the optimality of the length of these intervals within the class of methods that provide frequentist coverage guarantees. Donoho introduced in [12] intervals that are up to a multiplicative factor minimax optimal for this problem among the class of fixed-length intervals with guaranteed coverage. The intervals studied here are variable length and may hence be shorter than those of [12]. To the best of our knowledge, minimax optimality of variable-length intervals for this setting is an open problem. Furthermore, instead of minimax, a more appropriate notion of optimality here might be one with respect to a reasonable distribution on $\boldsymbol{x}$, such as the operational prior distribution.

Appendices. Appendices A–D assume that the forward model has been transformed to have identity covariance, i.e., $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{Kx}, \boldsymbol{I})$, as described in section 3.2.

Appendix A. Computational simplification.

Consider again the original problem in the primal form to obtain the lower endpoint $\underline{\theta}$:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{Kx}\|^2 \leq z_{1-\alpha/2}^2 + s^2, \\ & \boldsymbol{Ax} \leq \boldsymbol{b}, \end{array} \tag{A.1}$$

where $s^2 = \min_{\boldsymbol{x}\,:\,\boldsymbol{Ax}\leq\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{Kx}\|^2$.

Let us consider the singular value decomposition $\boldsymbol{K} = \boldsymbol{UDV}^T$. We first note that $\|\boldsymbol{y} - \boldsymbol{Kx}\|^2 = \|\boldsymbol{U}^T\boldsymbol{y} - \boldsymbol{DV}^T\boldsymbol{x}\|^2$, since $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{UU}^T = \boldsymbol{I}$. Further, as $n > p$, let us denote by $\tilde{\boldsymbol{y}}_{1:p}$ the first p entries of $\tilde{\boldsymbol{y}} = \boldsymbol{U}^T\boldsymbol{y}$ and by $\tilde{\boldsymbol{y}}_{p+1:n}$ the rest of the entries of $\tilde{\boldsymbol{y}}$. Then we can write $\|\boldsymbol{U}^T\boldsymbol{y} - \boldsymbol{DV}^T\boldsymbol{x}\|^2 = \|\tilde{\boldsymbol{y}}_{1:p} - \boldsymbol{D}_{1:p,:}\boldsymbol{V}^T\boldsymbol{x}\|^2 + \|\tilde{\boldsymbol{y}}_{p+1:n}\|^2$, where $\boldsymbol{D}_{1:p,:}$ denotes the first p rows of $\boldsymbol{D}$. This suggests a simplification of the primal problem where, instead of (A.1), we solve the following equivalent problem to obtain the lower endpoint $\underline{\theta}$:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \|\tilde{\boldsymbol{y}}_{1:p} - \boldsymbol{D}_{1:p,:}\boldsymbol{V}^T\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + \tilde{s}^2, \\ & \boldsymbol{Ax} \leq \boldsymbol{b}, \end{array} \tag{A.2}$$

where now $\tilde{s}^2 = \min_{\boldsymbol{x}\,:\,\boldsymbol{Ax}\leq\boldsymbol{b}} \|\tilde{\boldsymbol{y}}_{1:p} - \boldsymbol{D}_{1:p,:}\boldsymbol{V}^T\boldsymbol{x}\|^2$. This is equivalent to the original problem because $s^2 = \tilde{s}^2 + \|\tilde{\boldsymbol{y}}_{p+1:n}\|^2$. When $n \gg p$, solving problem (A.2), including the associated slack problem, is much faster than solving problem (A.1), because the norms involve p-variate vectors instead of n-variate vectors. An analogous simplification can obviously be used with the upper endpoint $\overline{\theta}$ as well. These simplifications proved crucial for our ability to perform the empirical coverage studies presented in this paper.

Appendix B. Dual derivation.

Consider the primal optimization problem to obtain the lower endpoint $\underline{\theta}$:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{Kx}\|^2 \leq z_{1-\alpha/2}^2 + s^2, \\ & \boldsymbol{Ax} \leq \boldsymbol{b}, \end{array} \tag{B.1}$$

where s^2 is the slack factor. For notational convenience, we let $z_{1-\alpha/2}^2 + s^2 = q^2$.

We first write an equivalent problem as follows:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{y} - \boldsymbol{Kx} = \boldsymbol{r}, \\ & \|\boldsymbol{r}\|^2 \leq q^2, \\ & \boldsymbol{Ax} \leq \boldsymbol{b}, \end{array} \tag{B.2}$$

where the optimization is now over both $\boldsymbol{x}$ and $\boldsymbol{r}$.

The Lagrangian of the above problem can be written as

$$L(\boldsymbol{x},\boldsymbol{r},\boldsymbol{w},\lambda,\boldsymbol{c}) = \boldsymbol{h}^T\boldsymbol{x} + \boldsymbol{w}^T(\boldsymbol{y}-\boldsymbol{K}\boldsymbol{x}-\boldsymbol{r}) + \lambda(\|\boldsymbol{r}\|^2 - q^2) + \boldsymbol{c}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}), \tag{B.3}$$

where $\boldsymbol{w}$, $\lambda \geq 0$ and $\boldsymbol{c} \geq \boldsymbol{0}$ are dual variables [5].

The dual function is obtained by minimizing the Lagrangian with respect to the primal variables $\boldsymbol{x}$ and $\boldsymbol{r}$:

$$g(\boldsymbol{w},\lambda,\boldsymbol{c}) = \inf_{\boldsymbol{x},\boldsymbol{r}} L(\boldsymbol{x},\boldsymbol{r},\boldsymbol{w},\lambda,\boldsymbol{c}). \tag{B.4}$$

We first rewrite the Lagrangian to group the terms corresponding to $\boldsymbol{x}$ and $\boldsymbol{r}$ together:

$$L(\boldsymbol{x},\boldsymbol{r},\boldsymbol{w},\lambda,\boldsymbol{c}) = (\boldsymbol{h} - \boldsymbol{K}^T\boldsymbol{w} + \boldsymbol{A}^T\boldsymbol{c})^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{r} + \lambda\|\boldsymbol{r}\|^2 + \boldsymbol{w}^T\boldsymbol{y} - \lambda q^2 - \boldsymbol{c}^T\boldsymbol{b}. \tag{B.5}$$

Next, we note that we can restrict ourselves to the case where $\boldsymbol{h} - \boldsymbol{K}^T\boldsymbol{w} + \boldsymbol{A}^T\boldsymbol{c} = \boldsymbol{0}$, since otherwise the Lagrangian is unbounded below as a linear function in $\boldsymbol{x}$. By minimizing with respect to $\boldsymbol{r}$ and substituting back, we obtain the dual function

$$g(\boldsymbol{w},\lambda,\boldsymbol{c}) = -\frac{1}{4\lambda}\|\boldsymbol{w}\|^2 + \boldsymbol{w}^T\boldsymbol{y} - \lambda q^2 - \boldsymbol{c}^T\boldsymbol{b}, \tag{B.6}$$

where $\boldsymbol{h} - \boldsymbol{K}^T\boldsymbol{w} + \boldsymbol{A}^T\boldsymbol{c} = \boldsymbol{0}$, $\lambda \geq 0$, and $\boldsymbol{c} \geq \boldsymbol{0}$. The dual optimization problem is then the problem of maximizing $g(\boldsymbol{w},\lambda,\boldsymbol{c})$ with respect to the dual variables $\boldsymbol{w}$, λ, and $\boldsymbol{c}$. Maximization with respect to λ can be carried out in closed form. We can thus eliminate λ to obtain the following dual problem for the remaining variables $\boldsymbol{w}$ and $\boldsymbol{c}$:

$$\begin{aligned} &\text{maximize} && \boldsymbol{w}^T\boldsymbol{y} - \sqrt{z_{1-\alpha/2}^2 + s^2}\,\|\boldsymbol{w}\| - \boldsymbol{b}^T\boldsymbol{c} \\ &\text{subject to} && \boldsymbol{h} + \boldsymbol{A}^T\boldsymbol{c} - \boldsymbol{K}^T\boldsymbol{w} = \boldsymbol{0}, \\ &&& \boldsymbol{c} \geq \boldsymbol{0}, \end{aligned} \tag{B.7}$$

which gives us (3.4). The dual problem (3.5) corresponding to the upper endpoint $\overline{\theta}$ follows from an analogous derivation.

Appendix C. Coverage for fixed dual variables. Assume $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}, \boldsymbol{I})$ with functional of interest $\theta = \boldsymbol{h}^T\boldsymbol{x}$ and state vector $\boldsymbol{x}$ satisfying $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$. Consider a lower endpoint of the form $\underline{\theta} = \boldsymbol{w}^T\boldsymbol{y} - z_{1-\alpha/2}\|\boldsymbol{w}\| - \boldsymbol{b}^T\boldsymbol{c}$ for some fixed $\boldsymbol{w}$ and $\boldsymbol{c}$ satisfying the constraints in program (3.4). We can bound the miscoverage probability as follows:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\varepsilon}}(\underline{\theta} \geq \theta) &= \mathbb{P}_{\boldsymbol{\varepsilon}}(\boldsymbol{w}^T\boldsymbol{y} - z_{1-\alpha/2}\|\boldsymbol{w}\| - \boldsymbol{b}^T\boldsymbol{c} \geq \theta) \\ &= \mathbb{P}_{\boldsymbol{\varepsilon}}(\boldsymbol{w}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x} - z_{1-\alpha/2}\|\boldsymbol{w}\| \geq \boldsymbol{h}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{c}) \\ &\overset{(1)}{\leq} \mathbb{P}_{\boldsymbol{\varepsilon}}(\boldsymbol{w}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x} - z_{1-\alpha/2}\|\boldsymbol{w}\| \geq \boldsymbol{h}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{A}\boldsymbol{x}) \\ &\overset{(2)}{=} \mathbb{P}_{\boldsymbol{\varepsilon}}(\boldsymbol{w}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x} - z_{1-\alpha/2}\|\boldsymbol{w}\| \geq 0) \overset{(3)}{=} \alpha/2, \end{aligned}$$

where (1) follows from the constraints $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ and $\boldsymbol{c} \geq \boldsymbol{0}$; (2) uses the fact that $\boldsymbol{w}$ and $\boldsymbol{c}$ need to satisfy the constraint $\boldsymbol{h} + \boldsymbol{A}^T\boldsymbol{c} - \boldsymbol{K}^T\boldsymbol{w} = \boldsymbol{0}$; and (3) follows from $\boldsymbol{w}^T\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{K}\boldsymbol{x}, \|\boldsymbol{w}\|^2)$ for any fixed $\boldsymbol{w}$. Thus, for fixed $\boldsymbol{w}$ and $\boldsymbol{c}$ that satisfy the constraints, we have $\mathbb{P}_{\boldsymbol{\varepsilon}}(\underline{\theta} \geq \theta) \leq \alpha/2$.

Appendix D. Simplification with full column rank and no constraints.

We prove in this section that when $\operatorname{rank}(\boldsymbol{K}) = p$ and there are no external constraints on $\boldsymbol{x}$, the solutions of problems (3.1) and (3.3) yield the interval $[\hat{\theta}_{\mathrm{LS}} - z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}_{\mathrm{LS}}), \hat{\theta}_{\mathrm{LS}} + z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}_{\mathrm{LS}})]$, where $\hat{\theta}_{\mathrm{LS}} = \boldsymbol{h}^T\hat{\boldsymbol{x}}_{\mathrm{LS}}$ is the plug-in estimator of θ, $\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}$ is the unregularized least-squares estimator of $\boldsymbol{x}$, and $\mathrm{se}(\hat{\theta}_{\mathrm{LS}}) = \sqrt{\boldsymbol{h}^T(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{h}}$ is the standard error of $\hat{\theta}_{\mathrm{LS}}$.

Consider the lower endpoint of the interval. In the absence of external constraints on $\boldsymbol{x}$, the optimization problem (3.1) reduces to

$$\text{(D.1)} \qquad \begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + s^2, \end{array}$$

where the slack factor s^2 is now defined as the objective function value of the corresponding unconstrained least-squares problem:

$$\text{(D.2)} \qquad s^2 = \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2.$$

Since we assume that $\boldsymbol{K}$ has full column rank, the solution to the above problem is exactly $\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}$. Plugging in this value of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$, we obtain that the squared slack is given by the residual sum of squares

$$\text{(D.3)} \qquad s^2 = \|\boldsymbol{y} - \boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}\|^2.$$

We can then write the constraint in problem (D.1) as follows:

$$\text{(D.4)} \qquad \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 \leq z_{1-\alpha/2}^2 + \|\boldsymbol{y} - \boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}\|^2.$$

We can further manipulate the difference

$$\text{(D.5)} \qquad \begin{aligned} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|^2 - \|\boldsymbol{y} - \boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}\|^2 &= \boldsymbol{x}^T\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{x} - 2\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y} \\ &= \|\boldsymbol{x} - \hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{K}^T\boldsymbol{K}}, \end{aligned}$$

where we have used the weighted-norm notation $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}$, to arrive at the following program for the lower endpoint of the interval:

$$\text{(D.6)} \qquad \begin{array}{ll} \text{minimize} & \boldsymbol{h}^T\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x} - \hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{K}^T\boldsymbol{K}} \leq z_{1-\alpha/2}^2. \end{array}$$

We proceed to show that the optimal value of this problem is given by $\hat{\theta}_{\mathrm{LS}} - z_{1-\alpha/2}\mathrm{se}\,(\hat{\theta}_{\mathrm{LS}})$. We begin by writing down the Karush–Kuhn–Tucker (KKT) conditions [5] of the problem. The Lagrangian of the problem is given by

$$\text{(D.7)} \qquad L(\boldsymbol{x}, \lambda) = \boldsymbol{h}^T\boldsymbol{x} + \lambda\left(\|\boldsymbol{x} - \hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{K}^T\boldsymbol{K}} - z_{1-\alpha/2}^2\right),$$

where $\lambda \geq 0$ is a dual variable. The KKT conditions for the primal and dual optimal pair $(\boldsymbol{x}^\star, \lambda^\star)$ are thus

$$\text{(D.8)} \qquad \boldsymbol{h} = -2\lambda^\star\boldsymbol{K}^T\boldsymbol{K}(\boldsymbol{x}^\star - \hat{\boldsymbol{x}}_{\mathrm{LS}}),$$

using first-order optimality with respect to $\boldsymbol{x}$, along with

$$\lambda^\star\left(\|\boldsymbol{x}^\star-\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{K}^T\boldsymbol{K}}-z^2_{1-\alpha/2}\right)=0 \tag{D.9}$$

from the complementary slackness condition. We find that $\lambda^\star=\frac{1}{2z_{1-\alpha/2}}\|\boldsymbol{h}\|_{(\boldsymbol{K}^T\boldsymbol{K})^{-1}}\geq 0$ and $\boldsymbol{x}^\star=\hat{\boldsymbol{x}}_{\mathrm{LS}}-\frac{z_{1-\alpha/2}}{\|\boldsymbol{h}\|_{(\boldsymbol{K}^T\boldsymbol{K})^{-1}}}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{h}$ satisfy the KKT conditions and therefore provide a primal-dual optimal pair. Substituting the value of $\boldsymbol{x}^\star$ into the objective, we arrive at the desired lower endpoint. The upper endpoint results from a similar argument.

Appendix E. Coverage of Gaussian central credible intervals. This section provides derivation of the frequentist coverage of the credible interval (4.3) used in the operational retrievals. Since $\hat{\theta}$ is an affine transformation of $\boldsymbol{y}$, it is Gaussian with mean $\mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{\theta}]$ and variance $\mathrm{var}_{\boldsymbol{\varepsilon}}(\hat{\theta})$. Therefore, $(\hat{\theta}-\theta-\mathrm{bias}(\hat{\theta}))/\mathrm{se}(\hat{\theta})$ has standard Gaussian distribution. Then

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\varepsilon}}(\theta\in[\underline{\theta},\overline{\theta}]) &= \mathbb{P}_{\boldsymbol{\varepsilon}}(\hat{\theta}-z_{1-\alpha/2}\sigma\leq\theta\leq\hat{\theta}+z_{1-\alpha/2}\sigma)\\
&= \mathbb{P}_{\boldsymbol{\varepsilon}}(-z_{1-\alpha/2}\sigma\leq\hat{\theta}-\theta\leq z_{1-\alpha/2}\sigma)\\
&= \mathbb{P}_{\boldsymbol{\varepsilon}}\left(-\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}-z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\leq\frac{\hat{\theta}-\theta-\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}\leq-\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}+z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\right)\\
&= \Phi\left(-\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}+z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\right)-\Phi\left(-\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}-z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\right)\\
&= \Phi\left(\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}+z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\right)-\Phi\left(\frac{\mathrm{bias}(\hat{\theta})}{\mathrm{se}(\hat{\theta})}-z_{1-\alpha/2}\frac{\sigma}{\mathrm{se}(\hat{\theta})}\right),
\end{aligned}$$

using $\Phi(x)=1-\Phi(-x)$ to obtain the last equality. This establishes (4.5).

Appendix F. Miscoverage probability with probabilistic constraints. Given the setup in section 6.3.1, we can bound the error probabilities as follows:

$$\begin{aligned}
\mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]) &= \mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,,x_i\in[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for all }i)\\
&\qquad+\mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,,x_i\notin[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for some }i)\\
&= \mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,|\,x_i\in[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for all }i)\cdot\mathbb{P}(x_i\in[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for all }i)\\
&\qquad+\mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,,x_i\notin[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for some }i)\\
&\leq \mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,|\,x_i\in[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for all }i)\\
&\qquad+\mathbb{P}(x_i\notin[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for some }i)\\
&\leq \mathbb{P}(\theta\notin[\underline{\theta},\overline{\theta}]\,|\,x_i\in[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)]\text{ for all }i)\\
&\qquad+\sum\nolimits_i\mathbb{P}(x_i\notin[\underline{x}_i(\alpha_i),\overline{x}_i(\alpha_i)])\\
&\leq \gamma+\sum\nolimits_i\alpha_i, \qquad\qquad (\mathrm{F}.1)
\end{aligned}$$

where i ranges over those variables that have probabilistic constraints.

We note that when this framework is used to incorporate probabilistic constraints on multiple variables, using the union bound to control the miscoverage probability, as is done in (F.1), might be loose and additional structure among the probabilistic constraints, such as independence, could provide additional gain.

Acknowledgments. We are grateful to Amy Braverman, Mike Gunson, Ryan Tibshirani, Sivaraman Balakrishnan, as well as the participants of the OCO-2 March 2019 Uncertainty Quantification Breakout Meeting, the Statistical and Applied Mathematical Sciences Institute Spatial X Working Group, and the Carnegie Mellon University Statistical Methods for the Physical Sciences Research Group for helpful discussions and feedback at various stages throughout this work. We would also like to thank the Associate Editor and the two anonymous referees for their insightful feedback that helped us improve various aspects of this work.

REFERENCES

[1] M. J. Bayarri and J. O. Berger, *The interplay of Bayesian and frequentist analysis*, Statist. Sci., 19 (2004), pp. 58–80, https://doi.org/10.1214/088342304000000116.

[2] J. Berger, *The case for objective Bayesian analysis*, Bayesian Anal., 1 (2006), pp. 385–402, https://doi.org/10.1214/06-BA115.

[3] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, *Large-Scale Inverse Problems and Quantification of Uncertainty*, John Wiley & Sons, 2011.

[4] H. Boesch, et al., *Orbiting Carbon Observatory-2 & 3: Level 2 Full Physics Retrieval Algorithm Theoretical Basis*, NASA Jet Propulsion Laboratory, OCO D-55207, Version 2.0, Rev. 3, 2019.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] B. Connor, et al., *Quantification of uncertainties in OCO-2 measurements of XCO2: Simulations and linear error analysis*, Atmos. Meas. Tech., 9 (2016), pp. 5227–5238, https://doi.org/10.5194/amt-9-5227-2016.

[7] B. J. Connor, H. Boesch, G. Toon, B. Sen, C. Miller, and D. Crisp, *Orbiting Carbon Observatory: Inverse method and prospective error analysis*, J. Geophys. Res. Atmos., 113 (2008), https://doi.org/10.1029/2006JD008336.

[8] N. Cressie, *Mission CO$_2$ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide*, J. Amer. Statist. Assoc., 113 (2018), pp. 152–168.

[9] N. Cressie, R. Wang, M. Smyth, and C. E. Miller, *Statistical bias and variance for the regularized inverse problem: Application to space-based atmospheric CO$_2$ retrievals*, J. Geophys. Res. Atmos., 121 (2016), pp. 5526–5537.

[10] D. Crisp, et al., *The Orbiting Carbon Observatory (OCO) mission*, Adv. Space Res., 34 (2004), pp. 700–709.

[11] D. Crisp, et al., *Orbiting Carbon Observatory-2 & 3: Level 1B Algorithm Theoretical Basis*, NASA Jet Propulsion Laboratory, OCO-2 D-55206, Version 2.0, Rev. 0, 2021.

[12] D. L. Donoho, *Statistical estimation and optimal recovery*, Ann. Statist., 22 (1994), pp. 238–270.

[13] A. Eldering, C. W. O'Dell, P. O. Wennberg, D. Crisp, M. Gunson, C. Viatte, C. Avis, A. Braverman et al., *The Orbiting Carbon Observatory-2: First 18 months of science data products*, Atmos. Meas. Tech., 10 (2017), pp. 549–563, https://doi.org/10.5194/amt-10-549-2017.

[14] A. Eldering, T. E. Taylor, C. W. O'Dell, and R. Pavlick, *The OCO-3 mission: Measurement objectives and expected performance based on 1 year of simulated data*, Atmos. Meas. Tech., 12 (2019), pp. 2341–2370.

[15] A. Eldering, P. Wennberg, D. Crisp, D. Schimel, M. Gunson, A. Chatterjee, J. Liu, F. Schwandner, Y. Sun, C. O'dell et al., *The Orbiting Carbon Observatory-2 early science investigations of regional carbon dioxide fluxes*, Science, 358 (2017), eaam5745.

[16] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, 2000.

[17] S. N. Evans and P. B. Stark, *Inverse problems as statistics*, Inverse Probl., 18 (2002), R55.

[18] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, 2005, https://doi.org/10.1137/1.9780898719697.

[19] J. Hobbs, A. Braverman, N. Cressie, R. Granat, and M. Gunson, *Simulation-based uncertainty quantification for estimating atmospheric CO$_2$ from satellite data*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 956–985, https://doi.org/10.1137/16M1060765.

[20] J. Hobbs, M. Katzfuss, D. Zilber, J. Brynjarsdóttir, A. Mondal, and V. Berrocal, *Spatial retrievals of atmospheric carbon dioxide from satellite observations*, Remote Sens., 13 (2021), 571, https://doi.org/10.3390/rs13040571.

[21] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[22] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, J. Comput. Phys., 296 (2015), pp. 348–368.

[23] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005.

[24] R. E. Kass and L. Wasserman, *The selection of prior distributions by formal rules*, J. Amer. Statist. Assoc., 91 (1996), pp. 1343–1370, https://doi.org/10.1080/01621459.1996.10477003.

[25] M. Kiel, C. W. O'Dell, B. Fisher, A. Eldering, R. Nassar, C. G. MacDonald, and P. O. Wennberg, *How bias correction goes wrong: Measurement of X_{CO2} affected by erroneous surface pressure estimates*, Atmos. Meas. Tech., 12 (2019), pp. 2241–2259.

[26] S. Kulawik, et al., *Consistent evaluation of ACOS-GOSAT, BESD-SCIAMACHY, CarbonTracker, and MACC through comparisons to TCCON*, Atmos. Meas. Tech., 9 (2016), pp. 683–709, https://doi.org/10.5194/amt-9-683-2016.

[27] M. Kuusela, *Uncertainty Quantification in Unfolding Elementary Particle Spectra at the Large Hadron Collider*, Ph.D. thesis, EPFL, 2016.

[28] M. Kuusela and V. M. Panaretos, *Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification*, Ann. Appl. Statist., 9 (2015), pp. 1671–1705.

[29] M. Kuusela and P. B. Stark, *Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra*, Ann. Appl. Statist., 11 (2017), pp. 1671–1710.

[30] O. Lamminpää, J. Hobbs, J. Brynjarsdóttir, M. Laine, A. Braverman, H. Lindqvist, and J. Tamminen, *Accelerated MCMC for satellite-based measurements of atmospheric CO_2*, Remote Sens., 11 (2019), https://doi.org/10.3390/rs11172061.

[31] M. Maahn, D. D. Turner, U. Löhnert, D. J. Posselt, K. Ebell, G. G. Mace, and J. M. Comstock, *Optimal estimation retrievals and their uncertainties: What every atmospheric scientist should know*, Bull. Amer. Meteorol. Inst., 101 (2020), pp. E1512–E1523, https://doi.org/10.1175/BAMS-D-19-0027.1.

[32] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487, https://doi.org/10.1137/110845598.

[33] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[34] H. Nguyen, N. Cressie, and J. Hobbs, *Sensitivity of optimal estimation satellite retrievals to misspecification of the prior mean and covariance, with application to OCO-2 retrievals*, Remote Sens., 11 (2019), 2770, https://doi.org/10.3390/rs11232770.

[35] D. Nychka, *Confidence intervals for smoothing splines*, J. Amer. Statist. Assoc., 83 (1988), pp. 1134–1143.

[36] C. W. O'Dell, et al., *The ACOS CO_2 retrieval algorithm - Part 1: Description and validation against synthetic observations*, Atmos. Meas. Tech., 5 (2012), pp. 99–121, https://doi.org/10.5194/amt-5-99-2012.

[37] C. W. O'Dell, et al., *Improved retrievals of carbon dioxide from Orbiting Carbon Observatory-*2 *with the version* 8 *ACOS algorithm*, Atmos. Meas. Tech., 11 (2018), pp. 6539–6576, https://doi.org/10.5194/amt-11-6539-2018.

[38] P. Patil, M. Kuusela, and J. Hobbs, *Supplement to "Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals,"* 2022.

[39] A. K. Ramanathan, H. M. Nguyen, X. Sun, J. Mao, J. B. Abshire, J. M. Hobbs, and A. J. Braverman, *A singular value decomposition framework for retrievals with vertical distribution information from greenhouse gas column absorption spectroscopy measurements*, Atmos. Meas. Tech., 11 (2018), pp. 4909–4928.

[40] P. J. Rayner and D. M. O'Brien, *The utility of remotely sensed CO_2 concentration data in surface source inversions*, Geophys. Res. Lett., 28 (2001), pp. 175–178.

[41] C. D. Rodgers, *Inverse Methods for Atmospheric Sounding: Theory and Practice*, World Scientific, 2000.

[42] D. Ruppert and R. J. Carroll, *Spatially-adaptive penalties for spline fitting*, Aust. N. Z. J. Stat., 42 (2000), pp. 205–223.

[43] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression,* Cambridge University Press, 2003.

[44] B. W. Rust and W. R. Burrus, *Mathematical Programming and the Numerical Solution of Linear Equations*, American Elsevier, 1972.

[45] B. W. Rust and D. P. O'Leary, *Confidence intervals for discrete approximations to ill-posed problems*, J. Comput. Graph. Stat., 3 (1994), pp. 67–96.

[46] J. A. Scales and L. Tenorio, *Prior information and uncertainty in inverse problems*, Geophys., 66 (2001), pp. 389–397, https://doi.org/10.1190/1.1444930.

[47] P. B. Stark, *Inference in infinite-dimensional inverse problems: Discretization and duality*, J. Geophys. Res. Solid Earth, 97 (1992), pp. 14055–14082.

[48] P. B. Stark (1995). *Simultaneous Confidence Intervals for Linear Estimates of Linear Functionals,* Tech. Report 417, University of California, Berkeley.

[49] P. B. Stark, *Constraints versus priors*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 586–598, https://doi.org/10.1137/130920721.

[50] P. B. Stark and L. Tenorio, *A primer of frequentist and Bayesian inference in inverse problems*, in Large-Scale Inverse Problems and Quantification of Uncertainty, L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, eds., John Wiley & Sons, 2011, pp. 9–32.

[51] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, 1999.

[52] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[53] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, 2005, https://doi.org/10.1137/1.9780898717921.

[54] L. Tenorio, *An Introduction to Data Analysis and Uncertainty Quantification for Inverse Problems*, SIAM, 2017, https://doi.org/10.1137/1.9781611974928.

[55] L. Tenorio, A. Fleck, and K. Moses, *Confidence intervals for linear discrete inverse problems with a non-negativity constraint*, Inverse Problems, 23 (2007), pp. 669–681.

[56] D. van Dyk, A. Connors, D. N. Esch, P. Freeman, H. Kang, M. Karovska, V. Kashyap, A. Siemiginowska, and A. Zezas, *Deconvolution in high-energy astrophysics: Science, instrumentation, and methods*, Bayesian Anal., 1 (2006), pp. 189–235, https://doi.org/10.1214/06-BA107.

[57] I. S. Weir, *Fully Bayesian reconstructions from single-photon emission computed tomography data*, J. Amer. Statist. Assoc., 92 (1997), pp. 49–60, https://doi.org/10.1080/01621459.1997.10473602.

[58] L. Wu, O. Hasekamp, H. Hu, J. Landgraf, A. Butz, J. aan de Brugh, I. Aben, D. F. Pollard, D. W. T. Griffith, D. G. Feist, D. Koshelev, F. Hase, G. C. Toon, H. Ohyama, I. Morino, J. Notholt, K. Shiomi, L. Iraci, M. Schneider, M. de Mazière, R. Sussmann, R. Kivi, T. Warneke, T.-Y. Goo, and Y. Té, *Carbon dioxide retrieval from OCO-2 satellite observations using the RemoTeC algorithm and validation with TCCON measurements*, Atmos. Meas. Tech., 11 (2018), pp. 3111–3130, https://doi.org/10.5194/amt-11-3111-2018.

[59] D. Wunch, et al., *Comparisons of the Orbiting Carbon Observatory-2 (OCO-2) X_{CO_2} measurements with TCCON*, Atmos. Meas. Tech., 10 (2017), pp. 2209–2238, https://doi.org/10.5194/amt-10-2209-2017.

[60] B. Zhang, N. Cressie, and D. Wunch, *Inference for errors-in-variables models in the presence of spatial and temporal dependence with an application to a satellite remote sensing campaign*, Technometrics, 61 (2019), pp. 187–201, https://doi.org/10.1080/00401706.2018.1476268.

Supplement to "Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals"

Pratik Patil[1,2], Mikael Kuusela[1] and Jonathan Hobbs[3]

[1]Department of Statistics and Data Science, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Jet Propulsion Laboratory, California Institute of Technology

S1 Supplement to Section 2.1

Figure S1 shows an example sounding for OCO-2.

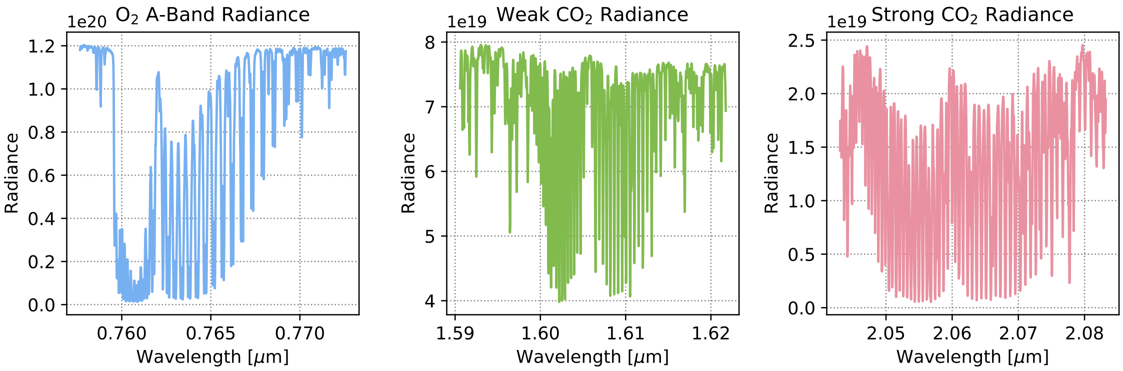


Figure S1: Example sounding from OCO-2. A sounding includes 1016 radiances in each of the three infrared spectral bands.

S2 Supplement to Section 5.1

S2.1 Description of state vector elements

The specific elements of the state vector $\boldsymbol{x} \in \mathbb{R}^{39}$ are:

- Variables $x_1, \dots, x_{20}$ are the dry-air mole fractions of atmospheric CO_2, i.e., the number of moles of CO_2 per one mole of dry air, in parts per million (ppm) at 20 fixed pressure levels. In the sequel, we denote these as levels 1 to 20, with level 1 being highest in the atmosphere (pressure ~ 0.1 hPa) and level 20 being the surface.

- Variable x_{21} is the surface air pressure in hPa. It corresponds to the total weight of the air molecules in the atmospheric column.

- Variables $x_{22}, \dots, x_{27}$ relate to surface albedo, i.e., the fraction of total incoming solar radiation reflected off the Earth's surface. Albedo varies across the three OCO-2 spectral bands and also within each band. In the surrogate model, albedo is modeled as a linear function within each spectral band; see Section B.2 in [1]. The albedo for each band is therefore parameterized by an intercept and a slope which enter the state vector as nuisance variables (x_{22}, x_{24}, and x_{26} are the three intercepts and x_{23}, x_{25}, and x_{27} are the slopes).

- Variables $x_{28}, \ldots, x_{39}$ parameterize the atmospheric aerosol concentrations and distributions. The surrogate model assumes that there are 4 aerosol types, which have distinct absorption and scattering properties. The first two are location-dependent composite species. For our investigation, these are sulfate and dust. The latter two are two cloud species, one for ice clouds and another for liquid water clouds [2, 1]. Each type is parameterized by 3 parameters corresponding to the aerosol optical depth (AOD, i.e., the opaqueness of the aerosol species measured as the natural logarithm of the ratio of incoming to transmitted radiation) as well as the altitude and thickness of each aerosol layer. Variables x_{28}, x_{31}, x_{34}, and x_{37} are the log-AOD values, variables x_{29}, x_{32}, x_{35}, and x_{38} are the altitudes, and variables x_{30}, x_{33}, x_{36}, and x_{39} are the log-thicknesses.

S2.2 Forward model singular values

Figure S2 shows the singular values of the linearized forward model $\boldsymbol{K}$.

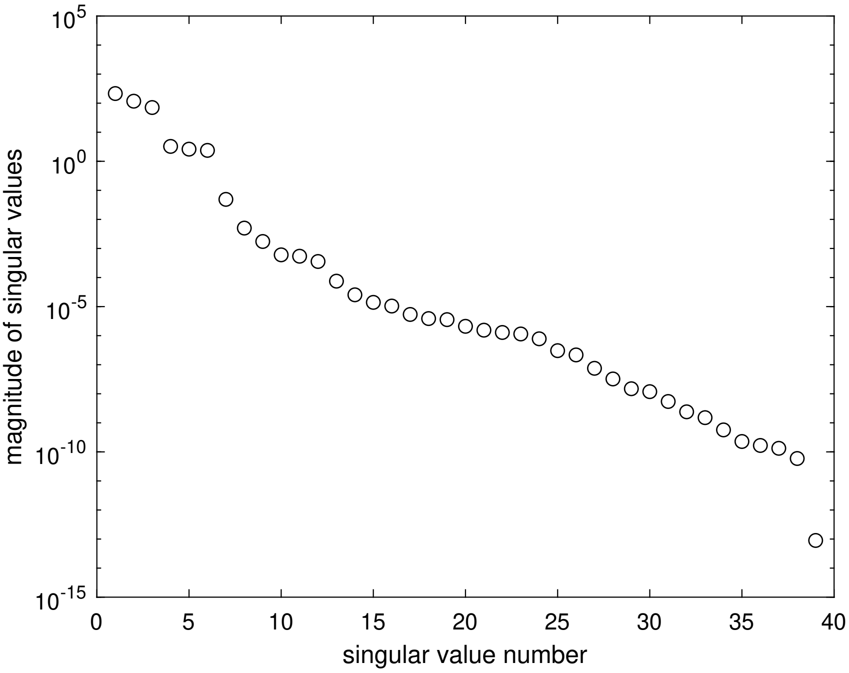
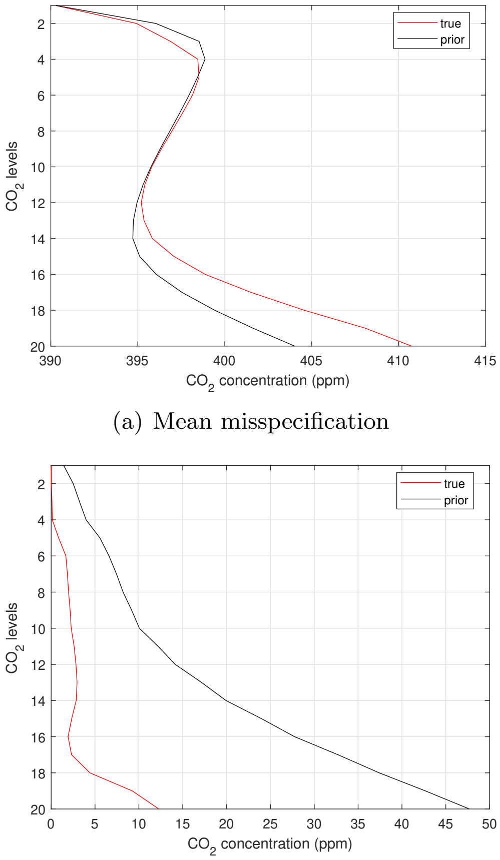


Figure S2: Singular value decay of the linearized forward model $\boldsymbol{K}$ near Lamont, OK, in October 2015.

S2.3 Exploratory analysis of the prior and generative models

In this section, we visualize and describe various components of the generative model and compare those to the prior model. We begin by comparing the prior mean and standard deviation to the true generative model mean and standard deviation for the CO_2 part of the state vector



(a) Mean misspecification

(b) Standard deviation misspecification

Figure S3: Visualization of the misspecification of the mean and the standard deviation for the CO$_2$ profile between the true generative process and the prior. Figure (a) visualizes the misspecification in the mean and Figure (b) shows the misspecification in the standard deviation.

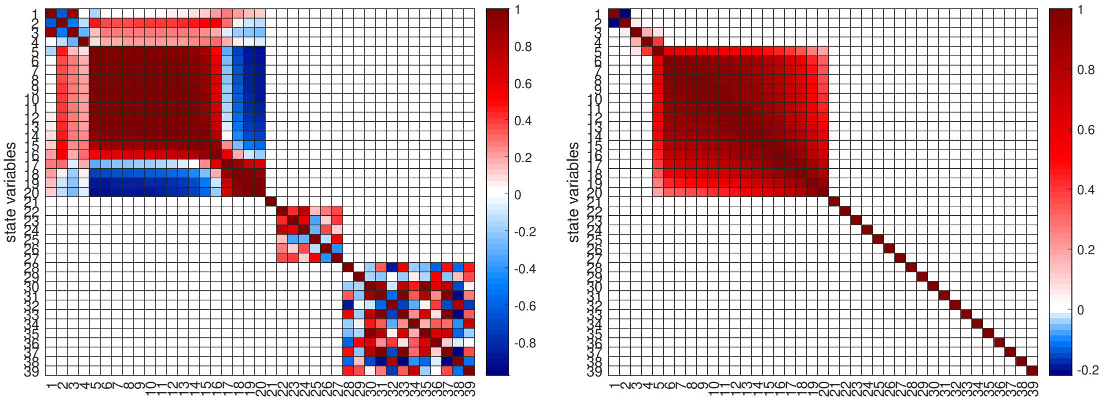
Table S1: Comparison of the true and prior mean and standard deviation for the nuisance variables $x_i, i = 21, \ldots, 39$. A detailed description of these variables is given in Section S2.1.

i	true mean (sd)	prior mean (sd)
	surface pressure	
21	972.3235 (1.7508)	970.6240 (4.000)
	albedo	
22	0.1267 (0.0204)	0.1267 (1.0000)
23	0.0001 (0.0001)	0.0000 (0.0005)
24	0.2484 (0.0044)	0.2484 (1.0000)
25	-0.0001 (0.0000)	0.0000 (0.0005)
26	0.2027 (0.0026)	0.2027 (1.0000)
27	-0.0000 (0.0000)	0.0000 (0.0005)
	aerosols	
28	-3.8786 (0.3380)	-3.7643 (2.0000)
29	0.8187 (0.0683)	0.9000 (0.2000)
30	-2.4492 (0.0409)	-2.9957 (0.1823)
31	-6.1959 (0.8492)	-4.7370 (2.0000)
32	0.3255 (0.0101)	0.9000 (0.2000)
33	-3.9219 (0.0201)	-2.9957 (0.1823)
34	-4.3980 (0.2477)	-4.3820 (1.8000)
35	-0.0087 (0.0355)	0.3000 (0.2000)
36	-3.2080 (0.0112)	-3.2189 (0.2231)
37	-5.6803 (0.2517)	-4.3820 (1.8000)
38	1.0917 (0.0870)	0.7500 (0.4000)
39	-2.3052 (0.0004)	-2.3026 (0.0953)

in Figure S3. We observe that both the mean and the standard deviation have the largest misspecification near the surface. Table S1 contains the same information for the nuisance variables $x_{21}, \ldots, x_{39}$. Generally speaking, the prior standard deviation is by design larger than that of the true process, which provides some protection against the prior mean misspecification.

We next visualize the correlation structure in the generative process and the prior. Figure S4(a) shows a heat map of the correlation matrix for the true generative process. We observe that the state vector consists of four independent subgroups of variables corresponding to the CO$_2$ profile, surface pressure, albedo variables and aerosol variables. While variables across these groups are uncorrelated, there are large within-group correlations. Figure S4(b) shows a heat map of the correlation matrix for the prior process. Again, the CO$_2$ variables are independent of the nuisance variables, but in this case, there are no correlations between the nuisance variables. Notice also the differences in the correlation structure within the CO$_2$ profile between the generative process and the prior.

Lastly, we visualize the spatial correlation structure for the generative process over an 8 × 8

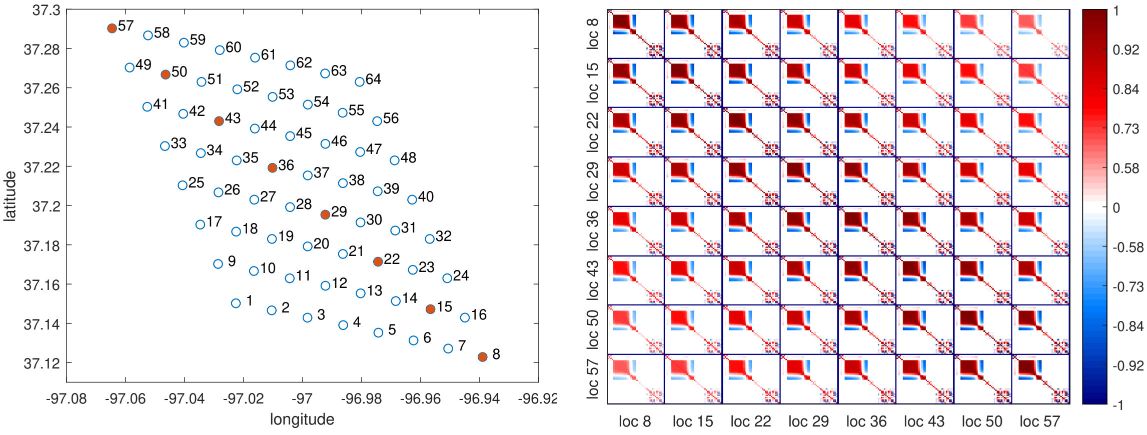


(a) True correlation

(b) Prior correlation

Figure S4: Visualization of the misspecification of the correlation structure between the true generative process and the prior. Figure (a) displays the correlation between the state vector elements in the true generative process. Figure (b) shows the same for the prior process.

spatial grid near Lamont, OK. These 64 sounding locations are shown in Figure S5(a) and Figure S5(b) displays the correlations across the locations along the diagonal in Figure S5(a). Nearby state vectors are strongly spatially correlated, but there is a fair amount of decorrelation when moving across the grid.



(a) Grid locations

(b) Grid correlation structure

Figure S5: Visualization of the correlation structure between the state vector elements for the generative process over an 8×8 spatial grid. Figure (a) shows the coordinates of the sounding locations near Lamont, OK, with the numbers giving an index for each location. Figure (b) visualizes the correlation structure between the state vector elements across the marked locations along the diagonal in Figure (a). Notice the nonlinear color scale in Figure (b).

S3 Supplement to Section 5.2: Illustrative instances

We pick some illustrative realizations to visualize the state vectors and confidence intervals produced by the two methods. We choose three representative cases corresponding to the minimum, nominal and maximum coverage for the operational method in Figure 3(b) which are also rows 1, 8 and 10 in Table 1, respectively.

Figure S6(a) shows the CO_2 part of the $\boldsymbol{x}$ instance having the smallest operational coverage. We observe, consistent with Section 5.2.1, that the operational interval is biased upward. It is overoptimistic about the amount of uncertainty, leading it to miss the true $X_{\mathrm{CO}2}$ value for this particular $\boldsymbol{\varepsilon}$ realization. From Table 1, we know that this happens for roughly 21% of $\boldsymbol{\varepsilon}$ realizations when the noncoverage probability should ideally be only 5%. The proposed interval, on the other hand, is wider and as a result ends up covering the true $X_{\mathrm{CO}2}$ value for the same $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ realizations. From Table 1, we know that the noncoverage probability for this interval is 5%, as it should be.

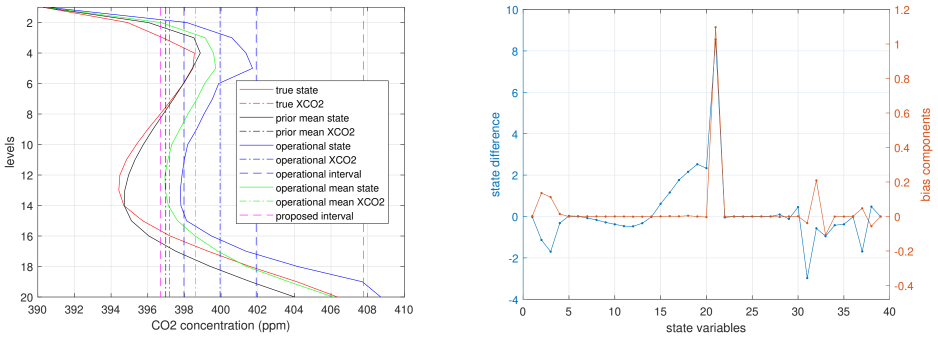
It is quite insightful to investigate the source of the upward bias for this $\boldsymbol{x}$ realization. Contrary to what one might at first imagine, it is not caused by a misspecification of the CO_2 profile in the prior. In fact, as shown in Figure S6(a), the prior CO_2 profile is very similar to the true CO_2 profile in $\boldsymbol{x}$ and the $X_{\mathrm{CO}2}$ value implied by the prior is almost the same as the true $X_{\mathrm{CO}2}$ value. Instead, it turns out that the bias is primarily caused by an upward fluctuation in the surface pressure variable x_{21}, as shown by Figure S6(b). The large positive difference in x_{21} between the true state and the prior gets multiplied by the relatively large positive bias multiplier of this variable (Figure 2(a); see also Section 5.2.1) leading to a large positive contribution to the overall bias. This results in a positively biased operational $X_{\mathrm{CO}2}$ point estimate and a miscalibrated confidence interval.

Next, instead of picking an adversarial $\boldsymbol{x}$, Figure S7(a) shows the $\boldsymbol{x}$ realization corresponding to the largest coverage for the operational method. For this $\boldsymbol{x}$, the operational $X_{\mathrm{CO}2}$ estimate is effectively unbiased and the operational interval covers the true $X_{\mathrm{CO}2}$ value for almost all $\boldsymbol{\varepsilon}$ realizations, resulting in substantial overcoverage. The proposed interval, on the other hand, again has the desired 95% coverage, as indicated by Table 1. For the $\boldsymbol{\varepsilon}$ realization shown in the figure, the operational and proposed intervals both cover the true $X_{\mathrm{CO}2}$ value.

Interestingly, the operational $X_{\mathrm{CO}2}$ estimator in Figure S7(a) is effectively unbiased even though the prior is badly misspecified for this CO_2 profile. Figure S7(b) explains this phenomenon. Due to the small bias multipliers of the low-altitude CO_2 values (Figure 2(a)), prior misspecification in the lower half of the CO_2 profile creates little bias, while the misspecification in the upper half of the profile is such that it creates both positive and negative contributions to the bias that cancel out. At the same time, there is a consistent positive contribution to the bias from x_{32}, which is negated by a downward fluctuation in the pressure variable x_{21} and a consistent negative bias contribution from x_{33}.

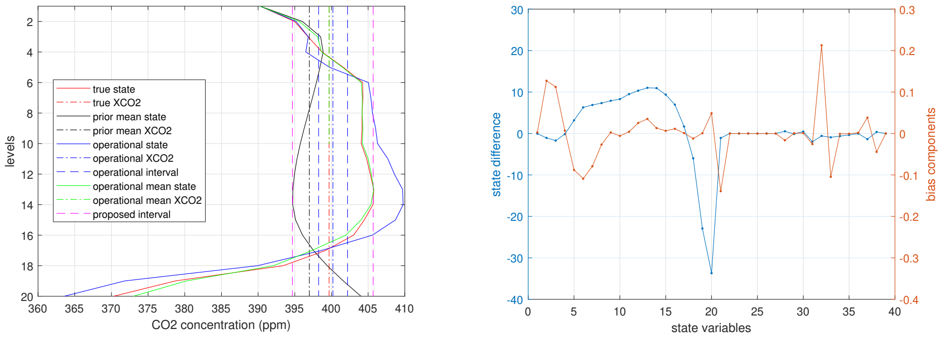
Finally, we compare in Figure S8(a) the intervals for the $\boldsymbol{x}$ realization that gives nominal coverage for the operational method in Table 1. For this $\boldsymbol{x}$, both intervals cover the true $X_{\mathrm{CO}2}$ value for 95% of $\boldsymbol{\varepsilon}$ realizations. The $\boldsymbol{\varepsilon}$ realization shown in the figure has both intervals covering the true $X_{\mathrm{CO}2}$ value. The operationally retrieved CO_2 profile has fluctuated quite substantially, but the anticorrelated fluctuations cancel out to produce a well-behaved confidence interval for $X_{\mathrm{CO}2}$. Overall, the operational method is slightly positively biased for $X_{\mathrm{CO}2}$, as expected based on the discussion in Section 5.2.2. Figure S8(b) reveals that the source of this bias is similar to the situation in Figure S6 in that the bias is primarily caused by the pressure variable x_{21}, albeit with a smaller magnitude. As before, the prior misspecification for the CO_2 profile near the surface (levels 18, 19 and 20) causes little harm due to the small bias multipliers of those variables (Figure 2(a)).

Overall, Figures S6–S8 show how challenging it is to understand, predict and explain the uncertainty quantification performance of the operational retrieval method. The method can



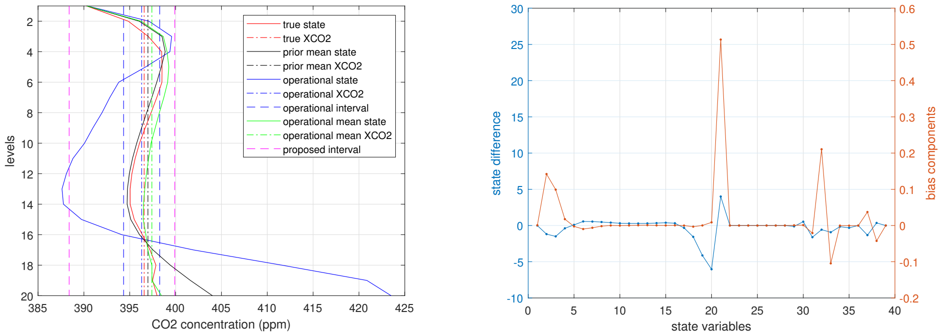
(a) Profile comparisons (b) Bias components

Figure S6: Figure (a) illustrates the operational and proposed intervals for the state realization that has the smallest coverage in Figure 3(b). Figure (b) visualizes the corresponding state differences $x_i - \mu_{a,i}$ and bias components $m_i(x_i - \mu_{a,i})$, where m_i is the ith bias multiplier.



(a) Profile comparisons (b) Bias components

Figure S7: Figure (a) illustrates the operational and proposed intervals for the state realization that has the largest coverage in Figure 3(b). Figure (b) visualizes the corresponding state differences $x_i - \mu_{a,i}$ and bias components $m_i(x_i - \mu_{a,i})$, where m_i is the ith bias multiplier.



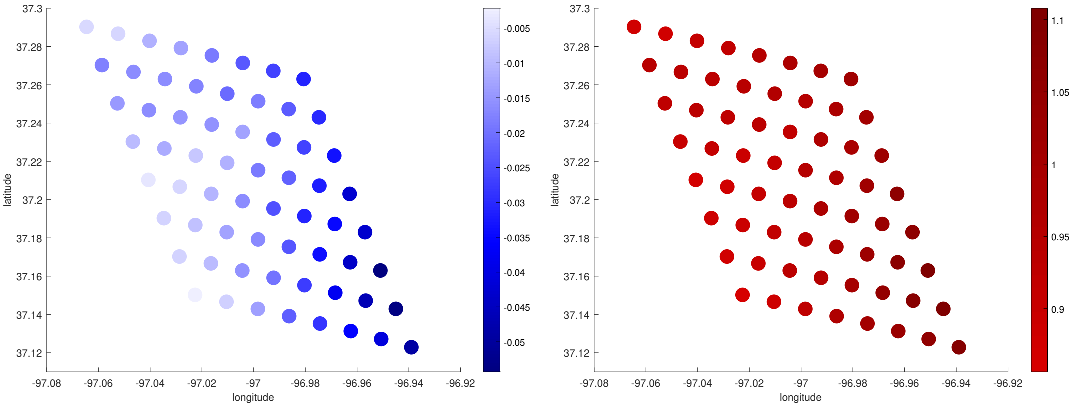
(a) Profile comparisons (b) Bias components

Figure S8: Figure (a) illustrates the operational and proposed intervals for a state realization that has nominal coverage in Figure 3(b). Figure (b) visualizes the corresponding state differences $x_i - \mu_{a,i}$ and bias components $m_i(x_i - \mu_{a,i})$, where m_i is the ith bias multiplier.

exhibit the very counterintuitive behavior where a relatively well-specified prior CO_2 profile (Figure S6) in fact has the worst coverage performance, while a badly misspecified prior CO_2 profile (Figure S7) has the highest coverage. The key to understanding the performance of the method, it turns out, is to understand the effect of the nuisance variables and how they interact with the misspecification of the CO_2 profile. Such analysis is obviously only possible when the true $\boldsymbol{x}$ is known, which would make it very difficult to perform a similar study for real-life operational retrievals. The proposed method, on the other hand, is free from these complications and exhibits consistent 95% coverage for all $\boldsymbol{x}$ realizations we have investigated.

S4 Supplement to Section 5.3

Figure S9 shows the spatial coverage and bias patterns for a case where the state vector realizations are such that all 64 intervals across the region have coverage below the nominal value.



(a) Operational coverage

(b) Operational bias

Figure S9: Operational retrieval over a grid of 8 × 8 soundings for an instance with undercoverage for the entire grid. Figure (a) shows the spatial coverage pattern relative to the nominal 95% in units of probability (i.e., -0.03, for example, corresponds to coverage 0.92, instead of the nominal 0.95). Figure (b) shows the corresponding bias pattern in ppm.

References

[1] Jonathan Hobbs, Amy Braverman, Noel Cressie, Robert Granat, and Michael Gunson. Simulation-based uncertainty quantification for estimating atmospheric CO_2 from satellite data. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):956–985, 2017.

[2] H. Boesch et al. *Orbiting Carbon Observatory-2 & 3: Level 2 Full Physics Retrieval Algorithm Theoretical Basis*. NASA Jet Propulsion Laboratory, January 2, 2019. OCO D-55207, Version 2.0 Rev 3.