NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

03 -19

Optimal Estimation Retrievals: Implications and Consequences when the Prior's Mean and Covariance are Misspecifed

Hai Nguyen, Noel Cressie, and Jonathan Hobbs

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

Optimal Estimation retrievals: Implications and consequences when the prior's mean and covariance are misspecified

Hai Nguyen[1,*], Noel Cressie[2,1], and Jonathan Hobbs[1]

[1]Jet Propulsion Laboratory, Pasadena, California

[2]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia

[*]Corresponding author. Email: hai.nguyen@jpl.nasa.gov

May 16, 2019

Optimal Estimation (OE) is a popular algorithm for remote sensing retrievals, partly due to its explicit parameterization of the sources of error and the ability to propagate them into estimates of retrieval uncertainty. These properties require specification of the prior distribution of the state vector. In many remote sensing applications, the true priors are multivariate and hard to characterize properly. Instead, priors are often constructed based on subject-matter expertise, existing empirical knowledge, and a need for computational expediency, resulting in a "working prior." This paper

explores the retrieval bias and the inaccuracy in retrieval uncertainty by explicitly separating the true prior (the probability distribution of the underlying state) from the working prior (the probability distribution used within the OE algorithm). We find that, in general, misspecifying the mean in the working prior will lead to biased retrievals, and misspecifying the covariance in the working prior will lead to inaccurate estimates of the retrieval uncertainty, though their effects vary depending on the state-space signal-to-noise ratio of the observing instrument. Our results point towards some attractive properties of a class of uninformative priors that is implicit for least-squares retrievals. Further, our derivations provide a theoretical basis, and an understanding of the trade-offs involved, for the popular OE practice of inflating a working prior covariance in order to reduce the prior's impact on a retrieval. Finally, our results also lead to practical recommendations for specifying the prior mean and covariance in OE.

Keywords: bias, inverse problem, efficiency, satellite retrievals, uncertainty quantification, validity

1 Introduction

Remote sensing from satellites involves the acquisition of surface and atmospheric states through measurement of electromagnetic radiation reflected from Earth's surface. Satellites are often designed to have global coverage, and a large number of physical processes (e.g., aerosols, carbon dioxide, sea surface height, land cover, leaf index) can be captured with instruments sensitive to the appropriate spectral bands. The functional relationship between the "hidden" geophysical variables of interest

and the observed spectral information can be expressed through radiative transfer equations, often called a forward model. The estimation of these variables from the observed spectral information (e.g., radiances) and the radiative transfer equations can be classified as an inverse problem.

One popular method for solving remote sensing inverse problems is called Optimal Estimation (OE; Rodgers, 2000), which regularizes the solution using Bayes' theorem. It entails specifying a (typically Gaussian) prior distribution for the natural variability of the hidden physical process, a (typically Gaussian) distribution for the spectral measurement errors, and an explicit (typically non-linear) forward model that relates the atmospheric state (or simply the state) functionally to noise-free radiances. Assuming all distributional parameters are known, the retrieved (or estimated) state from OE is then the maximum a posteriori (or MAP) estimate of the state given the observed, noisy radiances.

OE's specification of the sources of variability within a Bayesian framework allows the inverse problem to be regularized in addition to allowing the propagation of sources of error into a measure of the estimated state's uncertainty. For these reasons, OE has been the method of choice in many applications, including estimating total-column carbon dioxide for NASA's Orbiting Carbon Observatory-2 (OCO-2; O'Dell et al., 2018), sea surface temperature for the Spinning Enhanced Visible and Infra-Red Imager (SEVIRI; Merchant et al., 2013), total-column carbon dioxide and methane from the Greenhouse Gases Observing Satellite (GOSAT; Yoshida et al., 2013), temperature and ozone from the Tropospheric Emission Spectrometer (TES; Bowman et al., 2006), temperature and water vapor from the Atmospheric Infrared

Sounder (AIRS; Irion et al., 2018), and aerosols from the Meteosat Second Generation Spinning Enhanced Visible and Infrared Imager (MSG/SEVIRI; Govaerts et al., 2010).

1.1 The "working" prior

One of the advantages of OE relative to maximum-likelihood-based retrievals is OE's ability to propagate different sources of error into estimates of retrieval uncertainty. However, the validity of these uncertainty estimates implicitly requires that the prior distribution of the state used in the algorithm, which we call the "working prior" in this paper (Cressie, 2018), *matches* the true probability distribution of the state.

Rodgers (2000, Section 6.5.1) recognized that "if the *a priori* are inappropriate, [then] their errors are incorrect." He went on to acknowledge the difficulty of knowing the true distribution of the state, recommending that practitioners make a "reasonable estimate of a probability density function consistent with all our knowledge, one that is least committal about the state but consistent with whatever more or less detailed understanding we may have of the state vector prior to the measurement(s)" (Rodgers, 2000, Section 10.3.3.2). This approach is reflected in most implementations of OE retrievals. For instance, the OCO-2 retrieval uses a state vector that includes carbon dioxide, aerosol properties, and surface properties. The working prior's mean vector that is used in the retrieval algorithm is chosen using "a climatology based on the GLOBALVIEW dataset, and [they] change based on the time of year and the latitude of the site" (Wunch et al., 2011). The working prior's covariance matrix for the OCO-2 retrieval is assumed to be diagonal for all non-CO_2 state elements. For

the CO_2 elements, the prior covariance matrix has off-diagonal elements "estimated based on the Laboratoire de Météorologie Dynamique general circulation model, but the correlation coefficients were reduced arbitrarily to ensure numerical stability in taking its inverse" (Boesch et al., 2015). Furthermore, the diagonal elements of the CO_2 prior covariance matrix are "unrealistically large for most of the world, [they are] intended to be a minimal constraint on the retrieved XCO2."

Another example of the compromise between expediency and physical fidelity in designing the prior distribution of the state can be seen in Irion et al.'s (2018) OE retrieval of temperature and water vapor from the AIRS instrument. There, the state vector consists of surface temperature, atmospheric temperature, water vapor, CO_2, O_3, and cloud properties. The prior mean vector for surface temperature, atmospheric temperature, and water vapor is interpolated from European Centre for Medium-Range Weather Forecasts (ECMWF) data; for CO_2 it is the one used by the Total Carbon Column Observing Network (TCCON); and for cloud properties it is derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data. The prior covariance matrix for the OE retrieval is block diagonal with zero covariances between any of the respective constituents (e.g., between temperature and water vapor). Within the individual constituents (i.e., temperature, water vapor, CO_2, and O_3), the dependence along the vertical direction is modeled with an exponential covariance function (also called a Markov-process covariance; see Rodgers, 2000), where length scales are chosen "guided by previous experience with AIRS and TES retrievals" (Irion et al., 2018).

1.2 Twomey-Tikhonov versus Bayesian approach

It is apparent to us that the prior distributions for remote sensing, as they are widely designed in practice, draw from two separate traditions. In the first, the prior distribution is viewed as an *ad hoc* constraint or "regularizer" to ensure stability and uniqueness of the MAP solution. This is also known as the Twomey-Tikhonov approach (Rodgers, 2000, p. 108). In this tradition, it is perfectly valid to make the prior variance of a particular constituent unrealistically large so as to impose minimal external constraints on the retrieval. The second tradition is a Bayesian approach, where the prior's mean and covariance are assumed to come from the true probability distribution of the state. Here, the prior information is supposed to reflect as accurately as possible all knowledge about the variability of the state. Under the Bayesian approach, making variance terms unrealistically large to minimize the prior's impact on the retrieval, or making absolute covariance terms unrealistically small to ensure numerical stability, can have serious statistical consequences. In the Bayesian tradition, one should set the prior mean and covariance in accordance with a realistic understanding of the natural variability of the state.

Both the Twomey-Tikhonov approach and the Bayesian approach share the same equations (e.g., cost function, Levenberg-Marquardt update) that result in a retrieval of the state. However, there is a disconnect between the two when interpreting statistically the resulting estimated *uncertainties of the retrieval*. That is, when the prior distribution is misspecified, the estimated state's uncertainty may no longer be representative of the error one would see when comparing the retrievals to independent validation data. When the working prior means, variances, and covariances are

constructed under the Twomey-Tiknohov interpretation, with an eye towards expediency, in general the retrieval will be biased and the estimated retrieval uncertainty will not represent the true uncertainty. This can have serious consequences in subsequent scientific studies, such as flux inversion (e.g., Engelen et al., 2002).

1.3 Misspecification of the prior

The theoretical consequence of prior-distribution misspecification in OE retrievals is not well explored in the literature, with some studies made in special cases. Luo et al. (2007) investigated the impact of the prior and instrument characteristics on TES retrievals, and Hobbs et al. (2017) examined the relationship of total-column CO_2 bias and retrieval uncertainties with different specifications of OE and algorithmic parameters such as prior means, variances, covariances, starting values, and the convergence criterion. Kulawik et al. (2008, p. 3081) contend that different choices of priors might be appropriate, depending on different goals, noting that "[using] the most accurate prior will lead to the most accurate result; however conversion to a uniform prior can be useful for scientific analysis." Su et al. (2017) gave a derivation of the discrepancy arising from misspecification of the priors under a linearization assumption, although they focused on numerical case studies rather than on studying the theoretical properties arising therefrom. Cressie et al. (2017) examined the AIRS CO_2 retrieval algorithm and demonstrated that its least-squares cost function is equivalent to the OE cost function with an uninformative prior. Ramanathan et al. (2018) showed that a class of retrieval methods called the Singular Value Decomposition (SVD) retrieval is equivalent to an OE method with an uninformative prior

where the gain matrix is computed using a pseudo-inverse.

In this paper, we give an in-depth investigation of the consequences of misspecification of the prior mean vector and the prior covariance matrix *of the state vector* (that is, when the working prior is not the same as the true prior) by examining its effects on the retrieval bias and the retrieval uncertainty. In theory, it is also possible to misspecify the distribution of the measurement errors of the radiances, but that is another topic, and in what follows we assume that the radiances' measurement-error parameters are correctly specified.

The organization of our paper is as follows: In Section 2, we derive the multivariate equations for the bias and errors arising from prior misspecification. We give a simple example of a univariate state, to gain intuition into the properties implied by the multivariate equations. We also give the multivariate bias and errors for a particular choice of prior– the uninformative prior– versus the traditional prior used in OE retrievals, and we discuss the theoretical trade-offs between the choices therein. In Section 3, we design a simulation study using a surrogate OCO-2 forward model to evaluate empirically the consequences of prior misspecification, which we then compare to the theoretical derivations. In Section 4, we conclude with some observations and recommendations on choosing a prior in practice for Optimal Estimation of the state from satellite remote sensing data.

2 Derivation of retrieval equations

The OE framework, as formalized in Rodgers (2000), is a Bayesian approach to solve inverse problems in remote sensing. In this section, we review OE and derive the bias and error of an OE retrieval arising from mispecification of the prior.

In many OE applications, the forward model is non-linear, and solving for the optimal solution requires iterative optimization methods such as the Levenberg-Marquardt algorithm (e.g., Connor et al., 2008). The non-linear solver introduces complicating optimization-specific factors such as local minima, convergence criteria, linearization, and numerical stability. These can make it difficult to isolate the effect of prior misspecification within the resulting error analysis. Therefore, in this paper we shall focus on the leading case of a linear forward model. Our derivations are in fact highly relevant to non-linear problems, as this linearization approach is also used in quantifying the uncertainty of the OE retrieval (Rodgers, 2000, Section 5.5). When the forward model is moderately or highly non-linear, the conclusions derived from the linear case can be viewed as first-order approximations (Rodgers, 2000; Cressie et al., 2016).

2.1 Background

Consider the case where an N-dimensional radiance vector $\mathbf{y}$ is related to the r-dimensional (hidden) true state $\mathbf{x}$ by the following data model:

$$\mathbf{y} \;=\; \mathbf{F}(\mathbf{x}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{F}(\,\cdot\,)$ is the N-dimensional vector-valued forward model, $\mathbf{x}$ is the r-dimensional Gaussian true state with true mean $\mathbf{x}_T$ and true covariance matrix $\mathbf{S}_T$, and $\boldsymbol{\epsilon}$ is the N-dimensional Gaussian measurement-error vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}_\epsilon$, independent of $\mathbf{x}$. That is, $\mathbf{x} \sim \text{Gau}_r(\mathbf{x}_T, \mathbf{S}_T)$ and $\boldsymbol{\epsilon} \sim \text{Gau}_N(\mathbf{0}, \mathbf{S}_\epsilon)$, where $\text{Gau}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an n-dimensional Gaussian (or normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For the leading case of a linear forward model, (1) becomes,

$$\mathbf{y} \;=\; \mathbf{c} + \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}, \tag{2}$$

where the $N \times r$ matrix $\mathbf{K} = \frac{\partial \mathbf{F}}{\partial \mathbf{x}}$ is the Jacobian of the forward model, and $\mathbf{c}$ is an N-dimensional constant vector. The linear model in (2) could be thought of as the first-order term of the Taylor expansion of the non-linear model (1) around some known state vector (e.g., Cressie, 2018). Here, we assume that $\text{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathbf{S}_\epsilon$ is known. It is possible that the distributional parameters of $\boldsymbol{\epsilon}$ could be misspecified, but in this paper we focus on the more likely scenario of misspecification of the prior, $\text{Gau}_r(\mathbf{x}_T, \mathbf{S}_T)$, for the state vector $\mathbf{x}$.

Without loss of generality, we can assume that $\mathbf{c} = \mathbf{0}$ (since $\mathbf{c}$ is known and hence in principle can be subtracted from $\mathbf{y}$), in which case $\mathbf{y}$ is a vector of "centered" radiances. Our data model then becomes,

$$\mathbf{y} \;=\; \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}. \tag{3}$$

Rodgers (2000) proposes a loss function $L(\,\cdot\,)$ that is the negative logarithm of the

posterior distribution of $\mathbf{x}$ given $\mathbf{y}$; that is, after dropping constant terms,

$$L(\mathbf{x}) \equiv -2\log P(\mathbf{x}|\mathbf{y}) = (\mathbf{y} - \mathbf{K}\mathbf{x})' \, \mathbf{S}_\epsilon^{-1} \, (\mathbf{y} - \mathbf{K}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_T)' \, \mathbf{S}_T^{-1} \, (\mathbf{x} - \mathbf{x}_T). \tag{4}$$

The maximum a posteriori (MAP) solution (also the posterior mean in our case where the forward model is linear) is then given by,

$$\hat{\mathbf{x}}_T = \mathbf{x}_T + \mathbf{G}_T(\mathbf{y} - \mathbf{K}\mathbf{x}_T), \tag{5}$$

where $\mathbf{G}_T$ is called the gain matrix and is given by $\mathbf{G}_T = (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{S}_\epsilon^{-1}$. The uncertainty on $\hat{\mathbf{x}}$ is then given by,

$$\mathbf{\Sigma}_T \equiv \mathrm{var}_T(\hat{\mathbf{x}}_T - \mathbf{x}) = (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}, \tag{6}$$

where the subscript T on the variance operator indicates that statistical calculations are with respect to the true prior parameters $\{\mathbf{x}_T, \mathbf{S}_T\}$. The formulation above assumes that the prior mean and covariance matrix, $\{\mathbf{x}_T, \mathbf{S}_T\}$, are known perfectly. In practice, this is rarely the case. As discussed in Section 1, we draw a distinction between the (often unknown) true prior parameters $\{\mathbf{x}_T, \mathbf{S}_T\}$ and the specified working prior parameters $\{\mathbf{x}_w, \mathbf{S}_w\}$, which are used in algorithms and are often constructed from a mixture of educated guesses, empirical studies, need for computational expediency, and subject-matter expertise. Since the distribution of the state is assumed Gaussian, we abuse notation slightly by referring to $\{\mathbf{x}_T, \mathbf{S}_T\}$ as the true prior and $\{\mathbf{x}_w, \mathbf{S}_w\}$ as the working prior. Researchers have long recognized that retrieval un-

certainty in (6) is incorrect when $\{\mathbf{x}_w, \mathbf{S}_w\} \neq \{\mathbf{x}_T, \mathbf{S}_T\}$ (e.g., Rodgers, 2000; Kulawik et al., 2008; Cressie et al., 2016; Su et al., 2017; Cressie, 2018). To understand the effects of prior misspecification, we shall examine separately the effect on the retrieval bias (Section 2.2) and the effect on the retrieval uncertainty (Section 2.3). For ease of reference, we provide a list of the common mathematical symbols used in this paper and their meaning in Table 1.

2.2 Bias arising from prior misspecification

Having specified the working prior $\{\mathbf{x}_w, \mathbf{S}_w\}$, the MAP estimate $\hat{\mathbf{x}}_w$ is,

$$\hat{\mathbf{x}}_w = \mathbf{x}_w + \mathbf{G}_w \left(\mathbf{y} - \mathbf{K}\mathbf{x}_w\right), \tag{7}$$

where the subscript w on the the retrieved value $\hat{\mathbf{x}}_w$ and the gain matrix $\mathbf{G}_w$ indicates that they both depend on the working prior. The working gain matrix $\mathbf{G}_w$ has the following form:

$$\mathbf{G}_w = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{S}_\epsilon^{-1}. \tag{8}$$

When the working prior $\{\mathbf{x}_w, \mathbf{S}_w\}$ is separated notationally from the true prior $\{\mathbf{x}_T, \mathbf{S}_T\}$, it is easy to compute the bias from (7) as a function of the working prior

Table 1: Reference guide for mathematical symbols

Symbol	Definition
$\mathbf{y}$	Observed N-dimensional vector of radiances
$\mathbf{x}$	True (hidden) r-dimensional vector of state elements
$\boldsymbol{\epsilon}$	N-dimensional vector of radiance error
$\mathbf{K}$	Jacobian of the forward model
$\mathbf{x}_T$	True prior mean vector of the state vector $\mathbf{x}$
$\mathbf{x}_w$	Working prior mean vector of $\mathbf{x}$
$\hat{\mathbf{x}}_T$	Retrieved state vector under the true prior
$\hat{\mathbf{x}}_w$	Retrieved state vector under a working prior
$\mathbf{S}_\epsilon$	Covariance matrix for the radiance measurement error vector $\boldsymbol{\epsilon}$
$\mathbf{S}_T$	True prior covariance matrix of the state vector $\mathbf{x}$
$\mathbf{S}_w$	Working prior covariance matrix of $\mathbf{x}$
$\mathbf{G}_T$	Gain matrix under the true prior
$\mathbf{G}_w$	Gain matrix under the working prior
$\mathbf{b}_T(\,\cdot\,)$	True retrieval bias for OE estimates (as a function of the working prior)
$\boldsymbol{\Sigma}_w(\,\cdot\,)$	Working retrieval uncertainty from the OE algorithm
$\boldsymbol{\Sigma}_T(\,\cdot\,)$	True retrieval uncertainty for OE estimates (as a function of the working prior)

as follows:

$$\begin{aligned}
\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w) &\equiv \mathrm{E}_T(\hat{\mathbf{x}}_w - \mathbf{x}), \\
&= \mathrm{E}_T(\mathbf{x}_w + \mathbf{G}_w (\mathbf{y} - \mathbf{K}\mathbf{x}_w) - \mathbf{x}), \\
&= (\mathbf{I} - \mathbf{G}_w\mathbf{K})(\mathbf{x}_w - \mathbf{x}_T), \\
&\equiv (\mathbf{I} - \mathbf{A}_w)(\mathbf{x}_w - \mathbf{x}_T), \qquad (9)
\end{aligned}$$

where the subscript T indicates that all statistical calculations are with respect to the true prior $\{\mathbf{x}_T, \mathbf{S}_T\}$, and $\mathbf{A}_w \equiv \mathbf{G}_w\mathbf{K}$ is the working averaging kernel. From (8), it is straightforward to show that $(\mathbf{I} - \mathbf{A}_w) = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}$, which when substituted into (9) gives,

$$\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w) = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}(\mathbf{x}_w - \mathbf{x}_T). \qquad (10)$$

The key difference between the bias formula in (10) and its treatment in Section 3.4.2 of Rodgers (2000) is that our result is general for any working prior $\{\mathbf{x}_w, \mathbf{S}_w\}$. From (10), we see that the expected bias is equal to the product of the difference in prior means, $(\mathbf{x}_w - \mathbf{x}_T)$, and the matrix $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}$. Assume that the working prior covariance matrix $\mathbf{S}_w$ is positive-definite; since $\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}$ is positive-semidefinite, then the matrix $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}$ is positive-definite. Thus, the retrieval bias $\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w) = \mathbf{0}$ if the working prior mean is correct (i.e., $\mathbf{x}_w = \mathbf{x}_T$). Clearly, $\mathbf{x}_w = \mathbf{x}_T$ is a sufficient condition for unbiasedness. But note that OE retrievals can be unbiased when the working prior covariance matrix $\mathbf{S}_w$ is incorrect, as long as

the working prior mean $\mathbf{x}_w$ is correct.

Looking closely at (10), we see that a bias term $(\mathbf{x}_w - \mathbf{x}_T)$ is multiplied by $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}$. Recall $\mathbf{S}_w$ is positive-definite and $\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}$ is positive-semidefinite, it is easy to show that $\mathbf{0} < (\mathbf{S}_w^{-1}+\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1} \leq \mathbf{I}$, where $\mathbf{B} \leq \mathbf{A}$ means that $\mathbf{A}-\mathbf{B}$ is positive-semidefinite, and $\mathbf{B} < \mathbf{A}$ means that $\mathbf{A}-\mathbf{B}$ is positive-definite. Therefore, we can interpret this multiplicative term as 'shrinking' the bias depending on the relative strength between the working prior covariance $\mathbf{S}_w$ and the measurement-error contribution $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$. Mathematically, the latter matrix could be interpreted as the variance of the maximum-likelihood estimate of $\mathbf{x}$ using a frequentist approach (Section 2.6). Physically, it could also be interpreted as an expression of the measurement-error variability in the lower-dimensional state-space. When $\mathbf{S}_w$ is much 'smaller' than $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$ (that is, we have a lot of confidence and hence tight constraints on the trace or determinant of $\mathbf{S}_w$), then $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}$ 'approaches' $\mathbf{I}$, and hence the bias 'approaches' $(\mathbf{x}_w - \mathbf{x}_T)$. Another implication of (10) is that we can greatly reduce the bias resulting from an incorrect working prior, by relaxing constraints and being overly conservative in choosing our working prior covariance $\mathbf{S}_w$. That is, if we let $\mathbf{S}_w$ be unrealistically 'large' relative to $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$, then the bias 'approaches' $\mathbf{0}$. More formally, let $\mathbf{S}_w \to \boldsymbol{\infty}$, which we define as $\min(\lambda_1(\mathbf{S}_w), \ldots, \lambda_r(\mathbf{S}_w)) \to \infty$, with $\lambda_i(\mathbf{S}_w)$ being the i-th eigenvalue of $\mathbf{S}_w$. Then $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1} \to \mathbf{0}$, and

$$\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w) \quad \to \quad \mathbf{b}_T(\mathbf{x}_w, \boldsymbol{\infty}) \equiv \mathbf{0}. \tag{11}$$

The result in (11) is noteworthy, since the choice, $\mathbf{S}_w \to \boldsymbol{\infty}$ (equivalently, $\mathbf{S}_w^{-1} \to \mathbf{0}$),

constitutes a type of uninformative prior that is implicit in the frequentist maximum-likelihood formulation, a popular alternative choice for atmospheric retrievals (e.g., the AIRS CO_2 retrieval algorithm; Susskind et al., 2003). Cressie et al. (2017) showed that the AIRS least-squares (i.e., maximum-likelihood) retrieval can be considered an OE retrieval with an uninformative prior, in support of (11). In the rest of this paper, we shall use "OE" to refer to the case where estimates arise from an *informative prior*, and we shall use "maximum likelihood" or "least squares" to refer to the case of an *uninformative prior*. From (11), we see that maximum-likelihood methods have an advantage over OE in that their retrievals are always unbiased, while OE retrievals with an informative prior are *biased whenever the working prior mean* $\boldsymbol{x}_w$ *is misspecified*. However, as seen in Section 2.5, maximum-likelihood methods are statistically inefficient, often considerably so.

In summary, we can conclude that the choice of prior mean $\mathbf{x}_w$ is very important when OE is used to retrieve the state $\mathbf{x}$, with a bias arising when the working prior mean differs from the true prior mean. The magnitude of this bias varies between $||(\mathbf{x}_w - \mathbf{x}_T)||$ and 0, depending on the working prior covariance matrix $\mathbf{S}_w$. For algorithms using a working prior where $\mathbf{S}_w \to \boldsymbol{\infty}$, the bias $\mathbf{b}_T$ approaches $\mathbf{0}$ regardless of the choice of the working prior mean $\mathbf{x}_w$.

2.3 Inaccurate uncertainty arising from prior misspecification

In the previous section, we saw that for OE, a misspecified prior mean $\mathbf{x}_w$ results in a biased retrieval. We now consider the effect of misspecification of the prior on the retrieval uncertainty. From the working prior, the OE algorithm produces its own internal estimate of the retrieval uncertainty, $\mathbf{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w)$, as follows:

$$\mathbf{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w) \equiv \mathrm{var}_w(\hat{\mathbf{x}}_w - \mathbf{x}) = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}, \tag{12}$$

where the subscript w shown on $\mathbf{\Sigma}_w(\,\cdot\,)$ indicates that it is calculated with respect to the *working* prior. It is seen later in this subsection that the quantity (12) is equal to $\mathrm{var}_T(\hat{\mathbf{x}}_w - \mathbf{x})$ given by (14), provided $\mathbf{S}_w$ is the *same* as the true prior covariance $\mathbf{S}_T$. Rodgers (2000) recognized that this condition is very restrictive and one that is unlikely to be achieved in practice. Therefore, he recommended restraint and circumspection in the interpretation of (12), noting that to "estimate [the retrieval uncertainty] correctly, the actual statistics of the fine structure must be known. It is not enough to simply use some *ad hoc* matrix that has been constructed as a reasonable *a priori* constraint in the retrieval. If that real covariance matrix is not available, it may be better to abandon the estimation of the smoothing error, and consider the retrieval as an estimate of the smoothed version of the state, rather than an estimate of the complete state." (Rodgers, 2000, Section 3.2.1)

Here, we make Rodgers' warning mathematically precise, in addition to providing some guidance on choosing a 'good' prior. The true retrieval uncertainty is derived

as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w) &= \mathrm{var}_T(\hat{\mathbf{x}}_w - \mathbf{x}) \\
&= \mathrm{var}_T(\mathbf{x}_w + \mathbf{G}_w(\mathbf{y} - \mathbf{K}\mathbf{x}_w) - \mathbf{x}) \\
&= \mathrm{var}_T((\mathbf{G}_w\mathbf{K} - \mathbf{I})\mathbf{x} + \mathbf{G}_w\epsilon) \\
&= (\mathbf{G}_w\mathbf{K} - \mathbf{I})\mathbf{S}_T(\mathbf{G}_w\mathbf{K} - \mathbf{I})' + \mathbf{G}_w\mathbf{S}_\epsilon^{-1}\mathbf{G}_w' \,, \qquad (13)
\end{aligned}$$

since $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are independent, and recall from (8) that $\mathbf{G}_w = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}\mathbf{S}_\epsilon^{-1}$. Substituting this into (13), we see that,

$$\boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w) = (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}')(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}. \quad (14)$$

We note here that both the working retrieval uncertainty and the true retrieval uncertainty in (12) and (14), respectively, are dependent only on $\mathbf{S}_w$ and $\mathbf{S}_T$. This means that the accuracy of $\mathrm{var}_w(\hat{\mathbf{x}}_w - \mathbf{x})$ is *not affected* by misspecification of the prior mean $\mathbf{x}_w$. Now in practice the mean squared error (MSE) is an alternative measure of validation performance. It is the sum of the 'squared' retrieval bias and the true retrieval uncertainty as given by,

$$MSE \equiv E_T((\hat{\mathbf{x}}_w - \mathbf{x}_T)(\hat{\mathbf{x}}_w - \mathbf{x}_T)') = \mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w)\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w)' + \boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w).$$

Hence, the retrieval MSE is affected by both misspecifications, $\mathbf{x}_w \neq \mathbf{x}_T$ and $\mathbf{S}_w \neq \mathbf{S}_T$.

It is straightforward to show that when $\mathbf{S}_w = \mathbf{S}_T$, (12) and (14) are the same:

$$\begin{aligned}\boldsymbol{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T) &= (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_T^{-1}\mathbf{S}_T\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \\ &= (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} = \boldsymbol{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w), \qquad (15)\end{aligned}$$

since $\mathbf{S}_w = \mathbf{S}_T$. When $\mathbf{S}_w \neq \mathbf{S}_T$, we show in Section 2.5 that $\boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)$ is 'larger' than $\boldsymbol{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T)$, and hence $\hat{\mathbf{x}}_w$ is less accurate than $\hat{\mathbf{x}}_T$.

To gain some intuition into the bias and uncertainty under prior misspecification, in the next subsection we consider a univariate state (i.e., $r = 1$). This allows us to demonstrate some interesting theoretical trade-offs between two particular classes of priors. Then the general case of a multivariate state vector r is presented in Sections 2.5 and 2.6.

2.4 Univariate case study

To understand further the behavior of the true bias and true uncertainty of the retrieval, we consider a simple univariate forward model, which we use to help interpret the multivariate formulas given by (10) and (14). In this subsection, we assume that both the radiance y and the state x are scalars and that the data model is,

$$y = kx + \epsilon, \qquad (16)$$

where $x \sim \text{Gau}(x_T, \sigma_T^2)$ and $\epsilon \sim \text{Gau}(0, \sigma_\epsilon^2)$ independently, and k, x_T, σ_T^2, and σ_ϵ^2 are one-dimensional versions of the terms $\mathbf{K}$, $\mathbf{x}_T$, $\mathbf{S}_T$, and $\mathbf{S}_\epsilon$, respectively. The OE

retrieval and its uncertainty can be obtained as a special case of (5) and (6). Then the true retrieval bias (10) becomes,

$$b_T(x_w, \sigma_w^2) = \left(\frac{1}{\sigma_w^2} + \frac{k^2}{\sigma_\epsilon^2}\right)^{-1} \frac{1}{\sigma_w^2}(x_w - x_T). \tag{17}$$

In what follows, we pay particular attention to the state-space signal-to-noise ratio (SNR), which is the ratio of the variability of the signal (σ_T^2) to the measurement-error variability expressed in the state space (σ_ϵ^2/k^2). Note that in the remote sensing literature, SNR is typically computed within radiance space; it is usually defined as the ratio of the reference radiance intensity to the standard deviation of the radiance noise $\boldsymbol{\epsilon}$. To make it clear that our SNR refers to the state space, we shall refer to the ratio $\frac{\sigma_T^2}{(\sigma_\epsilon^2/k^2)}$ as the state-space SNR. To see the effects on the true retrieval bias (17), we consider three cases of state-space SNR: 0.5, 1, and 2. We fix the parameters $k = 1, x_w = 0$, $x_T = 1$, and $\sigma_T^2 = 1$, and consequently the three cases correspond to $\sigma_\epsilon^2 \in \{0.5, 1, 2\}$.

The bias b_T, as a function of the working prior variance σ_w^2, is plotted in the left panel of Figure 1. It is clear that the bias is negative and largest when unquestioning confidence ($\sigma_w = 0$) is put on the incorrect prior mean $x_w = 0$; recall that the true prior mean is $x_T = 1$. In this case, the bias is simply $x_w - x_T = -1$. As σ_w^2 increases from 0, the bias decreases monotonically towards 0. The rate at which the bias is reduced depends on the state-space SNR. The case of SNR = 2 shows a bias decreasing to 0 faster than the case of SNR = 1, which decreases to 0 faster than the case of SNR = 0.5.

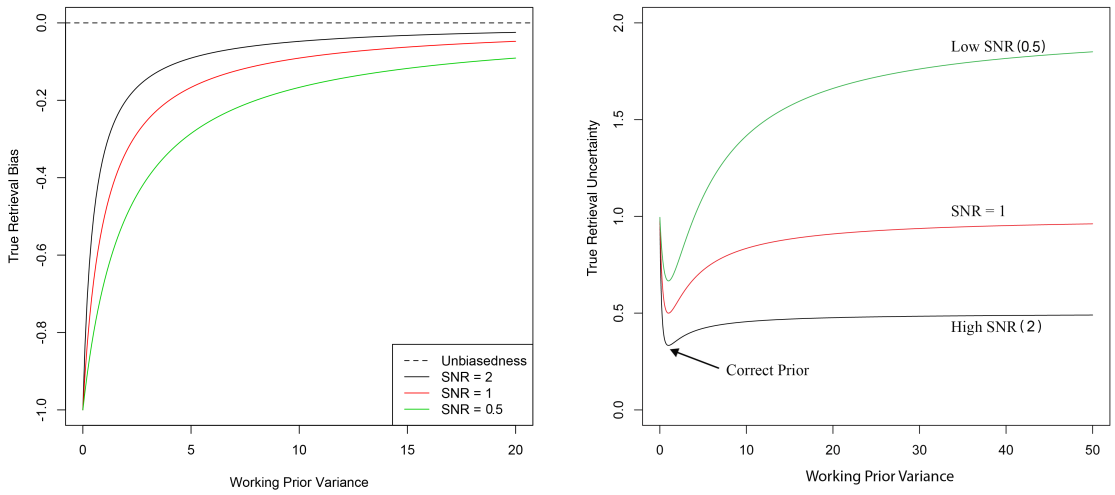

Figure 1: Left panel: True retrieval bias (vertical axis) resulting from OE as a function of σ_w^2 (horizontal axis) for a univariate model where $x_w = 0, x_T = 1$, and $\sigma_T^2 = 1$, for three choices of state-space SNRs. Right panel: The true retrieval uncertainty s_T^2 (vertical axis) given by (18) as a function of σ_w^2 (horizontal axis) for the same three choices of state-space SNR.

Assume the univariate retrieval model given by (16); then by substituting $r = 1$ into (14), we obtain the univariate true retrieval uncertainty:

$$s_T^2(x_w, \sigma_w^2) = \left(\frac{1}{\sigma_w^2} + \frac{k^2}{\sigma_\epsilon^2}\right)^{-1} \left(\frac{1}{\sigma_w^4}\sigma_T^2 + \frac{k^2}{\sigma_\epsilon^2}\right) \left(\frac{1}{\sigma_w^2} + \frac{k^2}{\sigma_\epsilon^2}\right)^{-1}, \tag{18}$$

which is plotted in the right panel of Figure 1 as a function of σ_w^2, for SNR $\in \{0.5, 1, 2\}$. We see that for all three SNRs, the true uncertainty s_T^2 is smallest when the working prior variance σ_w^2 is equal to the true prior variance $\sigma_T^2 = 1$. That is, $s_T^2(x_T, \sigma_T^2) \leq s_T^2(x_w, \sigma_w^2)$ for all $\{x_w, \sigma_w^2\}$. This inequality demonstrates the *statistical efficiency* (i.e., smallest uncertainty) of the retrieval when using the true prior; it is easy to show that statistical efficiency holds for $\sigma_w^2 = \sigma_T^2$ and all choices of $\{k, x_w, \sigma_\epsilon^2, x_T, \sigma_T^2\}$. In Section 2.5, we prove the result in the multivariate context where the state dimension $r > 1$.

In Section 2.2, we saw that the uninformative working prior (i.e., $\sigma_w^2 \to \infty$) that is implicit in maximum-likelihood methods has the advantage of yielding unbiased estimates (Figure 1, left panel). However, the right panel of Figure 1 indicates that an uninformative working prior (i.e., $\sigma_w^2 \to \infty$) yields statistically inefficient retrievals, since σ_w^2 has to be equal to $\sigma_T^2 = 1$ to achieve statistical efficiency.

Another major conclusion we can draw from the right panel of Figure 1 is that the uninformative working prior results in a retrieval that is fairly close in performance to that of the true prior when the state-space SNR is high (here, the blue curve, where SNR = 2). This agrees well with intuition because when SNR is high, there is more information in the data, and we can afford *not* to inject additional information in

the form of a small working prior variance σ_w^2. In contrast, when SNR is low (here, the green curve, where SNR = 0.5), an uninformative working prior does not work nearly as well; with less information in the data, a smaller working prior variance σ_w^2 is needed for a retrieval that has acceptable variability.

So far, we have discussed the behavior of the true retrieval uncertainty as a function of the working prior variance. We now compare the true retrieval uncertainty $s_T^2(x_w, \sigma_w^2)$ and the working retrieval uncertainty $s_w^2(x_w, \sigma_w^2)$, obtained from the retrieval algorithm. Assume the univariate retrieval model given by (16); then by substituting $r = 1$ into (12), we obtain the univariate working retrieval uncertainty:

$$s_w^2(x_w, \sigma_w^2) \quad = \quad \left(\frac{1}{\sigma_w^2} + \frac{k^2}{\sigma_\epsilon^2} \right)^{-1}. \tag{19}$$

In Figure 2, we plot (18) and (19) in three panels for the three choices of state-space SNRs, namely SNR $\in \{0.5, 1, 2\}$. One conclusion we can draw is that the working retrieval uncertainty (red line) can either underestimate or overestimate the true retrieval uncertainty (black line), depending on whether $\sigma_w^2 > \sigma_T^2$ or $\sigma_w^2 < \sigma_T^2$, and the only two instances where they are the same are when $\sigma_w^2 = \sigma_T^2$ or when $\sigma_w^2 \to \infty$ (uninformative working prior). Consequently, the OE retrieval uncertainty estimate is only *statistically valid* when the working prior variance σ_w^2 is correct ($= \sigma_T^2$) or when it is uninformative. Figure 2 also succinctly illustrates the trade-off between OE and maximum likelihood; maximum likelihood ($\sigma_w^2 \to \infty$) has the advantage of uncertainty estimates always being *valid*, though at the cost of the retrievals not being *statistically efficient* (i.e., does not achieve the minimum shown for the black line

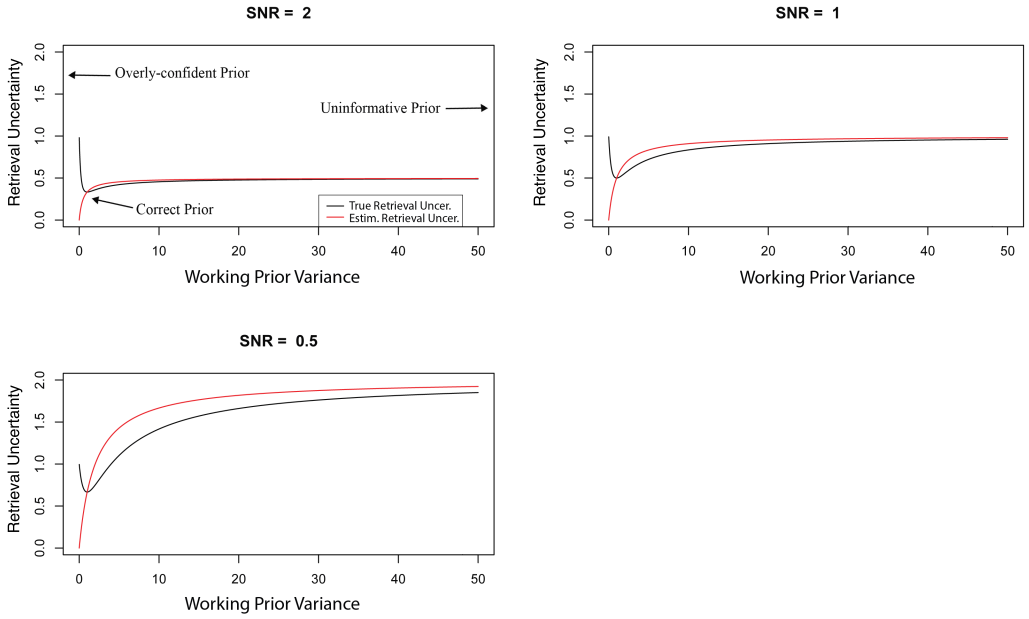

Figure 2: Working retrieval uncertainty s_w^2 given by (19) (red lines) and the true retrieval uncertainty s_T^2 given by (18) (black lines) as a function of the working prior variance σ_w^2 for three choices of state-space SNRs: 2 (top left), 1 (top right), and 0.5 (bottom left). In the univariate model, the true prior variance is $\sigma_T^2 = 1$.

in each of the three panels). This makes sense intuitively, since OE uses information from both the data and the prior, while maximum likelihood only uses information from the data. Assuming that the working prior variance is correct, then OE is clearly more efficient than maximum likelihood due to having the extra component of prior information. Since maximum likelihood is completely insulated from any potentially incorrect assumption about the prior (both mean and variance), its uncertainty estimates are always valid.

We now return to the fully general multivariate retrieval and its uncertainty. The next two subsections address efficiency and uncertainty validity of OE retrievals in the multivariate case.

2.5 Efficiency of OE under the true prior

Generalizing from the univariate case, we wish to show that the OE retrieval under the true prior, where $\{\mathbf{x}_w, \mathbf{S}_w\} = \{\mathbf{x}_T, \mathbf{S}_T\}$, has the 'smallest' true retrieval uncertainty for all possible choices of $\{\mathbf{x}_w, \mathbf{S}_w\}$. That is, we wish to show that $\mathbf{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T) \leq \mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)$, for all $\mathbf{x}_w$ and $\mathbf{S}_w$. From (12) and (14), this efficiency result is equivalent to the following proposition:

Proposition 1. *Under the definitions given in Section 2.1,*

$$(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \leq (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}. \quad (20)$$

Proof. See Appendix A.

This result indicates that $\mathbf{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T)$ is the 'smallest variance' possible for all

estimators arising from the cost function given by (4), and hence we say that the OE retrieval is efficient under the true prior and is generally *inefficient* under any working prior for which $\mathbf{S}_w \neq \mathbf{S}_T$. Proposition 1 holds regardless of whether a Bayesian approach or a Twomey-Tikhonov approach is used to choose $\mathbf{S}_w$.

We note that in many applications, the state vector $\mathbf{x}$ is converted to a different geophysical quantity through a linear combination. For instance, the OCO-2 instrument retrieves a 58-dimensional (57-dimensional for ocean observations) state vector that consists of a 20-level CO_2 profile, surface air pressure, surface albedos, aerosol profile, temperature scaling, humidity scaling, wavelength offset and scaling, and fluorescence. Typically, this state vector $\mathbf{x}$ is convolved into a single value called total-column carbon dioxide (XCO_2) using a linear pressure weighting vector $\mathbf{h}$; that is, $XCO_2 = \mathbf{h}'\mathbf{x}$. Since the matrix inequality, $\mathbf{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T) \leq \mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)$, is defined as $\mathbf{a}'\,\mathbf{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T)\,\mathbf{a} \leq \mathbf{a}'\,\mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)\,\mathbf{a}$ for all column vectors $\mathbf{a}$, it follows that this efficiency proposition holds true for geophysical products that are linear combinations of the state vector $\mathbf{x}$ (e.g., XCO_2 from the OCO-2 retrieval).

It is also important to note that validation studies often use the mean squared error (MSE) as a measure of uncertainty. Recall that the MSE can be written as

$$MSE = \mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w)\mathbf{b}_T(\mathbf{x}_w, \mathbf{S}_w)' + \mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w).$$

Proposition 1 shows that the second term, $\mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)$, is at a global minimum if $\mathbf{S}_w = \mathbf{S}_T$. In Section 2.2, we showed that if $\mathbf{x}_w = \mathbf{x}_T$, the bias is equal to $\mathbf{0}$, which implies that the first term is at a global minimum when $\mathbf{x}_w = \mathbf{x}_T$. Combining the

two results, we see that the MSE is at a global minimum when $\{\mathbf{x}_w, \mathbf{S}_w\} = \{\mathbf{x}_T, \mathbf{S}_T\}$.

Clearly, one of the advantages of the OE estimator with an informative prior is the potential to have the best of both worlds. That is, from (15) and (20), we see that when an OE algorithm uses the correct prior covariance matrix, its retrievals are statistically *efficient*, and its retrieval uncertainties are *valid* (validity is discussed below in Section 2.6). However, we note that this is by no means guaranteed, as indicated in Figure 2 where it is seen that using a 'bad' working prior (e.g., using an overly 'large' prior when the state-space SNR is low) results in the worst of both worlds, namely OE retrievals that are inefficient with retrieval uncertainties that are not valid. To avoid this, we give some recommendation in Section 4 on how to design a working prior based on these theoretical results.

2.6 Validity of the OE retrieval uncertainties

We have seen in the univariate case that when the working prior variance σ_w^2 approaches infinity, the working retrieval uncertainty approaches the true retrieval uncertainty. In the multivariate case, this property is equivalent to $\mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w) \rightarrow \mathbf{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w)$ when $\mathbf{S}_w \rightarrow \boldsymbol{\infty}$ (i.e., the uninformative prior). Unfortunately, using this uninformative prior does not take into account any knowledge one might have about the true prior covariance matrix $\mathbf{S}_T$, resulting in a retrieval that is inefficient (Section 2.5).

We define *validity* of retrieval uncertainty as:

$$\mathbf{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w) = \mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w),$$

which we now discuss for OE. Cressie et al. (2017) proved this validity property for $\mathbf{S}_w \to \boldsymbol{\infty}$ and applied it to the AIRS CO_2 retrieval algorithm. For completeness, we sketch the proof below using the notation summarized in Table 1. Let $\mathbf{S}_w \to \boldsymbol{\infty}$ in (14); then

$$\begin{aligned} \boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w) &\to (\mathbf{0} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{0} \cdot \mathbf{S}_T + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{0} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \\ &= (\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}. \end{aligned} \tag{21}$$

Similarly, let $\mathbf{S}_w \to \boldsymbol{\infty}$ in (12); then

$$\boldsymbol{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w) \to (\mathbf{0} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} = (\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}, \tag{22}$$

which is identical to (21). That is, using an uninformative working prior always produces valid retrieval uncertainties, which is the result given in Cressie et al. (2017). Contrast this with OE retrievals where an informative working prior is used, which has the potential for efficiency and validity (but may result in neither). The uninformative prior gives up efficiency in exchange for guaranteed validity.

In principle then, an OE practitioner could try to leverage some of the properties that result from using an uninformative prior by intentionally making $\mathbf{S}_w$ 'larger' than the best current understanding of $\mathbf{S}_T$. This is precisely what happens in many OE applications where some components of the prior covariance matrix are assigned unrealistically large values, such as the CO_2 components of the prior covariance matrix in OCO-2's XCO_2 retrieval (Boesch et al., 2015). According to the theory developed in

this section, such a strategy trades off a marginal decrease in efficiency of the retrieval for a marginal increase in validity of the retrieval uncertainty. Hence, when designing a working prior covariance $\mathbf{S}_w$, this trade-off should be guided by the state-space signal-to-noise ratio, which can be obtained by comparing the state-space measurement-error variability, $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$, to the science team's intuitive understanding of $\mathbf{S}_T$. More discussion and recommendations are given in Section 4.

3 Simulated data using true priors and CO_2 retrievals using misspecified priors

Having explored the theoretical implications of prior misspecification in Section 2, in this section we demonstrate the consequences of prior misspecification in a simulation using data from an Observing System Simulation Experiment (OSSE) for CO_2 retrievals with a linearized, streamlined version of the OCO-2 forward model (also called a surrogate model; see Hobbs et al., 2017). The OCO-2 satellite was launched by NASA in July 2014 with the goal of providing high-resolution estimates of total-column carbon dioxide (XCO_2). It is a near-infrared (IR) instrument measuring reflected solar radiation in three IR bands, resulting in a radiance vector of dimension $N = 3048$.

In our simulation, we make use of the OCO-2 surrogate model in Hobbs et al. (2017), which "makes some simplification for interpretability and computational efficiency while attempting to maintain the key components of the state vector and RT [radiative transfer] that contribute substantially to uncertainty in XCO_2." The

surrogate model has $N = 3048$ and $r = 39$; that is, $\mathbf{x}$ is a 39-dimensional state vector consisting of a 20-level CO_2 profile, surface air pressure, surface albedo, and aerosol profiles. For an overview of the surrogate model and its parameterization of the state vector, see Section 3 of Hobbs et al. (2017).

In this OSSE, we first designated a known distribution as the true prior, and we repeatedly sampled 1000 times the true state $\mathbf{x}$ from this true prior distribution. Here, the true prior, $\{\mathbf{x}_T, \mathbf{S}_T\}$, that we used is the sample mean and sample covariance of 5000 *retrieved states* obtained after simulation from a non-linear control case (Hobbs et al., 2017, Section 4.3). Each true state $\mathbf{x}$ from the OSSE was then put into the linearized surrogate forward model to produce a noise-free radiance vector. Then a vector of radiance measurement error was sampled and added to the noise-free vector to produce the noisy radiance data vector $\mathbf{y}$. Finally, from $\mathbf{y}$ we obtained the retrieved state vector, $\hat{\mathbf{x}}_w$, using a working prior distribution; see (7).

We note that the surrogate forward model in Hobbs et al. (2017) is non-linear, so in the simulation we use a linearized version of it. That is, we assume that $\mathbf{F}(\mathbf{x}) = \mathbf{c} + \mathbf{K}\mathbf{x}$, where $\mathbf{K}$ is a Jacobian matrix chosen from one of the 5000 retrievals from the control case in Hobbs et al. (2017), and $\mathbf{c} = \mathbf{F}(\mathbf{x}_T) - \mathbf{K}\mathbf{x}_T$. For the simulation, we select the Jacobian matrix $\mathbf{K}$ using a typical Jacobian from the non-linear OCO-2 surrogate model used by Hobbs et al. (2017). Because the forward model here is the same over all 1000 samples, and it is linear, this simulation exercise can be considered an OSSE 'simplification' of the atmosphere.

Hence, the OSSE produces 1000 true states $\mathbf{x}$, 1000 corresponding noisy radiance data vectors $\mathbf{y}$, and 1000 corresponding retrieved states $\hat{\mathbf{x}}_w$. The working prior

$\{\mathbf{x}_w, \mathbf{S}_w\}$ that we use to obtain $\hat{\mathbf{x}}_w$ is based on the operational prior for OCO-2, which depends on latitude and time of the OCO-2 sounding and on a climatology obtained from the GLOBALVIEW dataset. We chose one such in the OSSE; see the Supplementary Materials. Interested readers can find the priors $\{\mathbf{x}_T, \mathbf{S}_T\}$ and $\{\mathbf{x}_w, \mathbf{S}_w\}$, the pressure weighting vector $\mathbf{h}$, the Jacobian $\mathbf{K}$, and the measurement error matrix $\mathbf{S}_\epsilon$ in the Supplementary Materials.

In Table 2, we show the values of the true and working prior means for all 39 state elements. The standardized difference, defined by the element-wise difference of the working prior mean minus the true prior mean divided by the square root of the true prior variance, is displayed in the last column. The CO_2 elements here represent CO_2 mole-fraction concentrations at 20 different pressure levels in the atmosphere, though recall that these values are convolved into the scalar value called total-column carbon dioxide (XCO_2) using a linear pressure weighting vector $\mathbf{h}$. Here, the difference in XCO_2 from the working minus true prior means (computed as $\mathbf{h}' \cdot (\mathbf{x}_w - \mathbf{x}_T)$) is 3.23 ppm. The standardized differences indicate that the means for the CO_2 block are mostly similar, but the means for the Lambertian mean albedos for the Strong CO_2, Weak CO_2, and O_2 A bands include some very large misspecifications. These choices are deliberate, since we wish to demonstrate the ability of a 'large' $\mathbf{S}_w$ to mitigate a potentially large bias.

The OCO-2 working prior covariance matrix $\mathbf{S}_w$ is assumed to be diagonal for all non-CO_2 elements. To see how different the true and working prior covariances are, we show their correlation plots in Figure 3. Note that $\mathbf{S}_T$, unlike $\mathbf{S}_w$, has dependence between the aerosol, surface albedo, and water elements. We've chosen to show both

Table 2: True prior means and working prior means used in the simulation (first and second column). The standardized difference (SDiff) for each element is defined as the difference of the working prior mean minus true prior mean, divided by the square root of the true prior variance of that element (third column).

Name	**True**	**Working**	**SDiff**
CO_2 Volume Mixing Ratio [means in ppm]			
Vertical Level 1 (Top of Atmosphere)	389.7404	388.9731	-2.6829
Vertical Level 2	395.3024	392.9746	-4.7299
Vertical Level 3	398.1116	394.7076	-5.2564
Vertical Level 4	399.1278	396.0390	-3.8981
Vertical Level 5 (Tropopause)	398.0690	397.1398	-0.9599
Vertical Level 6	396.4378	398.3572	2.4179
Vertical Level 7	396.0817	398.4919	2.7922
Vertical Level 8	395.7496	398.4647	2.8952
Vertical Level 9	395.2420	398.4325	3.1810
Vertical Level 10	394.7879	398.3967	3.3577
Vertical Level 11	393.5765	398.3579	3.8944
Vertical Level 12	392.4954	398.3159	4.2996
Vertical Level 13	391.1232	398.2707	4.7768
Vertical Level 14	390.0250	398.2190	5.1063
Vertical Level 15	389.1317	398.1598	5.1538
Vertical Level 16	388.8229	398.0950	5.0912
Vertical Level 17	389.8204	398.0250	3.8831
Vertical Level 18	391.4878	397.9514	2.6177
Vertical Level 19	397.4609	397.8780	0.1217
Vertical Level 20 (Surface)	401.3001	397.8112	-0.6676
Surface Pressure [hPa]	998.7413	1002	1.4769
Lambertian Albedo [units of means are in the Suppl. Mat.]			
Strong CO_2 Band Mean Albedo	0.6496	0.1753	-273.7585
Strong CO_2 Band Albedo Spectral Slope	0	0	.00
Weak CO_2 Band Mean Albedo	0.6755	0.2560	-212.0764
Weak CO_2 Band Albedo Spectral Slope	0	0	.00
O_2 A-Band Mean Albedo	0.5183	0.1827	-146.3876
O_2 A-Band Mean Albedo Spectral Slope	0	0	0
Aerosols [units of means are in the Suppl. Mat.]			
Dust Log Aerosol Optical Depth	-2.4760	-3.3178	-4.5838
Dust Profile Height	0.8982	0.9000	0.0301
Dust Log Profile Thickness	-3.6365	-2.9957	1.8230
Sea Salt Log Aerosol Optical Depth	-3.8290	-4.0140	-0.9278
Sea Salt Profile Height	0.7478	0.9000	2.7018
Sea Salt Log Profile Thickness	-2.0716	-2.9957	-3.0608
Cloud Ice Log Aerosol Depth	-2.8718	-4.3820	-6.1065
Cloud Ice Profile Height	0.2208	0.3000	3.8450
Cloud Ice Log Profile Thickness	-3.2129	-3.2189	-0.1704
Cloud Water Log Aerosol Depth	-4.0925	-4.3820	-0.4233
Cloud Water Profile Height	0.5531	0.7500	1.1633
Cloud Water Log Profile Thickness	-2.3013	-2.3026	-0.0995

of these plots in correlation space because these matrices in the original covariance space have vastly different magnitudes for almost all elements of the state vector. For instance, the CO_2 variance at Earth's surface in the true prior is $(5.22 \text{ ppm})^2$, while the corresponding CO_2 variance at Earth's surface in the working prior is $(47.7 \text{ ppm})^2$. In the bottom row of Figure 3, we illustrate the relative sizes of the diagonals of $\mathbf{S}_w$ and $\mathbf{S}_T$ (i.e., the prior variances) by plotting their element-wise ratio at each of the 39 state elements in log scale. It is evident that the OCO-2 team operational prior inflates the working prior variance of most of the 39 elements by several orders of magnitude, with the Lambertian Albedo elements (index 22-27) being particular large relative to the corresponding components in the true prior covariance matrix. The only two exceptions to this are Dust Log Profile Thickness and Sea Salt Log Profile Thickness (index 30 and 33, respectively). This is probably because the OCO-2 team has particularly strong confidence in these two elements, reflected by their choice of these elements with small prior variances for them.

This decision to inflate most components of $\mathbf{S}_w$ by several orders of magnitude moves the working prior towards an uninformative prior (see Section 2.6), so that the working retrieval uncertainty should have better validity, although at the expense of statistical efficiency of the retrieval. The uninformative nature of the working prior covariance matrix is noted in the development of the OCO-2 retrieval algorithm (Boesch et al., 2015; O'Dell et al., 2012; Connor et al., 2008).

To see the different influences of the working prior mean vector and the working prior covariance matrix on the retrieval, the simulation experiment is divided into three parts, where we misspecify only the prior mean vector (Experiment 1: working

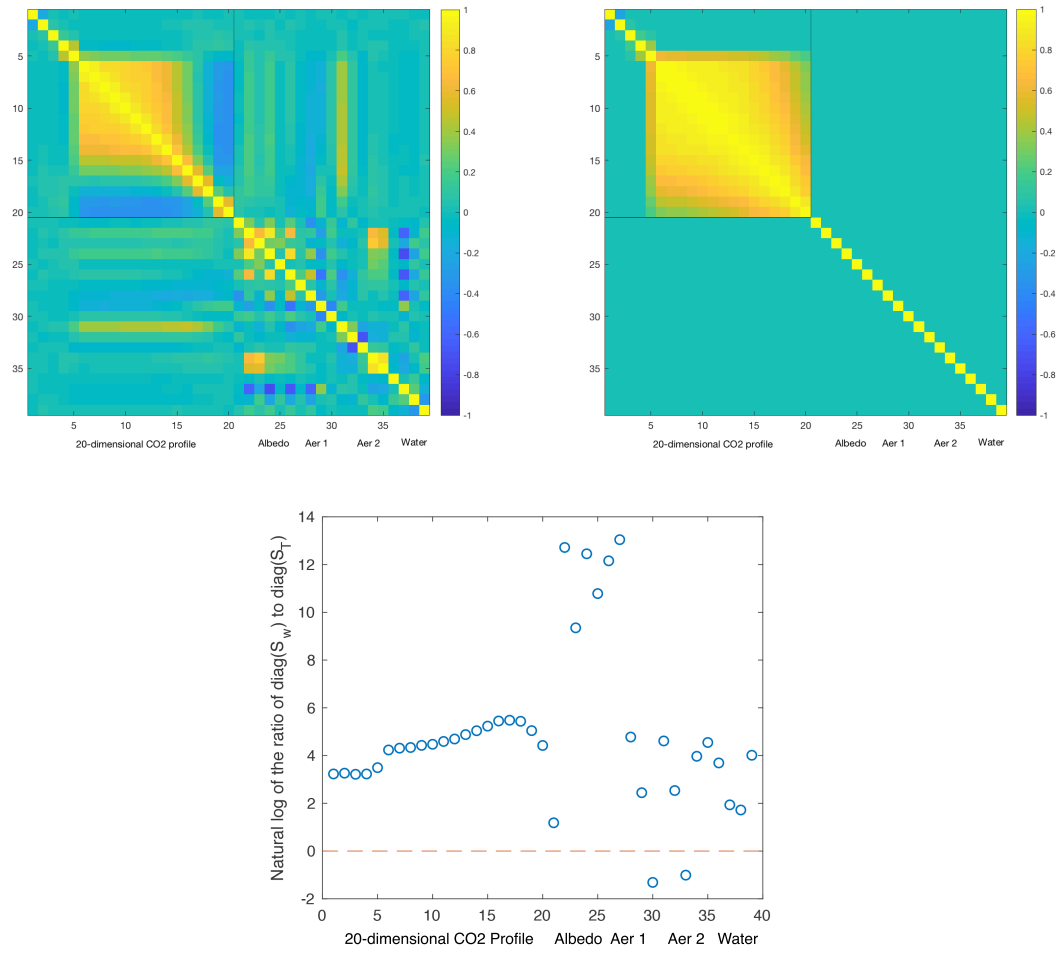

Figure 3: Top row: Plots of the true prior correlation matrix (left panel) and the working prior correlation matrix (right panel) used in the OSSE simulation. Bottom row: Natural log of the element-wise ratio of the diagonals of $\mathbf{S}_w$ to the diagonals of $\mathbf{S}_T$. The red dashed line indicates the dividing line at which the working prior variance is equal to the true prior variance.

prior = $\{\mathbf{x}_w, \mathbf{S}_T\}$), where we misspecify only the prior covariance matrix (Experiment 2: working prior = $\{\mathbf{x}_T, \mathbf{S}_w\}$), and where we misspecify both (Experiment 3: working prior = $\{\mathbf{x}_w, \mathbf{S}_w\}$). The steps for our simulation experiments are as follows:

0. Select a working prior from one of the three possibilities.
1. Sample a state $\mathbf{x}$ from the true prior distribution $\{\mathbf{x}_T, \mathbf{S}_T\}$.
2. Compute the radiance $\mathbf{y}$ using the model given by (3).
3. With the selected working prior, compute the retrieved XCO_2 and the retrieval uncertainty (specifically, $\mathbf{h}'\hat{\mathbf{x}}_w$ and $\mathbf{h}'\mathbf{\Sigma}_w(\mathbf{x}_w, \mathbf{S}_w)\mathbf{h}$) using (7) and (12), respectively.
4. Repeat steps 1-3 for 1000 iterations.

The summary statistics of the differences between the retrieved XCO_2 and the true XCO_2 under the three experiments are shown in Table 3. In Experiment 1, where only the prior mean is misspecified, the retrieval bias obtained from the simulation is 22.04 ppm! Table 3 shows that this agrees with a calculation based on the theoretical value given by (10). This large retrieval bias is somewhat counter-intuitive, given that the misspecification of the prior mean of XCO_2 (that is, $\mathbf{h}' \cdot (\mathbf{x}_w - \mathbf{x}_T)$) is only 3.23 ppm. However, we note that the working prior mean also includes surface pressure, aerosols, and albedo, and in this instance the misspecification of these non-CO_2 elements has pushed the retrieval bias above 22 ppm. Some sensitivity analysis showed that a large part of this discrepancy is due to the mean albedo components used for the Strong

CO_2, Weak CO_2, and O_2 A bands, which in the OSSE were deliberately misspecified as indicated by the SDiff column in Table 2.

Table 3: Simulation summary statistics for XCO_2. Both the bias and the uncertainty (here expressed as a standard deviation) have units of ppm. Estimates that are consistent with the corresponding confidence intervals are colored red. The true retrieval bias and true retrieval uncertainties are computed using the derivations in Section 2.

	Experiment 1	Experiment 2	Experiment 3
Working prior	$\{\mathbf{x}_w, \mathbf{S}_T\}$	$\{\mathbf{x}_T, \mathbf{S}_w\}$	$\{\mathbf{x}_w, \mathbf{S}_w\}$
Bias from simulation	22.04	0.01	.40
95% CI for bias	[22.02, 22.06]	[−0.02, 0.05]	[.36, .44]
True retrieval bias	22.04	0	.41
Uncertainty from simulation	0.30	0.61	0.60
95% CI for uncertainty	[0.29, 0.31]	[0.58, 0.64]	[0.57, 0.63]
True retrieval uncertainty	0.31	0.62	0.62
Working uncertainty	0.31	0.69	0.69
RMSE from simulation	22.04	.61	.72

Since there are 1000 simulated retrievals for each experiment, we could estimate a 95% confidence interval for the retrieval bias. We chose to use a nonparametric bootstrap based on 500 samples to do this (Efron, 1981). In Experiment 1, we misspecified only the prior mean, and the simulation gave a retrieval bias of 22.04 ppm. As can be seen from Table 3, the empirical 95% confidence interval for the retrieval bias in Experiment 1 is [22.02 ppm, 22.06 ppm], which is consistent with the theoretical retrieval bias of 22.04 ppm calculated from (10). We also display the corresponding statistics for the retrieval uncertainty (in units of standard deviation) in the lower half of Table 3. In Experiment 1, where $\mathbf{S}_w = \mathbf{S}_T$, the analytical derivations show that

the simulated retrieval uncertainty, the true retrieval uncertainty, and the working retrieval uncertainty should all be consistent with one another. From Table 3, we see that the true retrieval uncertainty is the same as the working retrieval uncertainty (0.31 ppm), both of which are consistent with the simulated retrieval uncertainty (0.30 ppm) and its 95% confidence interval.

In Experiment 2, we misspecified only the prior covariance matrix, and the simulation gave a retrieval bias of 0.02 ppm. As we noted in Section 2.2, $\mathbf{x}_w = \mathbf{x}_T$ is a sufficient condition for unbiasedness, so the true retrieval bias under this experiment should be 0. Indeed, the 95% confidence interval of the bias for this experiment is [−0.02 ppm, 0.05 ppm], which is consistent with the true value of 0. With regard to validity, the calculated working retrieval uncertainty (0.69 ppm) is about 12 % larger than the theoretical retrieval uncertainty (0.62 ppm). The retrieval uncertainty from simulation is 0.61 and the 95% confidence interval is [0.58 ppm, 0.64 ppm], which is consistent with the true retrieval uncertainty of 0.62 ppm but not the calculated working value of 0.69 ppm. This experiment reinforces our validity results in Section 2.3, namely that when an informative prior covariance matrix is misspecified, the working retrieval uncertainty is incorrect.

In Experiment 3, we misspecified both the prior mean vector and the prior covariance matrix. From Table 3, the outcome is a mixture of Experiment 1 and Experiment 2, namely that the working retrieval has both a bias and a retrieval uncertainty that is not valid. The trade-off between bias and variance is best captured in the square root of the mean squared error (or RMSE), which here is calculated from the simulation and is displayed on the last row of Table 3. The RMSE is largest (22.04 ppm)

when the working prior mean is incorrect, suggesting that in this experimental setup the RMSE is more sensitive to $\mathbf{x}_w$ than to $\mathbf{S}_w$. However, when a conservative $\mathbf{S}_w$ is applied, the same choice of $\mathbf{x}_w$ has a much smaller RMSE, of 0.72 ppm – see Table 3.

Experiment 3 provides a rationale behind the $\mathbf{S}_w$ used in the operational OCO-2 prior. As was noted earlier in this Section, our choice of $\mathbf{S}_w$ was modeled after the operational OCO-2 prior covariance matrix, where most elements are "unrealistically large for most of the world (all relatively clean-air sites), [in order to impose] a minimal constraint on the retrieved XCO2" (Boesch et al., 2015). In Experiment 1 where $\mathbf{x}_w$ is misspecified but $\mathbf{S}_w$ is not, the result is a bias of 22.04 ppm, but the same choice of $\mathbf{x}_w$ and a misspecified, conservative $\mathbf{S}_w$ in Experiment 3 results in a *greatly mitigated bias* of 0.41 ppm, about 50 times smaller! This implies that the operational OCO-2 retrieval, in its choice of working prior covariance matrix, is *quite robust to bias caused by using the wrong prior mean.* We note that this attractive bias property comes with efficiency and validity trade-offs, which are discussed in Sections 2.5 and 2.6.

4 Discussion and summary

In this paper, we give an in-depth investigation of the bias and uncertainty of retrievals from Optimal Estimation (OE), when the prior distribution of the state is misspecified. In many remote sensing applications, the true priors are multivariate and hard to characterize properly, and a pragmatic approach is typically taken in designing the working prior $\{\mathbf{x}_w, \mathbf{S}_w\}$. That approach is a mixture of computational need for expediency, subject-matter expertise, and existing empirical data. In other

words, the prior distributions within many OE application are typically constructed as a combination of the regularization approach (i.e., Twomey-Tikhonov constraint) and the Bayesian approach (i.e., distribution of the state). However, the retrieval uncertainties arising therefrom are almost universally interpreted within the Bayesian approach, albeit incorrectly. Here, our aim has been to show how this leads to biases and inaccuracies in OE retrievals and their uncertainties. We have done this by explicitly separating the true prior distribution, $\{\mathbf{x}_T, \mathbf{S}_T\}$, from the working prior distribution, $\{\mathbf{x}_w, \mathbf{S}_w\}$, and computing the true retrieval bias, $\mathrm{E}_T(\hat{\mathbf{x}}_w - \mathbf{x})$, and the true retrieval uncertainty, $\mathrm{var}_T(\hat{\mathbf{x}}_w - \mathbf{x})$. Our key findings can be summarized as follows:

- When the prior mean is misspecified (i.e., $\mathbf{x}_w \neq \mathbf{x}_T$), there is a resulting bias that is given by $(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{S}_w^{-1}(\mathbf{x}_w - \mathbf{x}_T)$. This bias can be reduced in magnitude by 'increasing' $\mathbf{S}_w$ (that is, making the working prior covariance matrix less informative).

- When the prior covariance is misspecified (i.e., $\mathbf{S}_w \neq \mathbf{S}_T$, where $\mathbf{S}_w^{-1} \neq \mathbf{0}$), then the working retrieval uncertainty of the retrieval will not be valid with respect to the true retrieval uncertainty.

- The extreme limiting case of making $\mathbf{S}_w$ less informative is $\mathbf{S}_w^{-1} = \mathbf{0}$ (equivalently $\mathbf{S}_w \to \boldsymbol{\infty}$). This is the uninformative prior that is implicitly used in a maximum-likelihood (also called least-squares) approach. We show that the uninformative prior results in a retrieval uncertainty that has the attractive property of being *valid* (i.e., having an accurate working retrieval uncertainty) and *unbiased*. However, the OE framework with an informative working prior

that is specified correctly has the advantage of being *efficient* (i.e., having the smallest possible retrieval-error variance, calculated using the true prior), *valid*, and a retrieval that is *unbiased*.

- Importantly, with a 'bad' choice of prior, OE can have the worst of both worlds, being both *not efficient* and *not valid*. A compromise between the potential efficiency of OE and the guaranteed validity of maximum likelihood is obtained by erring on the 'large' side when setting the prior covariance matrix. This practice of inflating the prior covariance matrix to 'relax' constraints on the retrieval can be interpreted as trading some amount of efficiency for an increase in validity. Given the complicated settings, perhaps the best that the OE practitioner can hope for is an estimator that is 'mostly' efficient and 'mostly' valid.

- The design of a working prior distribution should take into account the relative 'size' of the signal $\mathbf{S}_T$ and the noise component $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$. The latter can be computed easily, and it should be always be examined in order to have an idea of the contribution of the radiance noise in the state space. While the exact form of $\mathbf{S}_T$ is typically not known, in practice there are rough bounds available for the variability of each component of the state vector, and they can be compared to respective elements of $(\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}$ to obtain bounds on the state-space signal-to-noise ratio.

- When the signal components dominate (that is, when signal-to-noise ratios are larger than 1), then we could afford to use a less informative prior. If signal-to-noise ratios are much less than 1, then we recommend designing a more

constrained prior with $\mathbf{S}_w$ 'smaller' but hopefully close to $\mathbf{S}_T$.

Thus far, we have considered the impact of prior misspecification in the case of a *linear* forward model. For many applications, the forward model is non-linear, and the MAP solution is obtained using iterative least-squares methods such as the Levenberg-Marquardt algorithm. In this situation, the results we have derived still hold if the loss function is linearizable around a deterministic vector and if the iterative non-linear solver can find its way to this linearized neighborhood. These two requirements also apply to the usual OE uncertainty estimate, since it is also derived under a linearization assumption (Rodgers, 2000, Section 5.5). Cressie et al. (2016) gave a linear and quadratic Taylor approximation when the forward model is non-linear. Finally, the non-linear solver can be expected to also have error from convergence criteria and finding local rather than global minima.

Acknowledgement

The authors would like to thank Mike Turmon at the Jet Propulsion Laboratory for his insightful comments on the manuscript, and Rui Wang for initial discussions on prior misspecification. Nguyen's and Hobbs' research were performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. Cressie's research was performed under NASA ROSES NNH17ZDA001N and Australian Research Council Discovery Projects, DP150104576 and DP190100180.

Appendices

A Proof of Proposition 1

Proposition. *Under the definitions given in Section 2.1,* $\mathbf{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T) \leq \mathbf{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)$ *for all* $\mathbf{x}_w$ *and* $\mathbf{S}_w$*, or equivalently,*

$$(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \leq (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}.$$

Proof. The proof relies on the observation that this proposition is related to the Schur complement (e.g., Horn and Zhang, 2005). For a symmetric matrix

$$\mathbf{X} \equiv \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{pmatrix},$$

where $\mathbf{C}$ is invertible, the Schur complement of $\mathbf{C}$ in $\mathbf{X}$ is defined as $\mathbf{X}/\mathbf{C} \equiv \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}'$. The Schur-complement theorem states that if $\mathbf{C} > \mathbf{0}$, then $\mathbf{X} \geq \mathbf{0}$ if and only if its Schur complement $\mathbf{X}/\mathbf{C} \geq \mathbf{0}$ (e.g., Horn and Zhang, 2005, Theorem 1.12).

Now consider the matrix,

$$\mathbf{E} = \begin{pmatrix} \mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} & \mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} \\ \mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} & \mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} \end{pmatrix}. \tag{23}$$

We can rewrite $\mathbf{E}$ as the sum of two symmetric matrices,

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2,$$

where

$$\mathbf{E}_1 = \begin{pmatrix} \mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} & \mathbf{S}_w^{-1} \\ \mathbf{S}_w^{-1} & \mathbf{S}_T^{-1} \end{pmatrix},$$

and

$$\mathbf{E}_2 = \begin{pmatrix} \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} & \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} \\ \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} & \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} \end{pmatrix} = \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K} \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{24}$$

First, consider the term $\mathbf{E}_1$: we see that $\mathbf{S}_T^{-1} > \mathbf{0}$, since $\mathbf{S}_T$ is positive-definite, and that its Schur complement, $\mathbf{E}_1/\mathbf{S}_T^{-1} = \mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} - \mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} = \mathbf{0}$. Therefore by the Schur-complement theorem, $\mathbf{E}_1 \geq \mathbf{0}$. Second, consider the term $\mathbf{E}_2$: from (24), we see that $\mathbf{E}_2$ is the Kronecker product of $\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}$ and the 2×2 matrix of all 1's, both of which are positive-semidefinite. Since the Kronecker product of two positive-semidefinite matrices is also positive-semidefinite (Petersen and Pedersen, 2008, Section 10.2.1), then $\mathbf{E}_2 \geq \mathbf{0}$. Hence, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 \geq \mathbf{0}$.

From (23), given that $(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}) > \mathbf{0}$ and $\mathbf{E} \geq \mathbf{0}$, then by the Schur-

complement theorem, $\mathbf{E}/(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}) \geq \mathbf{0}$. Consequently,

$$\begin{aligned}
\mathbf{0} &\leq \mathbf{E}/(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}), \\
\mathbf{0} &\leq (\mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}) - (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}),
\end{aligned}$$

and hence

$$(\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \leq (\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_w^{-1}\mathbf{S}_T\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})(\mathbf{S}_w^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}.$$

That is,

$$\boldsymbol{\Sigma}_T(\mathbf{x}_T, \mathbf{S}_T) \leq \boldsymbol{\Sigma}_T(\mathbf{x}_w, \mathbf{S}_w)\,, \quad \text{for all } \{\mathbf{x}_w, \mathbf{S}_w\}\,. \quad \square$$

Supplementary Materials

Online supplementary materials for this paper are available at `https://hpc.niasra.uow.edu.au/files/SupplementaryMaterials.zip`.

References

Boesch, H., Brown, L., Castano, R., Christi, M., Crisp, D., Eldering, A., Fisher, B., Frankenberg, C., Gunson, M., Granat, R., et al. (2015). Orbiting Carbon Observatory (OCO-2) Level 2 Full Physics Algorithm Theoretical Basis Document. URL: `https://docserver.gesdisc.eosdis.nasa.gov/public/project/`

OCO/OCO2_L2_ATBD.V6.pdf.

Bowman, K. W., Rodgers, C. D., Kulawik, S. S., Worden, J., Sarkissian, E., Osterman, G., Steck, T., Lou, M., Eldering, A., Shephard, M., Worden, H., Lampel, M., Clough, S., Brown, P., Rinsland, C., Gunson, M., and Beer, R. (2006). Tropospheric Emission Spectrometer: Retrieval method and error analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 44(5):1297–1307.

Connor, B. J., Boesch, H., Toon, G., Sen, B., Miller, C., and Crisp, D. (2008). Orbiting Carbon Observatory: Inverse method and prospective error analysis. *Journal of Geophysical Research: Atmospheres*, 113(D5).

Cressie, N. (2018). Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide (with discussion). *Journal of the American Statistical Association*, 113(521):152–168.

Cressie, N., Wang, R., and Maloney, B. (2017). The Atmospheric Infrared Sounder retrieval, revisited. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1504–1507.

Cressie, N., Wang, R., Smyth, M., and Miller, C. E. (2016). Statistical bias and variance for the regularized inverse problem: Application to space-based atmospheric CO_2 retrievals. *Journal of Geophysical Research: Atmospheres*, 121(10):5526–5537.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Engelen, R. J., Denning, A. S., and Gurney, K. R. (2002). On error estimation

in atmospheric CO2 inversions. *Journal of Geophysical Research: Atmospheres*, 107(D22):ACL–10.

Govaerts, Y., Wagner, S., Lattanzio, A., and Watts, P. (2010). Joint retrieval of surface reflectance and aerosol optical depth from MSG/SEVIRI observations with an optimal estimation approach: 1. Theory. *Journal of Geophysical Research: Atmospheres*, 115(D2).

Hobbs, J., Braverman, A., Cressie, N., Granat, R., and Gunson, M. (2017). Simulation-based uncertainty quantification for estimating atmospheric CO_2 from satellite data. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):956–985.

Horn, R. A. and Zhang, F. (2005). Basic properties of the Schur complement. In Zhang, F., editor, *The Schur Complement and Its Applications*, pages 17–46. Springer, Boston, MA.

Irion, F., Kahn, B., Schreier, M., Fetzer, E., Fishbein, E., Fu, D., Kalmus, P., Wilson, R., Wong, S., and Yue, Q. (2018). Single-footprint retrievals of temperature, water vapor and cloud properties from AIRS. *Atmospheric Measurement Techniques*, 11:971–995.

Kulawik, S. S., Bowman, K. W., Luo, M., Rodgers, C. D., and Jourdain, L. (2008). Impact of nonlinearity on changing the a priori of trace gas profile estimates from the Tropospheric Emission Spectrometer (TES). *Atmospheric Chemistry and Physics*, 8(12):3081–3092.

Luo, M., Rinsland, C., Rodgers, C. D., Logan, J., Worden, H., Kulawik, S. S., Eldering, A., Goldman, A., Shephard, M., Gunson, M., and Lampel, M. (2007). Comparison of carbon monoxide measurements by TES and MOPITT: Influence of a priori data and instrument characteristics on nadir atmospheric species retrievals. *Journal of Geophysical Research: Atmospheres*, 112(D9).

Merchant, C., Le Borgne, P., Roquet, H., and Legendre, G. (2013). Extended optimal estimation techniques for sea surface temperature from the Spinning Enhanced Visible and Infra-Red Imager (SEVIRI). *Remote Sensing of Environment*, 131:287–297.

O'Dell, C. W., Connor, B., Boesch, H., O'Brien, D., Frankenberg, C., Castano, R., Eldering, A., Fisher, B., Gunson, M., McDuffie, J., Miller, C. E., Natraj, V., Oyafuso, F., Polonsky, I., Smyth, M., Taylor, T., Toon, G. C., Wennberg, P. O., and Wunch, D. (2012). The ACOS CO_2 retrieval algorithm – Part 1: Description and validation against synthetic observations. *Atmospheric Measurement Techniques*, 5:99–121.

O'Dell, C. W., Eldering, A., Wennberg, P. O., Crisp, D., Gunson, M. R., Fisher, B., Frankenberg, C., Kiel, M., Lindqvist, H., Mandrake, L., Merrelli, A., Natraj, V., Nelson, R. R., Osterman, G. B., Payne, V. H., Taylor, T. R., Wunch, D., Drouin, B. J., Oyafuso, F., Chang, A., McDuffie, J., Smyth, M., Baker, D. F., Basu, S., Chevallier, F., Crowell, S. M. R., Feng, L., Palmer, P. I., Dubey, M., García, O. E., Griffith, D. W. T., Hase, F., Iraci, L. T., Kivi, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Roehl, C. M., Sha, M. K., Strong, K., Sussmann, R., Te, Y., Uchino,

O., and Velazco, V. A. (2018). Improved retrievals of carbon dioxide from the Orbiting Carbon Observatory-2 with the version 8 ACOS algorithm. *Atmospheric Measurement Techniques Discussions*, 2018:1–57.

Petersen, K. B. and Pedersen, M. S. (2008). The Matrix Cookbook. *Technical University of Denmark*, 7(15):510.

Ramanathan, A. K., Nguyen, H. M., Sun, X., Mao, J., Abshire, J. B., Hobbs, J. M., and Braverman, A. J. (2018). A Singular Value Decomposition framework for retrievals with vertical distribution information from greenhouse gas column absorption spectroscopy measurements. *Atmospheric Measurement Techniques Discussions*, 2018:1–35.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Press, Singapore.

Su, Z., Yung, Y. L., Shia, R.-L., and Miller, C. E. (2017). Assessing accuracy and precision for space-based measurements of carbon dioxide: An associated statistical methodology revisited. *Earth and Space Science*, 4(3):147–161.

Susskind, J., Barnet, C. D., and Blaisdell, J. M. (2003). Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data in the presence of clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2):390–409.

Wunch, D., Toon, G. C., Blavier, J.-F. L., Washenfelder, R. A., Notholt, J., Connor, B. J., Griffith, D. W., Sherlock, V., and Wennberg, P. O. (2011). The Total Carbon

Column Observing Network. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369(1943):2087–2112.

Yoshida, Y., Kikuchi, N., Morino, I., Uchino, O., Oshchepkov, S., Bril, A., Saeki, T., Schutgens, N., Toon, G. C., Wunch, D., Roehl, C. M., Wennberg, P. O., Griffith, D. W. T., Deutscher, N. M., Warneke, T., Notholt, J., Robinson, J., Sherlock, V., Connor, B., Rettinger, M., Sussmann, R., Ahonen, P., Heikkinen, P., Kyrö, E., Mendonca, J., Strong, K., Hase, F., Dohe, S., and Yokota, T. (2013). Improvement of the retrieval algorithm for GOSAT SWIR XCO_2 and XCH_4 and their validation using TCCON data. *Atmospheric Measurement Techniques*, 6(6):1533–1547.