Supplementary materials for this article are available at https://doi.org/10.1007/s13253-018-0319-8.

Optimal Estimation Versus MCMC for CO_2 Retrievals

Jenny BRYNJARSDOTTIR, Jonathan HOBBS, Amy BRAVERMAN, and Lukas MANDRAKE

The Orbiting Carbon Observatory-2 (OCO-2) collects infrared spectra from which atmospheric properties are retrieved. OCO-2 operational data processing uses optimal estimation (OE), a state-of-the-art approach to inference of atmospheric properties from satellite measurements. One of the main advantages of the OE approach is computational efficiency, but it only characterizes the first two moments of the posterior distribution of interest. Here we obtain samples from the posterior using a Markov Chain Monte Carlo (MCMC) algorithm and compare this empirical estimate of the true posterior to the OE results. We focus on 600 simulated soundings that represent the variability of physical conditions encountered by OCO-2 between November 2014 and January 2016. We treat the two retrieval methods as ensemble and density probabilistic forecasts, where the MCMC yields an ensemble from the posterior and the OE retrieval result provide the first two moments of a normal distribution. To compare these methods, we apply both univariate and multivariate diagnostic tools and proper scoring rules. The general impression from our study is that when compared to MCMC, the OE retrieval performs reasonably well for the main quantity of interest, the column-averaged CO_2 concentration X_{CO_2}, but not for the full state vector $\mathbf{X}$ which includes a profile of CO_2 concentrations over 20 pressure levels, as well as several other atmospheric properties.

Supplementary materials accompanying this paper appear on-line.

Key Words: Orbiting Carbon Observatory; OCO-2; Remote sensing; Probabilistic forecast; Proper scoring rules; Carbon cycle science; Continuous ranked probability score; Energy score; PIT histogram; Average rank; Band depth rank; Kullback–Leibler divergence.

1. INTRODUCTION

One of the most urgent challenges in Earth science today is to better determine the physical mechanisms that govern carbon sources and sinks around the globe (Crisp et al.

Jenny Brynjarsdottir (✉), Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, Cleveland, OH, USA (E-mail: jenny.brynjarsdottir@case.edu). Jonathan Hobbs, Amy Braverman and Lukas Mandrake, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
(E-mail: Jonathan.M.Hobbs@jpl.nasa.gov; Amy.Braverman@jpl.nasa.gov; lukas.mandrake@jpl.nasa.gov).

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-018-0319-8

Published online: 01 March 2018

2004; Friedlingstein et al. 2006; O'Dell et al. 2012). Despite decades of research, there remain significant uncertainties in many elements of the global carbon cycle and its response to anthropogenic perturbations. Carbon flux inversion models are essential to global carbon cycle science (Battle et al. 2000; Bousquet et al. 2000; Schuh et al. 2010), and they require accurate measurements of CO_2 concentration with both high spatial resolution and global coverage. Satellite instruments are needed to achieve such a data record, current in situ observation networks are not sufficient. Furthermore, according to Miller et al. (2007), an absolute accuracy of 1–2 ppm is required in order to substantially reduce surface flux uncertainties. A big step toward having an ongoing measurement system that satisfies this need is the Orbiting Carbon Observatory-2 (OCO-2) mission (Crisp et al. 2004; Eldering et al. 2017) carried out by the National Aeronautics and Space Administration (NASA) and the Jet Propulsion Laboratory (JPL). Since its launch in July 2014 the OCO-2 instrument has provided measurements and uncertainty estimates of the column-averaged dry air mole fraction, X_{CO_2}, with an unprecedented spatial resolution.

The OCO-2 instrument measures radiances (i.e., reflected sunlight) in a range of wavelengths that are known to be affected by CO_2 and O_2 absorption. The vector of radiances, $\mathbf{Y}$, is then inverted to an estimate of a state vector $\mathbf{X}$ that represents atmospheric conditions at that time and location. This state vector includes CO_2 concentrations at 20 pressure levels of the atmospheric column and about 40 other elements such as surface pressure, albedo and aerosol information. The inversion is performed with a *retrieval algorithm* which includes a physical forward model $\mathbf{F}$, called the *full physics model*, that describes how radiances depend on the atmospheric state. In statistical parlance, the retrieved state vector is an estimate of a parameter vector $\mathbf{X}$ in the statistical model

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{b}) + \boldsymbol{\varepsilon} \tag{1}$$

where the constant vector $\mathbf{b}$ is assumed known and the random vector $\boldsymbol{\varepsilon}$ contains independent normal measurement errors with mean zero and known variances. A multivariate normal prior distribution is assigned to $\mathbf{X}$, and Bayes Theorem is used to estimate $\mathbf{X}$. The OCO-2 mission uses a procedure known as *optimal estimation* (OE) in the remote sensing literature (Rodgers 2000), that finds the posterior mode of $[\mathbf{X} \mid \mathbf{Y}]$, called $\hat{\mathbf{X}}$, using the Levenberg–Marquardt optimization algorithm. An estimate of the posterior covariance matrix $\hat{S}$ is also obtained, using linear approximations. Here, "$[\mathbf{X}]$" and "$[\mathbf{X} \mid \mathbf{Y}]$" denote the probability density function (pdf) of $\mathbf{X}$ and the conditional pdf of $\mathbf{X}$ given $\mathbf{Y}$, respectively. The main quantity of interest is the column-averaged dry air mole fraction, X_{CO_2}, which is a weighted average of the CO_2 concentrations in 20 vertical layers:

$$\hat{X}_{\mathrm{CO}_2} = \mathbf{h}^T \hat{\mathbf{X}}_p \tag{2}$$

where $\hat{\mathbf{X}}_p = (\hat{X}_1, \ldots, \hat{X}_{20})^T$ contains the estimated CO_2 *profile* and $\mathbf{h}$ is a vector of weights that partly depends on other elements of the state vector. The data product also includes the posterior variance of X_{CO_2}:

$$\hat{s}^2_{X_{\mathrm{CO}_2}} = \mathbf{h}^T \hat{S}_p \mathbf{h} \tag{3}$$

where the matrix $\hat{S}_p$ contains the first 20 rows and columns of $\hat{S}$.

The OE method is the state-of-the-art approach to retrievals of atmospheric properties from satellite measurements (Bösch et al. 2011; Crisp et al. 2012; O'Dell et al. 2012; Line et al. 2013; Eldering et al. 2017). One of its main advantages is computational efficiency. The OCO-2 instrument, for example, collects 24 observations every second. Using OE, the OCO-2 retrieval currently takes a few minutes per observation, which is achieved via high-performance computing systems at NASA. It is clear, however, that the OE method does not characterize the full posterior distribution $[\mathbf{X} \mid \mathbf{Y}]$, or the posterior of the quantity of interest $\left[X_{\mathrm{CO}_2} \mid \mathbf{Y}\right]$. For example, while the retrieved state vector $\hat{\mathbf{X}}$ gives the mode of $[\mathbf{X} \mid \mathbf{Y}]$, $\hat{X}_{\mathrm{CO}_2}$ in (2) is not necessarily the mode of $\left[X_{\mathrm{CO}_2} \mid \mathbf{Y}\right]$. In addition, users of the OCO-2 data product, e.g., carbon flux modelers, usually assume that the distribution of X_{CO_2} is normal with mean $\hat{X}_{\mathrm{CO}_2}$ and variance $\hat{s}^2_{X_{\mathrm{CO}_2}}$ (Engelen et al. 2002), and the flux inversion community has been increasingly interested in assimilating the full CO_2 profile $\mathbf{X}_p$ using the multivariate normal distribution $N(\hat{\mathbf{X}}_p, \hat{S}_p)$. The normality assumption is at best an approximation since the forward model $\mathbf{F}$ is not linear, and the posterior distribution $[\mathbf{X} \mid \mathbf{Y}]$ is therefore not multivariate normal. To date, there has been no systematic study of how well the OE algorithm performs in representing the actual posterior distributions of X_{CO_2} or the full state vector $\mathbf{X}$. Furthermore, it is important to characterize how well the OE approximation of the actual posterior performs as an estimate of the true state $X^{\mathrm{true}}_{\mathrm{CO}_2}$ or $\mathbf{X}^{\mathrm{true}}$, as compared to, e.g., the Bayes estimate (posterior mean) which is obtained via MCMC. The performance of OE will of course depend on the particular forward model $\mathbf{F}$, whether $\mathbf{F}(\mathbf{X}, \mathbf{b})$ is roughly linear around $\mathbf{X}^{\mathrm{true}}$, and the value of the true state (i.e., the physical conditions). This article addresses these issues for the retrievals performed by OCO-2, by estimating the posterior distributions using a Markov Chain Monte Carlo (MCMC) algorithm and comparing them to OE results. The application of the OE algorithm to a vector of observed radiances will hereafter be referred to as "OE retrieval" and application of MCMC to estimate posterior distributions will be called "MCMC retrievals".

Given the massive number of observations made by OCO-2, it is not possible to apply MCMC to every single retrieval so here we focus on 600 simulated soundings that represent the variability of physical conditions encountered by OCO-2 between November 2014 and January 2016. Furthermore, the full physics forward model ($\mathbf{F}$ in Eq. 1) used in the OCO-2 operational data processing is too computationally slow to be feasible to use in an MCMC algorithm where we require an evaluation of $\mathbf{F}(\mathbf{X}, \mathbf{b})$ in each iteration. However, a surrogate model exists, $\mathbf{F}^{\mathrm{surr}}$ (Hobbs et al. 2017), that is much faster but still includes the most essential physics and uses the same computational tools and tables (i.e., $\mathbf{b}$) as the operational data processing does. We perform both OE and MCMC retrievals on the selected soundings, using the surrogate forward model in both cases. We note that the analyses in this paper do not address all sources of error in a retrieval, e.g., parameter error and model discrepancy. No full exploration of the posterior distribution has been done for the OCO-2 retrieval before, for either full physics or surrogate forward model. In general, MCMC methods have rarely been applied to retrievals of atmospheric entities, exceptions include Haario et al. (2004), Wang et al. (2013), and Tukiainen et al. (2016).

When comparing and evaluating the two retrieval approaches as estimates of the true state we will view the methods as probabilistic forecasts. The OE retrieval will be treated as a *density forecast*, i.e., $N(\hat{\mathbf{x}}, \hat{S})$, whereas MCMC retrievals provide an *ensemble forecast*.

We note that even though $\hat{\mathbf{x}}$ is actually the posterior mode, $\hat{S}$ is only an approximation to the posterior covariance matrix, and the posterior is not actually Gaussian, the $N(\hat{\mathbf{x}}, \hat{S})$ distribution reflects the way the OCO-2 data product is used in practice. In our simulations, the true state is known, and we can therefore apply various forecast calibration diagnostics and proper scoring rules (Gneiting and Raftery 2007; Gneiting et al. 2008; Gneiting and Katzfuss 2014; Thorarinsdottir et al. 2016) to evaluate the two retrieval methods. Furthermore, we are concerned with the estimation of the multivariate state vector $\mathbf{X}$, the CO_2 profile vector $\mathbf{X}_p$, and the univariate quantity of interest X_{CO_2}. We therefore need comparison methods that can be applied to both density and ensemble forecasts, in both multivariate and univariate setting. For example, we utilize the continuous ranked probability score (CRPS) and its multivariate extension to compare the retrieval methods, and various rank histograms for diagnostics (see Sect. 4 for details). Viewing the OE retrieval as an approximation to the actual posterior we assess this approximation via the Kulback–Leibler divergence and asses normality of the actual posterior via probability plots of MCMC samples. Comparing the OE retrievals to the actual posteriors gives invaluable insight into whether and where the approximate OE retrievals provide satisfactory results and where they do not. For example, the OE retrieval may perform quite well in certain locations or at certain physical conditions (e.g., at certain values of aerosol optical depth, surface pressure, land surface type etc.) but fail in other locations.

We provide background on OCO-2, the forward model and Optimal Estimation in Sect. 2. The simulated data set is described in Sect. 3. A few details about the MCMC algorithm are given in Sect. 4.1 and in Sects. 4.2 and 4.3, and we describe the diagnostic tools and scoring rules we use to compare and evaluate the OE and MCMC retrievals. Results are presented in Sect. 5, and we conclude with a discussion in Sect. 6.

2. BACKGROUND

NASA's Orbiting Carbon Observatory-2 (OCO-2) is now in the process of collecting space-based measurements of atmospheric carbon dioxide, CO_2. The scientific basis of the OCO mission is described, e.g., by Crisp et al. (2004), Crisp et al. (2007), Crisp and Johnson (2005) and Crisp et al. (2014), and recent update is given in Eldering et al. (2017). The data product is publicly available, e.g., at `co2.jpl.nasa.gov`. Here we briefly describe the OCO-2 instrument (Sect. 2.1), the physical forward models (Sects. 2.2, 2.3), and the OE algorithm (Sect. 2.4).

2.1. Remote Sensing of CO_2

The OCO-2 instrument collects eight observations every 0.333 seconds with a surface footprint of less than 2.25 km down-track and between 0.1 and 1.3 km cross-track. The satellite flies in a polar, sun-synchronous orbit, providing global coverage with a 233-orbit 16-day repeat cycle. The instrument has three observation modes: nadir (preferred over land), glint (preferred over ocean) and target mode.

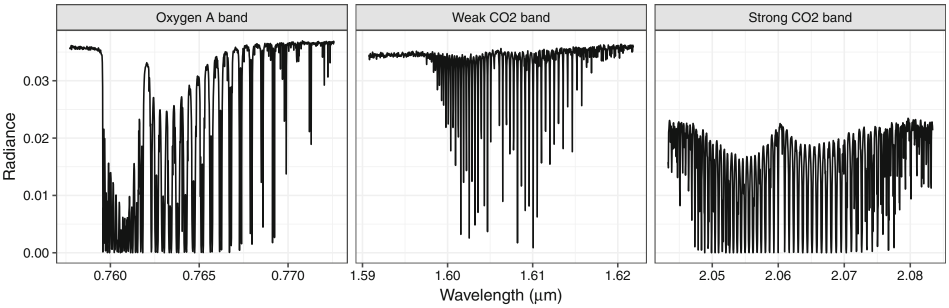

Figure 1. Example of an observed sounding $\mathbf{Y}$ over Madagascar in October 2015.

The instrument incorporates three high-resolution imaging grating spectrometers that make coincident measurements in three spectral bands. These bands are called O_2 A, weak CO_2, and strong CO_2 bands, and are centered around 0.765, 1.61, and 2.06 μm respectively. Each observation consists of 1016 radiances from each band and the 3048-dimensional observed vector $\mathbf{Y}$ is called a *sounding*. An example of an observed sounding is shown in Fig. 1. The sharp drops in radiances correspond to how much light is absorbed by molecules in the atmosphere at that wavelength. A physical forward model describes this absorption for a given amount of relevant molecules, represented by the state vector $\mathbf{X}$.

2.2. Full Physics Forward Model

The OCO-2 science team has developed a physical forward function for use in OCO-2 retrievals, referred to as the *Full Physics* forward model. This is a state-of-the-art physical model with minor simplifications to ease computation. For given inputs, the forward model $\mathbf{F}(\mathbf{X}, \mathbf{b})$ simulates solar spectra, a 3048-dimensional vector $\mathbf{Y}$, and radiance Jacobians. This involves numerically solving the radiative transfer equation, an ordinary integro-differential equation, for each sounding [more details are given in O'Dell et al. (2012)].

The input vector $\mathbf{b}$ is treated as known and includes tens of thousands of constants. Many elements of $\mathbf{b}$, such as gas absorption coefficients, are estimated from laboratory experiments. The OCO-2 retrieval algorithm uses precalculated lookup tables of absorption coefficients (ABSCO) for the calculation of gas absorption cross sections. The ABSCO tables contain molecular absorption cross sections over the range of relevant wavelengths, temperatures and pressures (Thompson et al. 2012). Other elements of $\mathbf{b}$, such as aerosol properties, are estimated from other remote sensing data as well as airborne/field measurement campaigns. Details about the nature and origins of various elements of the $\mathbf{b}$ vector are given in Crisp et al. (2014).

The state vector, $\mathbf{X}$, has about 60 elements, 20 of which are CO_2 concentration in 20 layers in the atmospheric column. The remaining elements include surface pressure, albedo, 4 different species of aerosols, weather variables such as temperature and wind speed, a water vapor multiplier, instrument wavelength shifts, and residual empirical orthogonal function (EOF) amplitudes. Prior means and variances for the state vector are determined using other measurement networks, such as the European Centre for Medium-Range Weather Forecasts

(ECMWF) daily forecasts of pressure, temperature profile, water vapor profile, and wind speed, the GLOBALVIEW dataset (GLOBALVIEW-CO2 2013) for the CO_2 profile and MERRA climatology (Rienecker et al. 2011) for aerosols.

2.3. SURROGATE FORWARD MODEL

The surrogate forward model, $\mathbf{F}^{\text{surr}}(\mathbf{X}, \mathbf{b})$, used in this paper is described in detail by Hobbs et al. (2017), we give a brief description here. The surrogate model assumes a 39-dimensional state vector, $\mathbf{X}$, which includes the 20 layers of CO_2 concentration, surface pressure, coefficients for 4 different species of aerosols, and albedo for the three spectral bands. It does not include the full physics state vector elements that account for temperature and humidity adjustments, wavelength offsets, and solar-induced fluorescence. These components are fixed, where necessary in the surrogate model. The surrogate model uses a similar wavelength resolution in the three bands to yield a 3048-dimensional radiance vector $\mathbf{Y}$.

As for the full physics model, the surrogate forward function $\mathbf{F}^{\text{surr}}$ is a numerical solution to the radiative transfer equation, which is an ordinary integro-differential equation defined along the vertical path through the atmosphere. In general this solution is nonlinear in the state $\mathbf{X}$. Other simplifications are made for computational efficiency, including additional numerical approximations in solving the radiative transfer equation, and fixed ABSCO tables for a given location and time which results in a smaller $\mathbf{b}$ vector. This yields a speed improvement on the order of 20 times over the full physics forward model.

2.4. OPTIMAL ESTIMATION (OE) RETRIEVAL

Details of the retrieval algorithm used by OCO-2 are given, e.g., by O'Dell et al. (2012), Crisp et al. (2012), and Crisp et al. (2014). Briefly, the retrieved state vector $\hat{\mathbf{x}}$ is the vector that minimizes the following cost function:

$$c = (\mathbf{y} - \mathbf{F}(\mathbf{x}, \mathbf{b}))^\top \Sigma_\varepsilon^{-1} (\mathbf{y} - \mathbf{F}(\mathbf{x}, \mathbf{b})) + (\mathbf{x} - \boldsymbol{\mu}_a)^\top \Sigma_a^{-1} (\mathbf{x} - \boldsymbol{\mu}_a) \tag{4}$$

where $\mathbf{y}$ is the observed sounding (3048-dim vector), $\mathbf{F}$ is the full physics forward model, $\mathbf{b}$ is a constant vector and Σ_ε is a diagonal matrix of measurement error variances, which also are assumed known. The vector $\boldsymbol{\mu}_a$ is the prior mean for the state vector and Σ_a is the prior covariance matrix. The minimization is performed with the Levenberg–Marquardt algorithm. Uncertainty estimates are given by the covariance matrix:

$$\hat{S} = (K^\top \Sigma_\varepsilon^{-1} K + \Sigma_a^{-1})^{-1} \tag{5}$$

where K is the linearization of $\mathbf{F}(\mathbf{x}, \mathbf{b})$, evaluated at the retrieved value:

$$K = \left. \frac{\delta \mathbf{F}(\mathbf{x}, \mathbf{b})}{\delta \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}} . \tag{6}$$

The main data product of interest is the column-averaged CO_2 dry air mole fraction, $\hat{X}_{CO_2}$, and its variance, $\hat{s}^2_{X_{CO_2}}$. This is a weighted average of the 20 CO_2 concentrations, which depends in part on elements of the retrieved state vector, i.e.,

$$\hat{X}_{CO_2} = \mathbf{h}^\top \hat{\mathbf{x}}_p \qquad \text{and} \qquad \hat{s}^2_{X_{CO_2}} = \mathbf{h}^\top \hat{S}_p \mathbf{h} \tag{7}$$

where $\hat{\mathbf{x}}_p = (\hat{x}_1, \ldots, \hat{x}_{20})^T$ is the retrieved CO_2 profile and $\hat{S}_p$ is the corresponding 20×20-dimensional retrieved covariance matrix. The pressure weighting function $\mathbf{h}$ represents the proportion of the mass of the full column of dry air represented by each vertical level in the CO_2 profile portion of the state vector. For OCO-2 these weights are nearly uniform, with a slight adjustment due to water vapor (Crisp et al. 2014).

The OE retrievals performed in this paper use surrogate forward model $\mathbf{F}^{\text{surr}}$ instead of $\mathbf{F}$. For prior distributions we use the same mean vectors and covariance matrices ($\boldsymbol{\mu}_a$ and Σ_a) used in the OCO-2 operational processing for given time and location. In the surrogate model, the water vapor is fixed, so the pressure weight vector $\mathbf{h}$ is constant from sounding to sounding. As in the operational retrieval, the measurement error covariance matrix Σ_ϵ is diagonal with variances proportional to the mean radiance, setting the signal-to-noise ratio comparable to that for the actual OCO-2 instrument (see also Hobbs et al. (2017)).

The approach described above is called *optimal estimation* (Rodgers 2000) in remote sensing circles, but is of course an application of a nonlinear normal model with a normal prior on the parameters:

$$\begin{aligned} \mathbf{Y} \mid \mathbf{X} &\sim N\left(\mathbf{F}(\mathbf{X}, \mathbf{b}),\ \Sigma_\varepsilon\right) \\ \mathbf{X} &\sim N\left(\boldsymbol{\mu}_a, \Sigma_a\right)\ . \end{aligned} \tag{8}$$

Therefore, $\hat{\mathbf{x}}$ is an estimate of the posterior mode (although often treated as the posterior mean) and $\hat{S}$ is an estimate of the posterior variance-covariance matrix.

We note that if $\mathbf{F}$ is a linear function of $\mathbf{X}$, then the posterior $[\mathbf{X} \mid \mathbf{Y}]$ is a multivariate normal distribution with mean $\hat{\mathbf{x}}$ and covariance matrix $\hat{S}$. But since $\mathbf{F}$ is not a linear function, the posterior is not normal. Therefore, the posterior mode $\hat{\mathbf{x}}$ is not necessarily the same as the posterior mean. Note also that OE retrieval estimates the posterior mode of the multivariate distribution $[\mathbf{X} \mid \mathbf{Y}]$ while the interest is mainly in the marginal posterior of a transformation of the first 20 elements $\left[X_{CO_2} \mid \mathbf{Y}\right]$. Even though $\hat{\mathbf{x}}$ is the mode of $[\mathbf{X} \mid \mathbf{Y}]$ the estimate of $\hat{X}_{CO_2}$ in Eq. (7) is not necessarily the mode of $\left[X_{CO_2} \mid \mathbf{Y}\right]$.

3. DATA USED IN THIS STUDY

To assess the OE approximation to the actual posterior in a meaningful way we need observations $\mathbf{Y}$ and state vectors $\mathbf{X}$ that are representative of the actual observations and atmospheric conditions the OCO-2 satellite encounters. For this and related mission simulation activities, the first 18 months of the OCO-2 Level 2 retrieved state vectors were analyzed with the t distributed stochastic neighbor embedding (tSNE) dimension-reduction procedure (Van Der Maaten and Hinton 2008). A cluster analysis was performed on the results,

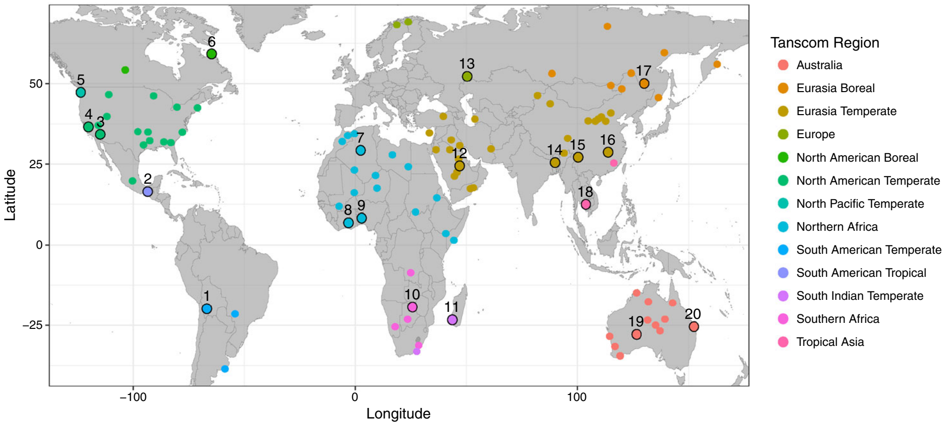

Figure 2. The 100 available data templates colored by the Transcom Region. The 20 templates selected for paper are circled black (Color figure online).

and a collection of 100 cluster centers and singleton outliers were identified as "template soundings". The analysis was performed separately on land nadir, land glint, and ocean glint data. The present study uses land nadir templates only.

By design, the collection of template soundings span the range of atmospheric CO_2 and aerosol concentrations estimated by OCO-2. In the process, the templates additionally span across seasons and geographical regions. For our MCMC experiments, we selected 20 land nadir templates that include all of the thirteen Transcom regions (Gurney et al. 2003) and cover a large range of total aerosol depth (AOD) values. The locations of the selected templates are shown in Fig. 2 and information about dates, aerosol species and total AOD are given in Table 1.

Within a template, we wish to study a set of plausible state vectors $\mathbf{X}$ that are randomly sampled from a common marginal distribution. The marginal distribution is estimated using a Gaussian mixture model, with data assembled from OCO-2 CO_2 retrievals and MERRA aerosol data within 300 km and in the same meteorological season (i.e., within approximately 2 months) as the template sounding. For each template, we simulate 30 random state vectors from this mixture model and simulate a synthetic radiance $\mathbf{Y}$ from the surrogate model, with measurement error, to pair with each $\mathbf{X}$. In total, we have 600 pairs of synthetic state vectors and soundings. The optimal estimation algorithm was applied to these synthetic soundings, using the surrogate forward model and the same priors as the mission uses for the template location and time. We note that the distribution from which the $\mathbf{X}$ are simulated is generally not the same as the prior distribution used in the retrieval.

4. METHODS

4.1. MCMC METHODS

We obtained samples from the posterior $[\mathbf{X} \mid \mathbf{Y}]$ in model (8), referred to here as the "MCMC retrieval", using the adaptive Metropolis algorithm of Haario et al. (2001). For each

Table 1. Information about the 20 selected data templates used in this paper.

	Transcom Region	Month/year	Aerosols	AOD
1	South American Temperate	1/2015	DU-SO	0.126
2	South American Tropical	2/2015	SO-SS	0.035
3	North American Temperate	7/2015	DU-SO	0.175
4	North American Temperate	6/2015	DU-SO	0.062
5	North Pacific Temperate	11/2014	SO-SS	0.040
6	North American Boreal	6/2015	DU-OC	0.042
7	Northern Africa	7/2015	DU-SO	0.182
8	Northern Africa	2/2016	DU-OC	0.285
9	Northern Africa	5/2015	DU-OC	0.249
10	Southern Africa	10/2015	BC-OC	0.039
11	South Indian Temperate	1/2016	SO-SS	0.202
12	Eurasia Temperate	12/2014	DU-SO	0.149
13	Europe	9/2015	DU-SO	0.047
14	Eurasia Temperate	1/2016	OC-SO	0.567
15	Eurasia Temperate	2/2015	DU-OC	0.021
16	Eurasia Temperate	9/2015	BC-SO	0.534
17	Eurasia Boreal	8/2015	DU-SO	0.094
18	Tropical Asia	11/2015	SO-SS	0.155
19	Australia	6/2015	DU-SS	0.044
20	Australia	12/2014	DU-SS	0.050

Locations span the whole globe, dates range from November 2014 to February 2016 and the total aerosol optical depth (AOD) ranges from 0.021 to 0.567. The aerosol species are: Black Carbon (BC), Dust (DU), Organic Carbon (OC), Sulfate (SO), and Sea Salt (SS).

of the 600 soundings we ran four independent chains, starting at different initial values, with 250,000 iterations per chain. The computation for each chain took about 24 h, but we ran many independent chains in parallel on a JPL computing cluster, reducing the overall time used for computations. Of the 2400 chains, 18 were terminated due to unsuccessful Cholesky factorization and 9 were terminated due to numerical issues in the forward model. For each chain, the first 100,000 iterations were set as burn-in and acceptance rate after burn-in ranged from 0.5 to 14.5% with a median of 3.72%. Furthermore, the chains were thinned by retaining every 100th MCMC sample leaving 1500 MCMC samples for inference, or 6000 in total if all four chains ran to completion. We performed convergence diagnostics both via visual inspection of traceplots of all 39 state vector elements and using the Gelman–Rubin convergence statistic (Gelman and Rubin 1992; Brooks and Gelman 1998). Of the 600 MCMC retrievals, 457 were deemed converged.

4.2. Assessing Retrieval Methods

In our assessments and comparisons of OE and MCMC retrievals, we do not make a distinction between the statistical concepts of estimation and prediction. The CO_2 retrieval is technically a (probabilistic) parameter estimation, and in practice the data product is usually viewed as a measurement with accompanying measurement error. The estimation of the atmospheric state comes in the form of posterior samples (MCMC retrieval) and multivariate normal posterior distributions (OE retrieval), which are analogous to an ensemble forecast and a density forecast. The true states of the atmosphere (known within the realm of our simulations) are treated here as the observations the MCMC and OE retrievals are supposed to predict. We can therefore utilize various probabilistic forecast assessment methods, many of which are frequently used in assessing weather forecasts (see, e.g., Gneiting and Katzfuss 2014, for a recent overview). In fact, the goals of our retrieval methods are the same as in forecasting: sharp distributions (low variance) and calibrated prediction, i.e., distributions that are statistically compatible to the true values.

In univariate settings (i.e., estimation of X_{CO_2}), we use diagnostic tools commonly used in forecast evaluation (Gneiting et al. 2007), probability integral transform (PIT) histogram for the OE retrieval and verification rank histogram for the MCMC retrieval to assess calibration. A PIT histogram (Dawid 1984; Diebold et al. 1998; Gneiting and Raftery 2007) is simply a histogram of the cumulative distribution functions (cdf) evaluated at the true values, which are uniformly distributed if the true values are realization of the posterior distributions. A verification rank histogram (Anderson 1996; Hamill and Colucci 1997) or Talagrand diagram (Talagrand et al. 1997) shows the ranks (or sample percentile rank) of the true values in the posterior samples. These two histograms are equivalent and therefore comparable for the OE and MCMC retrievals (Gneiting et al. 2007).

Diagnostic tools for multivariate ensemble and density forecasts are an area of active research. For ensemble forecasts we consider two approaches. Firstly, the average rank histogram, where ensembles are ordered according to average univariate ranks (Thorarinsdottir et al. 2016), and secondly, the band depth rank histogram where the concept of band dept is used for ordering (López-Pintado and Romo 2009; Sun and Genton 2011, 2012; Sun et al. 2012). For density forecasts (OE), we apply the Box density ordinate transform (BOT) proposed by Box (1980), O'Hagan (2003), and Gneiting et al. (2008), which in our case is simply a histogram of the percentile ranks of the true values:

$$u = 1 - \chi^2_{39}\left((\mathbf{x}^{\text{true}} - \hat{\mathbf{x}})^T \hat{S}^{-1} (\mathbf{x}^{\text{true}} - \hat{\mathbf{x}}) \right)$$

where $\hat{\mathbf{x}}$ and $\hat{S}$ are the OE estimates of the posterior mode and posterior covariance matrix, and $\chi^2_{39}(\cdot)$ is the cdf of a chi-square distribution with 39 degrees of freedom. For all three diagnostic tools, the histograms show a uniform distribution if observations are a random sample from the posteriors. Interpretation of average rank histograms is equivalent to the standard univariate rank histograms, outlying observations can give either high or low rank values. However, both band depth and BOTs are center-outward orderings so outlying values tend to lead to low-rank and in-lying observations tend to yield high values.

We apply proper scoring rules to compare the skill of OE and MCMC to recover the true values that take into account both sharpness and calibration (Gneiting and Raftery 2007; Gneiting et al. 2007). In the univariate setting we consider both absolute error and the continuous ranked probability score (CRPS) (Matheson and Winkler 1976; Gneiting and Raftery 2007; Gneiting and Katzfuss 2014) as well as the squared prediction error. The absolute and squared errors are simply AE $= |\hat{x} - x|$ and SPE $= (\hat{x} - x)^2$ where x is the true value and $\hat{x}$ is a point estimate, either OE estimate of X_{CO_2} or the MCMC estimate, defined here as the posterior median. In other words, only point estimates are included. The CRPS takes the whole predictive distribution in to account and is defined as

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} (F(u) - 1(x \leq u))^2 \, \mathrm{d}u = E_F|U - x| - \frac{1}{2} E_F|U - U'| \tag{9}$$

where x is the observed value and U and U' are independent random variables with cdf F. The CRPS is available in closed form for normal predictive distributions and can also be estimated from samples using estimates of the empirical cdf (Hersbach 2000). In multivariate setting there are not many scoring rules that can be applied to both density and ensemble forecasts. Here we consider the energy score, a multivariate extension of CRPS (Gneiting et al. 2008):

$$ES = E_F||\mathbf{U} - \mathbf{x}|| - 0.5 E_F||\mathbf{U} - \mathbf{U}'|| \tag{10}$$

where $\mathbf{x}$ is the true state vector and $\mathbf{Y}$ and $\mathbf{Y}'$ are independent samples from F. We estimate the energy score via Monte Carlo simulation for both ensemble and density forecasts.

4.3. KULLBACK–LEIBLER DIVERGENCE

The Kullback–Leibler divergence, though not mathematically a distance measure, can be used to compare two distributions and is defined as

$$D(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \tag{11}$$

in the univariate case. In our setting the distributions P and Q are the actual posterior and the OE approximation, respectively. Since Q is a normal distribution we can evaluate the density $q(x)$ at any x, but we only have samples from P. To estimate the divergence in (11), we obtain a kernel density estimate of $p(x)$ and calculate

$$\hat{D}(P||Q) = \frac{1}{M} \sum_{m=1}^{M} \log \frac{\hat{p}(x^{(m)})}{q(x^{(m)})} \tag{12}$$

where $\hat{p}$ is the estimated density and $\{x^{(m)}\}_{m=1}^{M}$ are MCMC samples from P.

5. RESULTS

We compare the results of the computationally efficient, but approximate, OE retrievals to results of MCMC retrievals. The MCMC retrievals provide samples from the posterior and we treat OE retrievals as implying a Gaussian posterior distribution with means and covariances given by the OE estimates. We apply methods laid out in Sects. 4.2 and 4.3 and will focus on three viewpoints, the main quantity of interest X_{CO_2} (Sect. 5.1), as well as the CO$_2$ profile $\mathbf{X}_p = (X_1, X_2, \ldots, X_{20})$ and the full state vector $\mathbf{X}$ (Sect. 5.2).

5.1. UNIVARIATE ASSESSMENTS AND COMPARISONS FOR X_{CO_2}

We begin by examining the estimated posterior distributions of X_{CO_2} for seven out of the 457 successful retrieval, see Fig. 3. These seven were chosen to show examples of

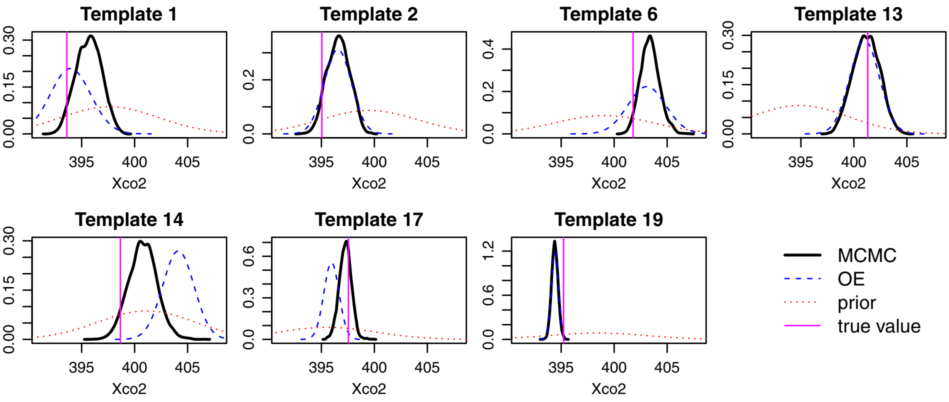

Figure 3. Posterior densities for one sample from seven templates as obtained by MCMC (black, solid) and Optimal Estimation (blue, dashed). The vertical lines (magenta, solid) show the true value of X_{CO_2} (Color figure online).

the variety in results and a few things are worth noting. First, the posterior distributions estimated with MCMC are not severely non-Gaussian. In fact, normal probability plots (not shown here) indicated substantial deviation from normality in only about 10 of the 457 successful retrievals. A normal approximation to the posterior of X_{CO_2} is therefore not unreasonable. Second, in some cases the MCMC and OE retrievals give essentially the same posterior (templates 2, 13, and 19) while in other cases they are quite different. Sometimes the MCMC estimated posterior is much sharper than the OE (templates 1 and 6), that is the posterior variance is smaller than what is implied by OE. This may seem surprising since the uncertainty in the product is commonly believed to be underestimated rather than overestimated, but keep in mind that this comparison is only between two computational methods and there are many sources of uncertainty (unknown parameters, model error, etc.) that neither the MCMC nor OE retrieval used here take into account. Third, sometimes the distributions given by MCMC and OE retrievals are centered at very different values (templates 1, 14, and 17). It may also seem surprising that the posterior modes estimated with these two methods can be so different, but keep in mind that OE finds the posterior mode of the joint posterior of the whole state vector $[\mathbf{X} \mid \mathbf{Y}]$ which after transformation to X_{CO_2} may or may not correspond to the posterior mode of the univariate posterior $[X_{\mathrm{CO}_2} \mid \mathbf{Y}]$, which is the distribution we obtain with MCMC methods. Furthermore, the posterior mode can be missed by the OE retrieval if the Levenberg–Marquardt algorithm did not converge properly, e.g., found a local minimum. Fourth, comparing the posterior distributions given by MCMC and OE retrievals to true values for these seven examples only indicates that most of the time both MCMC and OE retrievals recover the true values. Sometimes the MCMC retrieval covers the true value better (e.g., template 14) and sometimes the OE retrieval does (e.g., template 1).

We now turn to an overall assessment of the two methods using all 457 successful retrievals. The PIT histogram (OE retrieval) and verification rank histogram (MCMC retrieval) are shown in Fig. 4. The U-shaped histograms indicate that both methods are under-dispersed, i.e., true values are in the far tails of the posterior in more cases than can be explained by chance particularly the left tail. The OE retrieval is perhaps slightly less

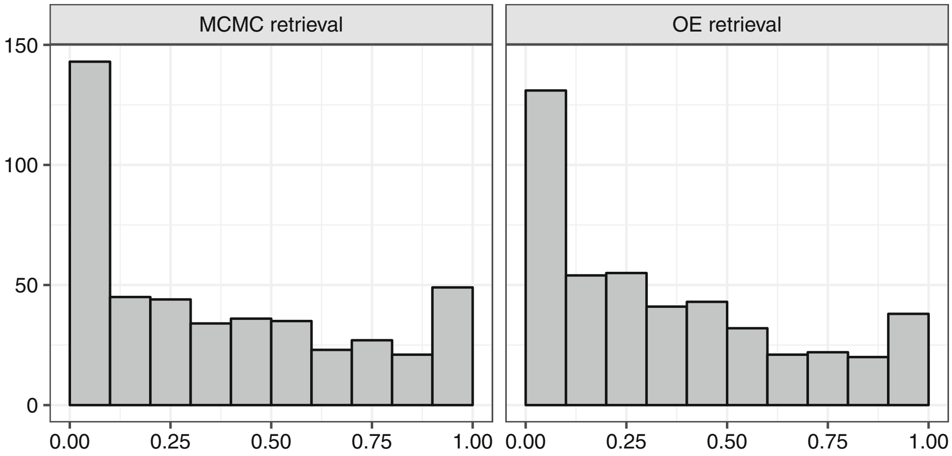

Figure 4. Verification rank histogram for the MCMC retrievals (left) and PIT histogram for OE retrievals (right).

Table 2. Average scores for column-averaged CO$_2$ concentration X_{CO_2}, CO$_2$ profile vector $\mathbf{X}_p$, and full state vector $\mathbf{X}$.

	X_{CO_2}, CRPS	X_{CO_2}, AE	X_{CO_2}, MSPE	$\mathbf{X}_p$, ES	$\mathbf{X}$, ES
OE	1.25	1.63	6.35	14.54	14.85
MCMC	0.97	1.28	4.09	13.29	13.58
Perc. MCMC $<$ OC	67.0%	62.1%	62.1%	60.8%	62.6%

The scores are the average continuous ranked probability score (CRPS), average absolute error (AE), mean square prediction error (MSPE), and the multivariate energy score (ES). In all cases, a lower score indicates a better prediction.

under-dispersed, which can be explained by the typically larger OE posterior variances. In fact, the standard deviations estimated by the OE are larger than those estimated by MCMC retrieval in 440 of our 457 cases and the average difference is 0.289 ppm (average ratio was about 1.34).

For a quantitative comparison we calculate the absolute error, squared prediction error, and CRPS using the R package scoringRules (Jordan et al. 2017). Average scores are shown in Table 2, note that a lower score indicates a better prediction. We see that even though OE retrieval seems slightly better calibrated (see Fig. 4), MCMC retrieval generally performs better both as a point estimate (absolute/squared error) and as a probabilistic estimate (CRPS).

We now give two examples of insights this and similar studies can give for the OCO-2 mission. First, Fig. 5 shows the CRPS and absolute error scores plotted against the aerosol species of the state vector. The state vector includes information about four aerosol species, two of which are fixed (water vapor and ice water) and two are chosen in the data processing operation. The possible aerosol species are Black Carbon (BC), Dust (DU), Organic Carbon (OC), Sulfate (SO), and Sea Salt (SS). The effect of which aerosol species are selected is that the parameter vector **b** changes, as the aerosol species have different absorption properties. For our dataset, the pair of flexible aerosol species are the same within template,

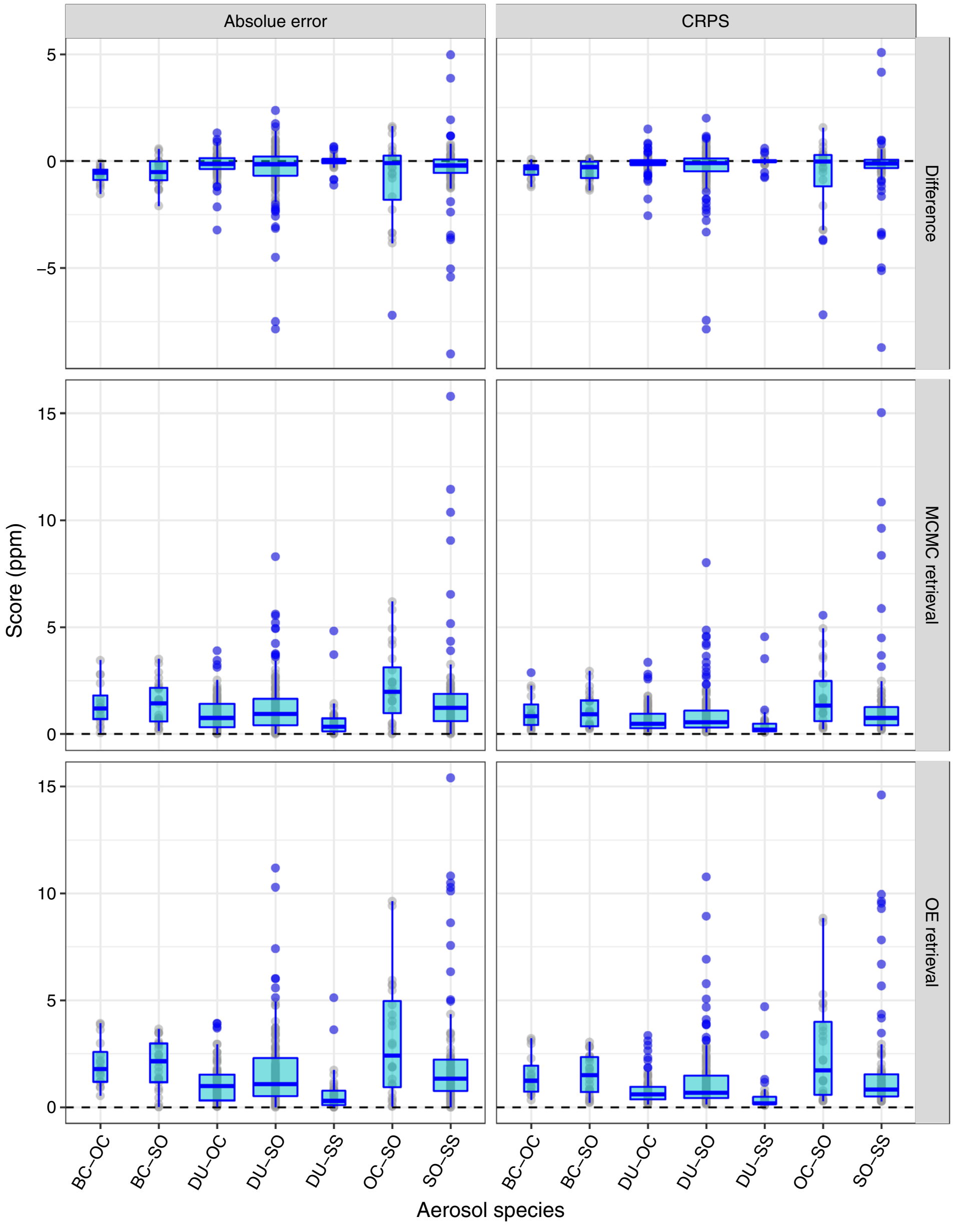

Figure 5. Absolute error (left) and CRPS (right) for the 457 successful retrievals of X_{CO_2}, plotted against the pair of aerosol species. The first row shows the differences in scores between MCMC and OE retrievals, the second and third row show the scores for MCMC and OE retrievals, respectively. The species are: Black Carbon (BC), Dust (DU), Organic Carbon (OC), Sulfate (SO), and Sea Salt (SS).

e.g., in template 1 we have dust and sulfate, DU/SO (Table 1). We see from Fig. 5 that the quality of the prediction varies between aerosol types (recall that a lower score indicates a better prediction). For example, the median scores for both MCMC and OE retrievals for templates with aerosol species DU/SS and DU/OC are lower than for others. Also, the scores

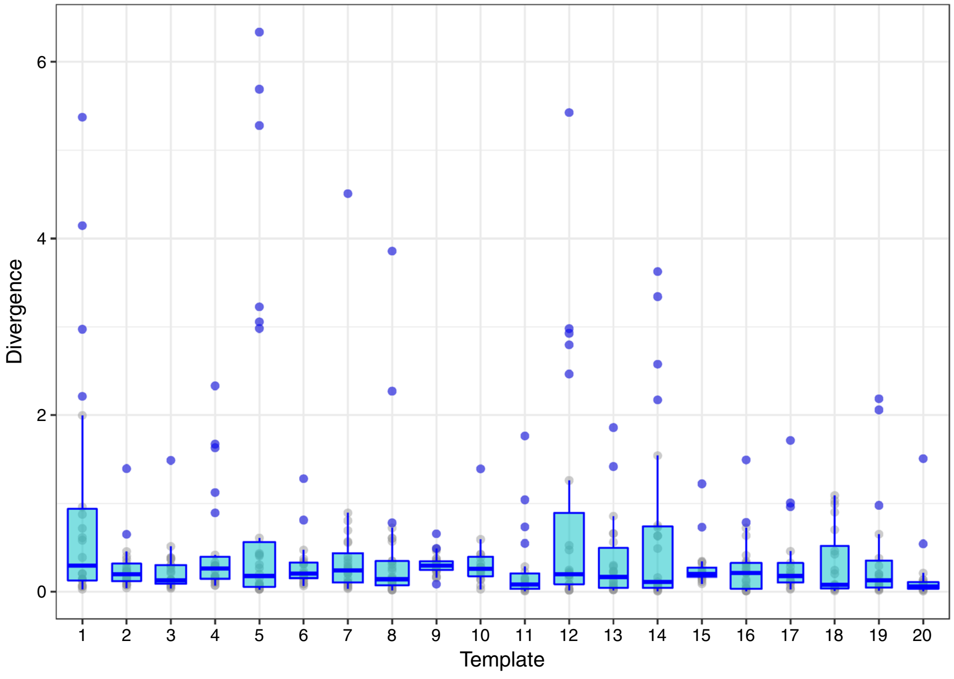

Figure 6. Kullback–Leibler divergence between MCMC retrievals and OE retrievals for the 20 selected templates. Outliers not shown in the graph, (template, divergence): (1, 15.0), (5, 10.5), (5, 17.6), (14, 13.0), and (17, 53.9).

for templates with species BC/OC and BC/SO are less variable than for any others. Largest scores occur for templates with species DU/SO and SO/SS. This indicates that both OE and MCMC retrievals tend to perform worse when the state vector includes sulfate paired with dust, organic carbon or sea salt than for other aerosol combinations. The difference in scores between retrieval methods (Fig. 5, top row) shows that MCMC performs better than OE for almost all soundings in templates with aerosol species BC/OC and BC/SO. In all cases the median difference is negative, i.e., MCMC performs better for all combinations of aerosol species. Overall, MCMC performs better in 67 and 62.1% of cases in terms of CRPS and absolute errors respectively. This indicates that while the OE retrieval tends to perform worse than the MCMC retrieval in terms of predicting the true value in general, the difference in performance is most persistent when the aerosol species include Black Carbon.

Second, Fig. 6 shows the estimated Kullback–Leibler divergences between posterior estimated via MCMC and the OE approximation $N(\hat{x}_{CO_2}, \hat{s}^2_{X_{CO_2}})$ for each template (location). The Fig. 6 reveals that in terms of Kullback–Leibler divergence, the OE approximation tends to be closer to the actual posterior for some templates (e.g., templates 3 and 11) than others (e.g., templates 1 and 5).

5.2. MULTIVARIATE COMPARISONS

We now turn our attention to the retrievals of the 20-dimensional CO_2 profile vector $\mathbf{X}_p$ and the 39-dimensional full state vector $\mathbf{X}$. Our first question is whether the posterior $[\mathbf{X} \mid \mathbf{Y}]$

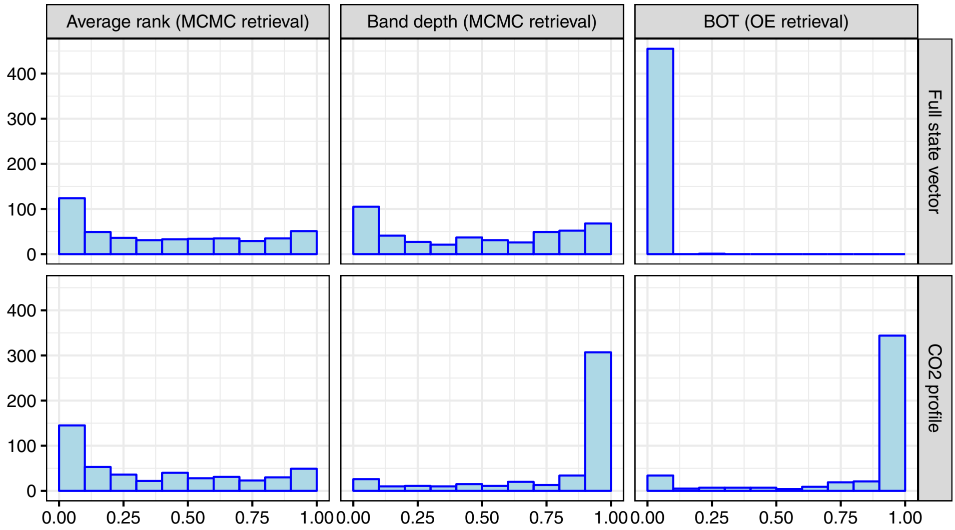

Figure 7. Forecasting diagnostics for the full state vector $\mathbf{X}$ (top row) and the CO_2 profile vector $\mathbf{X}_p$ (bottom row). Left: average rank histograms (MCMC retrieval), middle: band depth rank histograms (MCMC retrieval), and right: BOT histograms (OE retrieval).

or the marginal posterior $\left[\mathbf{X}_p \mid \mathbf{Y}\right]$ are (close to) multivariate normal. Chi-square probability plots (not shown here) revealed that for almost all soundings, $\left[\mathbf{X} \mid \mathbf{Y}\right]$ and $\left[\mathbf{X}_p \mid \mathbf{Y}\right]$ are *not* multivariate normally distributed, although plots for $\left[\mathbf{X}_p \mid \mathbf{Y}\right]$ tended to show less severe inconsistencies with normality.

Turning to multivariate prediction diagnostics, Fig. 7 shows the average rank and band depth rank-histograms for the MCMC retrieval and the BOT histogram for the OE retrieval. From the BOT histogram we see that the OE retrieval is very poorly calibrated in the multivariate case, but in different ways for $\mathbf{X}_p$ and $\mathbf{X}$. The low values for the whole state vector points to too many outlying values, i.e., the posterior $N_{39}(\hat{\mathbf{x}}, \hat{S})$ distributions tend to not cover the true value. The high values for the profile vector (Fig. 7, bottom right) may indicate the opposite is true there, $N_{20}(\hat{\mathbf{x}}_p, \hat{S}_p)$ is over dispersed in that dimension, although bias and miss-specification of correlations cannot be ruled out as culprits. The average rank and band depth histograms (Fig. 7, left and middle) show calibration diagnostics for MCMC retrieval. Neither reveal severe calibration issues for the full state vector, both indicate only slightly too many outlying values for $\mathbf{X}$. However, the band depth histogram shows too many in-lying or biased values for $\mathbf{X}_p$. The average rank histograms are similar for $\mathbf{X}$ and $\mathbf{X}_p$ and both indicate slightly too many outlying values.

Average energy scores are shown in Table 2. As in the univariate case the MCMC method performs better both for $\mathbf{X}_p$ and $\mathbf{X}$. MCMC gave a lower (i.e., better) energy score in 60.8 and 62.6% of soundings for CO_2 profile and full state vector respectively. Plotting energy scores against aerosol species reveals a similar story as Fig. 5, i.e., the difference in energy scores between the OE and MCMC retrievals vary across aerosol species.

6. DISCUSSION

We have performed extensive comparisons of the OCO-2 operational data production procedure, OE retrieval, and the actual posterior obtained with MCMC methods. The general indication from our results is that when compared to MCMC, the OE retrieval performs reasonably well for the main quantity of interest X_{CO_2} but not for the full state vector $\mathbf{X}$.

We found that, for the cases considered in this paper, a normal approximation to $\left[X_{\mathrm{CO}_2} \mid \mathbf{Y}\right]$ is not unreasonable. Even though the posterior of the CO_2 profile $\mathbf{X}$ is generally not multivariate normal, the averaging in calculating X_{CO_2} helps making its posterior reasonably close to normal. Prediction performance of the two retrieval methods is similar, the OE retrievals are slightly better calibrated but the MCMC retrievals generally have better forecasting scores, both absolute errors and CRPS.

The source of the difference in OE and MCMC retrievals of X_{CO_2} can be traced to at least two issues with the OE retrieval. First, the estimate $\hat{X}_{\mathrm{CO}_2}$ is not necessarily the mode of the posterior $\left[X_{\mathrm{CO}_2} \mid \mathbf{Y}\right]$ so treating it as the mean of the posterior may be a source of bias. In fact, the posterior median (from the MCMC retrieval) was on average closer to the true value than $\hat{X}_{\mathrm{CO}_2}$ (absolute error was on average smaller). Second, the estimated variances $\hat{s}^2_{\mathrm{CO}_2}$ are almost always larger than the variances derived from MCMC, and variances that are too large will increase the CRPS.

The OE retrieval almost always overestimated the posterior variances of X_{CO_2}, on average the OE standard deviation was about 34% larger than the actual posterior standard deviation. It is generally acknowledged that the variance in the OCO-2 data product is too small, which is mostly due to the omission of many sources of uncertainty, e.g., in parameters $\mathbf{b}$ and model discrepancy. Neither of these sources are present her; however, both the parameter values and forward models are the same for both generating soundings and for the retrievals. The difference in variances detected here is due to the way the OE retrieval estimates the posterior variance. The forward model is linearized around the estimated state vector $\hat{\mathbf{x}}$ (Eqs. 5, 6) and then the posterior variance of X_{CO_2} is calculated as if the posterior of the CO_2 profile is multivariate normal (Eq. 7). In our case this results in over estimating the posterior variance.

We found that for the multivariate cases the posteriors $\left[\mathbf{X}_p \mid \mathbf{Y}\right]$ and $\left[\mathbf{X} \mid \mathbf{Y}\right]$ are not well represented with a normal distribution. This has potential ramifications for users of the data product, e.g., when assimilating the whole CO_2 profile $\mathbf{X}_p$ in carbon flux inversions. The OE retrieval fails as a probabilistic estimator of the true state $\mathbf{X}$, as indicated by the BOT histogram in Fig. 7). As a probabilistic estimator of $\mathbf{X}_p$ the OE is too conservative (too large variances) but so is the MCMC retrieval.

This article addresses one of many sources of uncertainty in the OCO-2 data product, namely the computational approximations made to obtain a posterior distribution for X_{CO_2}. As in the operational retrieval, we treat many parameters as fixed, e.g., the $\mathbf{b}$ vector and the measurement error variances in Σ_ϵ. These parameters are too numerous to estimate (via MCMC or OE) but a subset could be chosen based on expert knowledge about which ones the forward model would be most sensitive to, and is a subject of our future work. An overall quantification of sources of uncertainty for OCO-2 is a work in progress and an active research area (Connor et al. 2016; Cressie et al. 2016; Hobbs et al. 2017).

MCMC methods are currently too computationally expensive to be applied routinely in the operational data processing for OCO-2. Future work will include alternative sampling or approximation methods to obtain a more accurate representation of the posterior of the quantity of interest. On the other hand, it is possible to apply MCMC methods on a limited number of strategically selected soundings in the operational data processing. A comparison between OE retrievals and the MCMC retrievals similar to those performed in this paper could then help identify locations or physical conditions where the OE retrieval falls short of representing the actual posterior and an MCMC retrieval would be beneficial.

ACKNOWLEDGEMENTS

The research described in this paper was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. Support was provided by the Orbiting Carbon Observatory-2 (OCO-2) mission. The authors thank Dr. Michael Gunson, Dr. Annmarie Eldering, and Dr. Noel Cressie for insightful discussions, support, and advice during the development of this work. Furthermore, we thank the Editor, Associate Editor and an anonymous reviewer for very helpful comments.

[Received September 2017. Accepted February 2018.]

REFERENCES

Anderson, J. L. (1996), “A method for producing and evaluating probabilistic forecasts from ensemble model integrations,” *Journal of Climate*, 9(7), 1518–1530.

Battle, M., Bender, M. L., Tans, P. P., White, J. W. C., Ellis, J. T., Conway, T., and Francey, R. J. (2000), “Global Carbon Sinks and Their Variability Inferred from Atmospheric O_2 and $\delta^{13}C$,” *Science*, 287(5462), 2467–2470.

Bösch, H., Baker, D., Connor, B., Crisp, D., and Miller, C. (2011), “Global characterization of CO_2 column retrievals from shortwave-infrared satellite observations of the Orbiting Carbon Observatory-2 mission,” *Remote Sensing*, 3(2), 270–304.

Bousquet, P., Peylin, P., Ciais, P., Le Quéré, C., Friedlingstein, P., and Tans, P. P. (2000), “Regional Changes in Carbon Dioxide Fluxes of Land and Oceans Since 1980,” *Science*, 290(5495), 1342–1346.

Box, G. E. P. (1980), “Sampling and Bayes Inference in Scientific Modelling and Robustness,” *Journal of the Royal Statistical Society Series A*, 143(4), 383–430.

Brooks, S. P., and Gelman, A. (1998), “General Methods for Monitoring Convergence of Iterative Simulations General Methods for Monitoring Convergence of Iterative Simulations,” *Journal of Computational and Graphical Statistics*, 7(4), 434–455.

Connor, B., Bösch, H., McDuffie, J. et al. (2016), “Quantification of uncertainties in OCO-2 measurements of XCO2: Simulations and linear error analysis,” *Atmospheric Measurement Techniques*, 9(10), 5227–5238.

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.

Crisp, D., Atlas, R., Breon, F.-M. et al. (2004), “The Orbiting Carbon Observatory (OCO) mission,” *Advances in Space Research*, 34(4), 700 – 709.

Crisp, D., Boesch, H., Brown, L. et al. (2014), OCO - 2 Level 2 Full Physics Retrieval Algorithm Theoretical Basis, Technical Report OCO D–65488, NASA Jet Propultion Laboratory, OCO D-65488, Pasadena.

Crisp, D., E.Miller, C., and DeCola, P. L. (2007), “NASA Orbiting Carbon Observatory: measuring the column averaged carbon dioxide mole fraction from space,” *Journal of Applied Remote Sensing*, https://doi.org/10.1117/1.2898457.

Crisp, D., Fisher, B. M., O'Dell, C. et al. (2012), "The ACOS CO_2 retrieval algorithm - Part II: Global X_{CO_2} data characterization," *Atmospheric Measurement Techniques*, 5(4), 687–707.

Crisp, D., and Johnson, C. (2005), "The orbiting carbon observatory mission," *Acta Astronautica*, 56(1-2), 193–197.

Dawid, A. P. (1984), "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society Series A*, 147(2), 278–292.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–883.

Eldering, A., O'Dell, C. W., Wennberg, P. O. et al. (2017), "The Orbiting Carbon Observatory-2: First 18 months of science data products," *Atmospheric Measurement Techniques*, 10(2), 549–563.

Engelen, R. J., Denning, A. S., and Gurney, K. R. (2002), "On error estimation in atmospheric CO2 inversions," *Journal of Geophysical Research: Atmospheres*, 107(D14), ACL 10–1 – ACL 10–13.

Friedlingstein, P., Cox, P., Betts, R. et al. (2006), "Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison," *Journal of Climate*, 19(14), 3337–3353.

Gelman, A., and Rubin (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7(4), 457–511.

GLOBALVIEW-CO2 (2013), Cooperative Global Atmospheric Data Integration Project. 2013, updated annually. Multi-laboratory compilation of synchronized and gap-filled atmospheric carbon dioxide records for the period 1979-2012 (obspack_co2_1_GLOBALVIEW-CO2_2013_v1.0.4_2013-12-23), Technical report.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007), "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2), 243–268.

Gneiting, T., and Katzfuss, M. (2014), "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 1(1), 125–151.

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008), "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds," *Test*, 17(2), 211–235.

Gurney, K. R., Law, R. M., Denning, A. S. et al. (2003), "TransCom 3 CO2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information," *Tellus, Series B: Chemical and Physical Meteorology*, 55(2), 555–579.

Haario, H., Laine, M., Lehtinen, M., Saksman, E., and Tamminen, J. (2004), "Markov chain Monte Carlo methods for high dimensional inversion in remote sensing," *J. R. Statist. Soc. B*, 66(3), 591–607.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7(2).

Hamill, T. M., and Colucci, S. J. (1997), "Verification of Eta RSM Short-Range Ensemble Forecasts," *Monthly Weather Review*, 125, 1312–1328.

Hersbach, H. (2000), "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, 15(5), 559–570.

Hobbs, J., Braverman, A., Cressie, N., Granat, R., and Gunson, M. (2017), "Simulation-based Uncertainty Quantification for estmating CO2 from satellite data," *SIAM/ASA Journal on Uncertainty Quantification*, 5(1), 956–985.

Jordan, A., Krueger, F., and Lerch, S. (2017), *scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts*.

Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., and Yung, Y. L. (2013), "A Systematic Retrieval Analysis of Secondary Eclipse Spectra. I. a Comparison of Atmospheric Retrieval Techniques," *The Astrophysical Journal*, 775(2), 137.

López-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104(486), 718–734.

Matheson, J. E., and Winkler, R. L. (1976), "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22(10), 1087–1096.

Miller, C. E., Crisp, D., DeCola, P. L. et al. (2007), "Precision requirements for space-based data," *Journal of Geophysical Research: Atmospheres*, 112(D10314).

O'Dell, C. W., Connor, B., Bösch, H. et al. (2012), "The ACOS CO_2 retrieval algorithm - Part 1: Description and validation against synthetic observations," *Atmospheric Measurement Techniques*, 5(1), 99–121.

O'Hagan, A. (2003), "HSSS model criticism," in *Highly structured stochastic systems*, eds. P. J. Green, N. L. Hjort, and S. Richardson, Oxford: Oxford University Press, pp. 423–444.

Rienecker, M. M., Suarez, M. J., Gelaro, R. et al. (2011), "MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications," *Journal of Climate*, 24(14), 3624–3648.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding* World Scientific.

Schuh, A. E., Denning, A. S., Corbin, K. D., Baker, I. T., Uliasz, M., Parazoo, N., Andrews, A. E., and Worthy, D. E. J. (2010), "A regional high-resolution carbon flux inversion of North America for 2004," *Biogeosciences*, 7(5), 1625–1644.

Sun, Y., and Genton, M. G. (2011), "Functional Boxplots," *Journal of Computational and Graphical Statistics*, 20(2), 316–334.

—(2012), "Adjusted functional boxplots for spatio-temporal data visualization and outlier detection," *Environmetrics*, 23(1), 54–64.

Sun, Y., Genton, M. G., and Nychka, D. W. (2012), "Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked?," *Stat*, 1(1), 68–74.

Talagrand, O., Vautard, R., and Strauss, B. (1997), Evaluation of probabilistic prediction systems, in *Proceedings of a workshop held at ECMWF on predictability, 2022 October 1997. European Centre for Medium- Range Weather Forecasts*, pp. 1–25.

Thompson, D. R., Benner, D. C., Brown, L. R. et al. (2012), "Atmospheric validation of high accuracy CO2 absorption coefficients for the OCO-2mission," *Journal of Quantitative Spectroscopy & Radiative Transfer*, 113, 2265–2276.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. (2016), "Assessing the calibration of high-dimensional ensemble forecasts using rank histograms," *Journal of Computational and Graphical Statistics*, 25(1), 105–122.

Tukiainen, S., Railo, J., Laine, M., Hakkarainen, J., Kivi, R., Heikkinen, P., Chen, H., and Tamminen, J. (2016), "Retrieval of atmospheric CH4 profiles from Fourier transform infrared data using dimension reduction and MCMC," *Journal of Geophysical Research: Atmospheres*, 121(17), 10,312–10,327.

Van Der Maaten, L. J. P., and Hinton, G. E. (2008), "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, 9, 2579–2605.

Wang, Y., Jiang, X., Yu, B., and Jiang, M. (2013), "A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data," *Journal of the American Statistical Association*, 108(502), 483–493.