ESA Climate Change Initiative “Plus” (CCI+)

Algorithm Theoretical Basis Document
Version 4 (ATBDv4) for the FOCAL XCO2
OCO-2 Data Product CO2_OC2_FOCA (v10)
for the Essential Climate Variable (ECV)
Greenhouse Gases (GHG)

Authors:
M.Reuter (mreuter@iup.physik.uni-bremen.de),
M.Hilker, S.Noël, M.Buchwitz, H.Bovensmann, and J.P.Burrows
Institute of Environmental Physics (IUP) / Institute of Remote Sensing (IFE),
University of Bremen (UB), Bremen, Germany

The further development of the FOCAL retrieval algorithm and corresponding OCO-2 data
processing and analysis is co-funded by:

ESA via CCI project GHG-CCI+
EUMETSAT via the FOCAL-CO2M study
The European Commission via the H2020 project
VERIFY (Grant Agreement No. 776810)
GHG-CCI+ project

ESA Climate Change Initiative “Plus” (CCI+)
Algorithm Theoretical Basis Document
Version 4 (ATBDv4) for the FOCAL XCO2
OCO-2 Data Product CO2_OC2_FOCA (v10)
for the Essential Climate Variable (ECV)
Greenhouse Gases (GHG)

Change log:

<table>
<thead>
<tr>
<th>Version Nr.</th>
<th>Date</th>
<th>Status</th>
<th>Reason for change</th>
</tr>
</thead>
</table>
Algorithm Theoretical Basis Document Version 4 (ATBDv4) -

Retrieval of XCO$_2$ from the OCO-2 satellite using the
Fast Atmospheric Trace Gas Retrieval (FOCAL)

ESA Climate Change Initiative "Plus" (CCI+)
for the Essential Climate Variable (ECV)

Greenhouse Gases (GHG)

Prepared by:

M. Reuter, M. Hilker, S. Noël, M. Buchwitz, O. Schneising, H. Bovensmann, and J. P. Burrows

Institute of Environmental Physics (IUP)
University of Bremen, FB1
PO Box 33 04 40
D-28334 Bremen
Germany
## Contents

1 Introduction 6

2 Algorithm Overview 8
   2.1 Physical Basis . . . . . . . . . . . . . . . . . . . . . . . . 8
   2.2 Input Data . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   2.3 Output Data . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   2.4 Computational Efficiency . . . . . . . . . . . . . . . . . . . 10

3 Radiative Transfer 12
   3.1 Radiance transmission . . . . . . . . . . . . . . . . . . . . . 12
   3.2 Upward irradiance (diffuse) transmission . . . . . . . . . . . 14
   3.3 Downward irradiance (diffuse) transmission . . . . . . . . . . . 15
   3.4 Solar radiation . . . . . . . . . . . . . . . . . . . . . . . . 16
   3.5 Single scattering radiance from the scattering layer . . . . . . 16
   3.6 Multiple scattering radiance from the surface due to direct
       illumination of the surface . . . . . . . . . . . . . . . . . . . 16
   3.7 Multiple scattering radiance from the scattering layer due to
       direct illumination of the surface . . . . . . . . . . . . . . . . 17
   3.8 Multiple scattering radiance from the surface due to diffuse
       illumination of the surface . . . . . . . . . . . . . . . . . . . . 18
   3.9 Multiple scattering radiance from the scattering layer due to
       diffuse illumination of the scattering layer . . . . . . . . . . . 18
   3.10 Radiance from solar induced fluorescence . . . . . . . . . . . . 19
   3.11 Approximations . . . . . . . . . . . . . . . . . . . . . . . . 19
   3.12 Pseudo-spherical geometry . . . . . . . . . . . . . . . . . . . 20

4 Retrieval 22
   4.1 Measurement vector $\mathbf{y}$ . . . . . . . . . . . . . . . . . . 22
   4.2 Measurement error covariance matrix $\mathbf{S}_e$ . . . . . . . . . . . 22
   4.3 Forward model $\mathbf{F}$ . . . . . . . . . . . . . . . . . . . . . 24
   4.4 State vector $\mathbf{x}$ . . . . . . . . . . . . . . . . . . . . . . . 25
   4.5 A priori error covariance matrix $\mathbf{S}_a$ . . . . . . . . . . . . 26
   4.6 Jacobian matrix $\mathbf{K}$ . . . . . . . . . . . . . . . . . . . . . 29
   4.7 Parameter vector $\mathbf{b}$ . . . . . . . . . . . . . . . . . . . . . 29
   4.8 A posteriori error covariance matrix $\hat{\mathbf{S}}$ . . . . . . . . . . . 30
   4.9 Levenberg-Marquardt damping parameter $\gamma$ . . . . . . . . . . . 30
   4.10 Convergence . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
5 Preprocessing
   5.1 Data collection and preparation .................................. 31
   5.2 Filtering .......................................................... 31
   5.3 Cross-section scaling ............................................. 35
   5.4 Noise Model ....................................................... 35
   5.5 Zero level offset correction ................................. 39

6 Postprocessing ......................................................... 43
   6.1 Filtering .......................................................... 43
   6.2 Bias correction .................................................... 45

7 Version History ......................................................... 51

References ............................................................... 54
1 Introduction

Satellite retrievals of the atmospheric column-average dry-air mole fraction of CO$_2$ ($X_{CO_2}$) based on hyper-spectral measurements in appropriate near (NIR) and short wave infrared (SWIR) O$_2$ and CO$_2$ absorption bands can help to answer pressing questions about the carbon cycle (e.g., Reuter et al., 2017a). However, the precision and even more the accuracy requirements for applications like surface flux inversion or emission monitoring are demanding (e.g., Miller et al., 2007; Chevallier et al., 2007; Bovensmann et al., 2010). As an example, large scale biases of a few tenths of a ppm can already hamper an inversion with mass-conserving global inversion models (Miller et al., 2007; Chevallier et al., 2007).

The Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY, Burrows et al., 1995; Bovensmann et al., 1999) became operational in 2002 and its radiance measurements allowed to start the time series of NIR/SWIR $X_{CO_2}$ retrievals. With an overlap of about three years, the Greenhouse Gases Observing Satellite (GOSAT, Kuze et al., 2009) allowed complementation and continuation of this time series in 2009.

The Orbiting Carbon Observatory-2 (OCO-2) was launched in 2014 also aiming at continuing and improving $X_{CO_2}$ observations from space. As part of the A-train satellite constellation, OCO-2 flies in a sun-synchronous orbit crossing the equator at 13:36 local time. It measures one polarization direction of the solar backscattered radiance in three independent wavelength bands: the O$_2$-A band at around 760 nm (band1) with a spectral resolution of about 0.042 nm and a spectral sampling of about 0.015 nm, the weak CO$_2$ band at around 1610 nm (band2) with a spectral resolution of about 0.080 nm and a spectral sampling of about 0.031 nm, and the strong CO$_2$ band at around 2060 nm (band3) with a spectral resolution of about 0.103 nm and a spectral sampling of about 0.040 nm. OCO-2 is operated in a near-push-broom fashion and has eight footprints across track and an integration time of 0.333 s. The instrument’s spatial resolution at ground is 1.29 km across track and 2.25 km along track. See Crisp et al. (2004) for more information on the OCO-2 instrument.

Multiple scattering of light at aerosols and clouds can be a significant error source for $X_{CO_2}$ retrievals. Therefore, so called full physics retrieval algorithms were developed aiming to minimize scattering related errors by explicitly fitting scattering related properties such as cloud water/ice content, aerosol optical thickness, cloud height, etc. However, the computational costs for multiple scattering radiative transfer (RT) calculations can be immense. Processing
all data of the Orbiting Carbon Observatory-2 (OCO-2) can require up to thousands of CPU cores and the next generation of CO₂ monitoring satellites will produce at least an order of magnitude more data. For this reason, the Fast atmOospheric traCe gAs retrievA FOCAL has been developed reducing the computational costs by orders of magnitude by approximating multiple scattering effects with an analytic solution of the RT problem of an isotropic scattering layer.

This algorithm theoretical basis document (ATBD) describes FOCAL in detail as used for the retrieval of XCO₂ from OCO-2. In parts, this document is compiled from text and figures of the publications of Reuter et al. (2017c,b). Reuter et al. (2017c) described the physical and mathematical basis of FOCAL’s radiative transfer (RT) and assessed the quality of a proposed FOCAL based OCO-2 XCO₂ retrieval algorithm by confronting it with accurate multiple scattering vector RT simulations covering, among others, some typical cloud and aerosol scattering scenarios. This initial FOCAL OCO-2 XCO₂ algorithm with the version number v01 has only been used for theoretical studies based on simulated measurements.

Reuter et al. (2017b) adapted this algorithm and confronted FOCAL for the first time with actually measured OCO-2 data and protocolled the steps undertaken to transform the input data (most importantly, the OCO-2 radiances) into a validated XCO₂ data product. This includes preprocessing, adaptation of the noise model, zero level offset correction, post-filtering, bias correction, comparison with the CAMS (Copernicus Atmosphere Monitoring Service) greenhouse gas flux inversion model, comparison with NASA’s operational OCO-2 XCO₂ product, and validation with ground based Total Carbon Column Observing Network (TCCON) data. Their FOCAL OCO-2 XCO₂ algorithm has the version number v06 and is the base for further developments also described in this ATBD.

The FOCAL OCO-2 XCO₂ algorithm (in the following for the sake of simplicity referred to as FOCAL) is being continuously developed further and the most recent version is v10. A version history itemizing the main changes from version to version can be found in Section 7.
2 Algorithm Overview

2.1 Physical Basis

The FOCAL OCO-2 XCO\textsubscript{2} algorithm described in this ATBD fits the OCO-2 measured radiance simultaneously in four fit windows: SIF (∼758.26–759.24 nm), O\textsubscript{2} (∼757.65–772.56 nm), wCO\textsubscript{2} (∼1595.0–1620.6 nm), and sCO\textsubscript{2} (∼2047.3–2080.9 nm). This is achieved by iteratively optimizing the state vector including, among others, the following geophysical parameters: five layered CO\textsubscript{2} and H\textsubscript{2}O concentration profiles, the pressure (i.e., height), scattering optical thickness at 760 nm, and the Ångström exponent of a scattering layer, solar induced chlorophyll fluorescence (SIF), and polynomial coefficients describing the spectral albedo in each fit window. The fit is performed using the optimal estimation formalism (Rodgers, 2000) and Levenberg-Marquardt minimization of the cost function.

The RT model (RTM) of FOCAL approximates multiple scattering effects at an optically thin isotropic scattering layer. It splits up the top of atmosphere (TOA) radiance into parts originating from direct reflection at the scattering layer or the surface and parts originating from multiple scattering of the diffuse radiant flux between scattering layer and surface. FOCAL’s relatively simple approximation of the RT problem allows unphysical inputs such as negative scattering optical thicknesses or albedos. This can be an advantage when analyzing measurements including noise and assuming Gaussian a priori error statistics. FOCAL accounts for polarization only implicitly by the retrieval of a variable scattering optical thickness.

The PPDF (photon path-length distribution function) method (e.g., Bril et al., 2007, 2012) gains its computational efficiency by applying the theorem of equivalence to replace computationally expensive multiple scattering RT computations with a set of fast transmission computations. This is conceptually similar to FOCAL which uses an effective transmission function for the diffuse flux. However, different from the PPDF method, FOCAL accounts for multiple scattering by solving the geometric series of successive (flux) scattering events.

In principle, the PPDF method can simulate arbitrary scattering phase functions (SPFs). This is not possible for FOCAL which can only simulate an isotropic scattering layer. However, splitting the radiance into direct and diffuse parts can be interpreted as a SPF with a sharp forward peak and which is isotropic otherwise. This still represents typical Mie SPFs not very well but much better than an entirely isotropic SPF.

Strictly, the theorem of equivalence only applies for spectral regions with constant scattering and reflection properties (Bennartz and Preusker, 2006)
making the PPDF shape, e.g., depending on surface albedo. This can make it complicated to transfer scattering information from one fit window into another. Reflection and scattering properties of FOCAL are allowed to vary within the fit windows and can be used to transfer information between fit windows, e.g., via the Ångström exponent.

Despite FOCAL is in principle able to account for scattering at an optically thin scattering layer, pre- and post-filtering as well as bias correction is still needed. The strict pre-filtering bases on sounding quality, cloud coverage, radiance level, and others.

In order to consider not only instrumental noise but also (pseudo) noise of the forward model, we set up a noise model that depends on the instrument noise and one free fit parameter which we determine from the residuals of a set of relatively unconstrained retrievals. The noise model suggests that forward model errors (plus potential pseudo noise of the instrument) have a magnitude of $0.8\% - 3.0\%$ of the continuum radiance. This means that in dark scenes the mismatch of simulated and measured radiance is still dominated by the noise of the instrument but in bright scenes (e.g., above deserts) the forward model error dominates.

Apparent or effective zero level offsets can have various reasons such as residual calibration errors or unconsidered spectroscopic effects. For the SIF, and both CO$_2$ fit windows, we found linear relationships between the retrieved zero level offsets and the continuum radiances with slopes between 0.7% and 1.9%. As FOCAL usually does not retrieve the zero level offset (ZLO) per sounding, we correct the measured radiance with the derived linear relationships before the retrieval.

Post-filtering checks for convergence, for fit window residuals being smaller than the thresholds derived from the noise model analyses, and for potential outliers. With about 88%, the rate of converging soundings is generally high. Soundings with too large residuals are more often found above the tropics and in high latitudes. The filter for potential outliers is most active in the region of the south Atlantic anomaly (SAA) and high latitudes. The total post-filtering throughput is about 35%.

We correct for biases in the post-filtered results with a method adapted from Noël et al. (2021, 2022) which bases on a random forest regressor. Its input data consists of a priori known parameters like land/sea fraction, footprint ID, solar zenith angle, satellite zenith angle and retrieved parameters such as the height of the scattering layer, polynomial coefficients, XCO2 uncertainty, and others. Its training data set consists of model XCO2 data which has been verified by TCCON.
2.2 Input Data

OCO-2 v10 L1b data (Eldering et al., 2015; Crisp et al., 2017) obtained from https://daac.gsfc.nasa.gov are the main input for the FOCAL v10 OCO-2 L2 retrieval. One year has a volume of about 6TB. FOCAL uses meteorological profiles from ECMWF ERA5 (http://www.ecmwf.int). These have a data volume of about 19TB per year. Gaseous absorption cross sections are calculated from NASA’s (National Aeronautics and Space Administration) tabulated absorption cross section database ABSCO v5.1 for O2, CO2, and H2O (Thompson et al., 2012), and HITRAN2016 for the water vapor isotopologue HDO (Gordon et al., 2017). We use a high resolution solar irradiance spectrum which we generated by fitting the solar irradiance spectrum of Kurucz (1995) with the high resolution solar transmittance spectrum used by O’Dell et al. (2012).

2.3 Output Data

Only those measurements which fulfill all quality criteria are stored in daily result files in Network Common Data Format (NetCDF). These files contain all the information required for, e.g., surface flux inverse modeling such as retrieved XCO2 values for individual ground pixels, their errors, corresponding averaging kernels, used a priori profiles, etc. Tab. 1 lists all parameters stored in the L2 result files. A detailed description of the file format and the primary parameters as well as a manual on how to correctly use them can be found in the product specification document (PSDv3, Buchwitz et al., 2014). The final L2 database has a data volume of about 2.9GB per year.

2.4 Computational Efficiency

The computational performance of FOCAL is similar to an absorption only retrieval and currently determined by the convolution of the simulated spectra with the instrumental line shape function (ILS). Currently the FOCAL processing scheme runs on a Linux cluster and uses typically about 300-400 Intel CPU cores. In this environment, FOCAL processes one year of pre-processed L1 data in about one week making it about 52 times faster than real-time.
Table 1: List of output parameters contained in daily FOCAL result files in NetCDF file format. Dimensions are defined as number of pixels per orbit (n) and number of profile layers (m=5). More details can be found in the product specification document (PSDv3, Buchwitz et al., 2014).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Type</th>
<th>Dimension</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>solar_zenith_angle</td>
<td>Float</td>
<td>n</td>
<td>Degrees</td>
<td>Solar zenith angle (0°=zenith)</td>
</tr>
<tr>
<td>sensor_zenith_angle</td>
<td>Float</td>
<td>n</td>
<td>Degrees</td>
<td>Satellite zenith angle (0°=nadir)</td>
</tr>
<tr>
<td>time</td>
<td>Double</td>
<td>n</td>
<td>Seconds</td>
<td>Seconds since 01.01.1970 00:00 UTC</td>
</tr>
<tr>
<td>longitude</td>
<td>Float</td>
<td>n</td>
<td>Degrees</td>
<td>Longitude of pixel centre</td>
</tr>
<tr>
<td>latitude</td>
<td>Float</td>
<td>n</td>
<td>Degrees</td>
<td>Latitude of pixel centre</td>
</tr>
<tr>
<td>pressure_levels</td>
<td>Float</td>
<td>n×(m+1)</td>
<td>hPa</td>
<td>Retrieval pressure levels</td>
</tr>
<tr>
<td>pressure_weight</td>
<td>Float</td>
<td>n×m</td>
<td>-</td>
<td>Pressure weights</td>
</tr>
<tr>
<td>sif_760nm</td>
<td>Float</td>
<td>n</td>
<td>mW/m²/sr/nm</td>
<td>Solar-induced chlorophyll fluorescence at 760nm</td>
</tr>
<tr>
<td>xh2o</td>
<td>Float</td>
<td>n</td>
<td>ppm</td>
<td>Retrieved XH₂O</td>
</tr>
<tr>
<td>xh2o_averaging_kernel</td>
<td>Float</td>
<td>n×m</td>
<td>-</td>
<td>Normalized column averaging kernel for XH₂O</td>
</tr>
<tr>
<td>xh2o_profile_aprior</td>
<td>Float</td>
<td>n×m</td>
<td>ppm</td>
<td>A priori H₂O profile</td>
</tr>
<tr>
<td>xh2o_quality_flag</td>
<td>Float</td>
<td>n</td>
<td>-</td>
<td>Quality flag for XH₂O retrieval (0=good)</td>
</tr>
<tr>
<td>xco2</td>
<td>Float</td>
<td>n</td>
<td>ppm</td>
<td>Retrieved XCO₂</td>
</tr>
<tr>
<td>xco2_averaging_kernel</td>
<td>Float</td>
<td>n×m</td>
<td>-</td>
<td>Normalized column averaging kernel for XCO₂</td>
</tr>
<tr>
<td>xco2_profile_aprior</td>
<td>Float</td>
<td>n×m</td>
<td>ppm</td>
<td>A priori CO₂ profile</td>
</tr>
<tr>
<td>xco2_quality_flag</td>
<td>Float</td>
<td>n</td>
<td>-</td>
<td>Quality flag for XCO₂ retrieval (0=good)</td>
</tr>
</tbody>
</table>
3 Radiative Transfer

This section describes the radiative transfer scheme used by FOCAL. Let, for now, the model atmosphere consist of a plane parallel, vertically heterogeneous, absorbing atmosphere, a surface with Lambertian reflectance, and an optically thin scattering layer of infinitesimal geometrical thickness and with an isotropic SPF (Fig. 1). Light hitting the scattering layer may either be transmitted without interaction, absorbed, or isotropically scattered. In the following, we derive an equation for the satellite measured radiance $I$ for a plane parallel geometry; in Sec. 3.12, we adapt our results for a pseudo spherical geometry.

We separate the radiance reaching the satellite instrument in the components $I_C$, $I_{SD}$, $I_{CD}$, $I_{SI}$, $I_{CI}$, and $I_{SIF}$:

$$I = I_C + I_{SD} + I_{CD} + I_{SI} + I_{CI} + I_{SIF}$$

$(I_C)$ is the radiance directly scattered from the scattering layer to the satellite. $I_{SD}$ represents the radiance originating from the surface due to direct illumination of the surface and includes components due to multiple scattering of the Lambertian surface flux $(I_{SD})$. $I_{CD}$ represents the radiance originating from the scattering layer due to direct illumination of the surface including components due to multiple scattering $(I_{CD})$. $I_{SI}$ represents the radiance originating from the surface due to diffuse illumination of the surface including components due to multiple scattering $(I_{SI})$. $I_{CI}$ represents the radiance originating from the scattering layer due to diffuse illumination of the surface including components due to multiple scattering $(I_{CI})$. $I_{SIF}$ is the radiance originating from solar induced chlorophyll fluorescence at 760 nm (SIF) transmitted through the scattering layer but ignoring multiple scattering because of the weak signal.

If not otherwise noted, in the following, $F$ stands for flux, $I$ for intensity (radiance), $T$ for transmittance, $\tau$ for vertical optical thickness, and $g$ for gaseous absorption. A superscript $s$ stands for the scattering layer in general. The subscripts $e$, $a$, and $s$ stand for extinction, absorption, and scattering of the scattering layer, respectively. As an example, the term $T^g_I$ represents a transmittance of intensity through a gaseous absorber.

3.1 Radiance transmission

The transmittance $T^g_I$ along a slant light path through a plane parallel atmospheric layer with gaseous absorption can be computed with Beer-Lambert’s
Figure 1: Schematic of the FOCAL radiative transfer forward model with an absorbing atmosphere, a surface with Lambertian reflectance, and an optically thin semi-transparent layer which can partly transmit, absorb, or scatter light in an isotropic way. $F_0$ is the solar incoming flux, $\theta_0$ and $\theta$ are the solar and satellite zenith angles, and $I$ is the radiance reaching the satellite instrument split into components as discussed in the main text. Red represents radiation originating from direct illumination of the surface. Green represents radiation originating from direct illumination of the scattering layer. Arrows represent radiance components reaching the satellite instrument originating from the surface (solid) or from the scattering layer (dashed). Waved lines represent diffuse radiant fluxes.
law

$$T_g^g(\tau_g, \zeta) = e^{-\zeta \int K(z) \, dz} = e^{-\zeta \, \tau_g} \quad (2)$$

with $K$ being the absorption coefficient, $z$ the height above the surface, $\tau_g$ the total vertical optical thickness, and $\zeta = 1 / \cos \theta$ the light path extension for the zenith angle $\theta$.

Considering light scattering and absorption within the scattering layer, the fraction of light transmitted through the scattering layer becomes

$$T_s^s(\tau_e, \zeta) = e^{-\tau_e \, \zeta} = 1 - S_I(\tau_s, \tau_e, \zeta) - A_I(\tau_a, \tau_e, \zeta); \quad (3)$$

with $\tau_e = \tau_a + \tau_s$ being the extinction optical thickness, i.e., the sum of absorption (not to be confused with gaseous absorption) and scattering optical thickness. $S_I$ and $A_I$ are the fraction of scattered and absorbed radiance within the scattering layer:

$$S_I(\tau_s, \tau_e, \zeta) = \frac{\tau_s}{\tau_e} [1 - T_g^s(\tau_e, \zeta)] \quad (4)$$

$$A_I(\tau_a, \tau_e, \zeta) = \frac{\tau_a}{\tau_e} [1 - T_g^s(\tau_e, \zeta)] \quad (5)$$

### 3.2 Upward irradiance (diffuse) transmission

The surface is assumed to scatter light in a Lambertian way, thus the bidirectional reflectance distribution function (BRDF) is:

$$B_L(\theta) = \frac{1}{\pi} \cos \theta. \quad (6)$$

Therefore, the transmittance of the scattered radiant flux originating from the Lambertian surface through a plane parallel atmospheric layer can be computed by integrating over the hemisphere (see, e.g., the textbook of Roedel and Wagner (2011)):

$$T_{Flam}^g(\tau_g) = \int_0^{2\pi} \int_0^{\pi/2} e^{-\tau_g \, \zeta} B_L(\theta) \sin \theta \, d\theta \, d\varphi. \quad (7)$$

Integration over the azimuth angle $\varphi$ and substituting $\zeta = 1 / \cos \theta$ gives

$$T_{Flam}^g(\tau_g) = 2 \int_1^{\infty} \frac{e^{-\tau_g} \, \zeta}{\zeta^3} \, d\zeta, \quad (8)$$
which is basically the definition of the third exponential integral $E_3$

$$T_{F,sm}^g(\tau_g) = 2 \ E_3(\tau_g). \tag{9}$$

Analogously, the flux transmitted through the atmosphere below the scattering layer (with gaseous optical thickness $\tau_g$) plus the scattering layer becomes

$$T_{F}^{gs}(\tau_g + \tau_e) = 2 \ E_3(\tau_g + \tau_e) \tag{10}$$

so that the relative additional extinction due to the scattering layer becomes

$$E_F(\tau_e, \tau_g) = 1 - \frac{E_3(\tau_g + \tau_e)}{E_3(\tau_g)}. \tag{11}$$

This can be separated into a fraction of scattered and absorbed flux within the scattering layer:

$$S_F(\tau_s, \tau_e, \tau_g) = \frac{\tau_s}{\tau_e} \ E_F(\tau_e, \tau_g) \tag{12}$$

$$A_F(\tau_a, \tau_e, \tau_g) = \frac{\tau_a}{\tau_e} \ E_F(\tau_e, \tau_g) \tag{13}$$

### 3.3 Downward irradiance (diffuse) transmission

The particles of the scattering layer are assumed to have an isotropic scattering phase function

$$P_S = \frac{1}{4\pi}. \tag{14}$$

Note that the surface BRDF is normalized to result one when integrating over a hemisphere while the isotropic scattering phase function is normalized to result one when integrating over the full sphere.

As we assume the scattering layer to be optically thin, multiple scattering within the scattering layer can be neglected and the reflectance function of the scattering layer becomes the scattering phase function. The transmittance of the scattered radiant flux originating from the scattering layer through a plane parallel atmospheric layer can be computed accordingly by replacing the Lambertian reflectance function by the phase function for isotropic scattering multiplied by two (i.e., normalized to result one when integrating over a hemisphere).

$$T_{Fiso}^g(\tau_g) = \int_0^{2\pi} \int_0^{\pi/2} e^{-\frac{\tau_g}{2\tau_e}} \ 2 \ P_S \ \sin \theta \ d\theta \ d\varphi. \tag{15}$$
Integration over the azimuth angle $\phi$ and substituting $\zeta = 1/\cos \theta$ gives

$$T_{F_{iso}}^g(\tau_g) = \int_1^\infty \frac{e^{-\tau_g \zeta}}{\zeta^2} d\zeta, \quad (16)$$

which defines the second exponential integral $E_2$:

$$T_{F_{iso}}^g(\tau_g) = E_2(\tau_g). \quad (17)$$

### 3.4 Solar radiation

Letting the solar incoming irradiant flux be $F_0$, the solar downward flux reaching the scattering layer becomes

$$F = \frac{F_0}{\zeta_0} T_i^g(\tau_i, \zeta_0). \quad (18)$$

Here $\tau_i$ is the gaseous optical thickness above the scattering layer and $\zeta_0$ the light path extension due to the solar zenith angle $\theta_0$. $T_i^g(\tau_i, \zeta_0)$ corresponds to the transmission along the slant light path from the sun to the scattering layer.

The radiance reaching the satellite transmits the upper layer a second time and the radiance components $I_C$, $I_{SD}$, $I_{CD}$, $I_{SI}$, and $I_{CI}$ become proportional to

$$I_0 = \frac{F_0}{\zeta_0} T_i^g(\tau_i, \zeta_0) T_i^g(\tau_i, \zeta) = \frac{F_0}{\zeta_0} T_i^g(\tau_i, \zeta_0 + \zeta) \quad (19)$$

with $\zeta$ being the light path extension due to the satellite zenith angle $\theta$.

### 3.5 Single scattering radiance from the scattering layer

$I_C$ is the radiance directly scattered from the scattering layer to the satellite

$$I_C = \frac{I_0 \zeta}{4\pi} S_i(\tau_s, \tau_e, \zeta_0). \quad (20)$$

### 3.6 Multiple scattering radiance from the surface due to direct illumination of the surface

$I_{SD}$ represents the radiance originating from the surface due to direct illumination of the surface and includes components due to multiple scattering of the Lambertian surface flux ($I_{SD}$). This means, solar radiation transmits directly through the scattering layer ($T_i^g(\tau_i, \zeta_0)$) and the atmosphere below ($T_i^g(\tau_i, \zeta_0)$) and illuminates the surface with the albedo $\alpha$. This produces a Lambertian
upward flux which is in parts transmitted, absorbed, and scattered into the upper hemisphere, or back scattered into the lower hemisphere when reaching the scattering layer. The back scattered part contributes to the illumination of the surface and so on.

The radiance component \(I_{SD_i}\) corresponds to the directly transmitted radiance from the surface through the lower atmosphere \((T^g_i(\tau_e, \zeta))\), the scattering layer \((T^s_i(\tau_e, \zeta))\), and the upper atmosphere after \(i\) diffuse reflections between surface and scattering layer \((\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_i) T^g_{lam}(\tau_g) T^g_{iso}(\tau_g))\).

Summing up all individual radiance components \(I_{SD_i}\) results in the following geometric series:

\[
I_{SD} = \frac{l_0 \alpha}{\pi} T^s_i(\tau_e, \zeta_0) T^s_i(\tau_e, \zeta) T^g_i(\tau_e, \zeta_0) T^g_i(\tau_e, \zeta) \sum_{i=0}^{\infty} \left( \alpha \frac{1}{2} S_F(\tau_s, \tau_e, \tau_i) T^g_{lam}(\tau_g) T^g_{iso}(\tau_g) \right)^i
= \frac{l_0 \alpha}{\pi} T^s_i(\tau_e, \zeta_0) T^s_i(\tau_e, \zeta) T^g_i(\tau_e, \zeta_0) T^g_i(\tau_e, \zeta) \frac{1}{1 - \alpha \frac{1}{2} S_F(\tau_s, \tau_e, \tau_i) T^g_{lam}(\tau_g) T^g_{iso}(\tau_g)}
\]

(21)

3.7 Multiple scattering radiance from the scattering layer due to direct illumination of the surface

\(I_{CD}\) represents the radiance originating from the scattering layer due to direct illumination of the surface and includes components due to multiple scattering of the Lambertian surface flux \((I_{CD_i})\). As for \(I_{SD}\), solar radiation transmits directly through the scattering layer \((T^s_i(\tau_e, \zeta_0))\) and the atmosphere below \((T^g_i(\tau_s, \zeta_0))\) and illuminates the surface with the albedo \(\alpha\). This produces a Lambertian upward flux which is in parts transmitted, absorbed, and scattered into the upper hemisphere, or back scattered into the lower hemisphere when reaching the scattering layer. The back scattered part contributes to the illumination of the surface and so on.

The radiance component \(I_{CD_i}\) originates from the scattering layer due to the diffuse surface flux transmitting the lower atmosphere \((T^g_{lam}(\tau_g))\) and getting scattered into the upper hemisphere \((\frac{1}{2} S_F(\tau_s, \tau_e, \tau_i))\) after \(i\) diffuse reflections between surface and scattering layer \((\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_i) T^g_{lam}(\tau_g) T^g_{iso}(\tau_g))\).

Summing up all individual radiance components \(I_{CD_i}\) results in the following
geometric series:

\[
I_{CD} = \frac{I_0 \alpha \zeta}{4\pi} T_i^s(\tau_e, \zeta_0) S_F(\tau_s, \tau_e, \tau_\perp) T_i^g(\tau_\perp, \zeta_0) T_{F_{lam}}^g(\tau_g) \frac{1}{1 - \frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)}
\] (22)

3.8 Multiple scattering radiance from the surface due to diffuse illumination of the surface

\(I_{SI}\) represents the radiance originating from the surface due to diffuse illumination of the surface by the scattering layer and includes components due to multiple scattering of the isotropic downward flux of the scattering layer (\(I_{SI_i}\)). Here we follow that part of the solar radiation which is isotropically scattered downward by the scattering layer (\(\frac{1}{2} S_i(\tau_s, \tau_e, \zeta_0)\)) and transmitted to the surface (\(T_{F_{iso}}^g(\tau_g)\)). The illuminated surface has the albedo \(\alpha\) and produces a Lambertian upward flux which is in parts transmitted, absorbed, and scattered into the upper hemisphere, or back scattered into the lower hemisphere when reaching the scattering layer. The back scattered part contributes to the diffuse illumination of the surface and so on.

The radiance component \(I_{SI}\) corresponds to the directly transmitted radiance from the surface through the lower atmosphere (\(T_i^g(\tau_l, \zeta)\)), the scattering layer (\(T_i^s(\tau_e, \zeta)\)), and the upper atmosphere after \(i\) diffuse reflections between surface and scattering layer (\(\frac{1}{2} S_F(\tau_s, \tau_e, \tau_\perp) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)\)).

Summing up all individual radiance components \(I_{SI}\), results in the following geometric series:

\[
I_{SI} = \frac{I_0 \alpha}{2\pi} S_I(\tau_s, \tau_e, \zeta_0) T_i^s(\tau_e, \zeta) T_{F_{iso}}^g(\tau_g) T_i^g(\tau_\perp, \zeta) \frac{1}{1 - \frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)}
\] (23)

3.9 Multiple scattering radiance from the scattering layer due to diffuse illumination of the scattering layer

\(I_{CI}\) represents the radiance originating from the scattering layer due to diffuse illumination of the scattering layer and includes components due to multiple scattering of the isotropic downward flux of the scattering layer (\(I_{CI_i}\)). Again we follow that part of the solar radiation which is isotropically scattered downward by the scattering layer (\(\frac{1}{2} S_i(\tau_s, \tau_e, \zeta_0)\)) and transmitted towards the surface (\(T_{F_{iso}}^g(\tau_g)\)). The illuminated surface has the albedo \(\alpha\) and produces a Lambertian
upward flux which is in parts transmitted, absorbed, and scattered into the upper hemisphere, or back scattered into the lower hemisphere when reaching the scattering layer. The back scattered part contributes to the diffuse illumination of the surface and so on.

The radiance component $I_{CI, i}$ originates from the scattering layer due to the diffuse Lambertian surface flux transmitting the lower atmosphere ($T_{F_{lam}, g}^g(\tau_g)$) and getting scattered into the upper hemisphere ($\frac{1}{2} S_F(\tau_s, \tau_e, \tau_i)$) after $i$ diffuse reflections between surface and scattering layer ($\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_i) T_{F_{lam}, g}^g(\tau_g) T_{F_{iso}, g}^g(\tau_g)$).

Summing up all individual radiance components $I_{CI, i}$ results in the following geometric series:

$$I_{CI} = \frac{I_0 \alpha \zeta}{8 \pi} \frac{S_I(\tau_s, \tau_e, \zeta_0) S_F(\tau_s, \tau_e, \tau_i) T_{F_{iso}, g}(\tau_g) T_{F_{lam}, g}(\tau_g)}{1 - \frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_i) T_{F_{lam}, g}(\tau_g) T_{F_{iso}, g}(\tau_g)}$$ (24)

### 3.10 Radiance from solar induced fluorescence

$I_{SIF}$ is the radiance originating from the Lambertian solar induced chlorophyll fluorescence flux $F_{SIF}^0$ at the surface transmitted through the atmosphere ($T_{I, g}(\tau_I + \tau_I, \zeta)$) and the scattering layer ($T_{I, g}(\tau_e, \zeta)$) but ignoring multiple scattering because of the weak signal.

$$I_{SIF} = \frac{F_{SIF}^0}{\pi} T_{I, g}(\tau_e, \zeta) T_{I, g}(\tau_I + \tau_I, \zeta)$$ (25)

### 3.11 Approximations

By means of the following approximations, we are reducing the complexity of the final result which further enhances the computational efficiency. Note that this also considerably reduces the complexity of the analytic partial derivatives needed to compute the Jacobian used by the retrieval.

Due to the high accuracy requirements for the retrieval of greenhouse gases, we are primarily interested in scenarios where scattering at aerosols and clouds is minimal, even if the retrieval algorithm is, in principle, capable of reducing scattering related errors. Additionally, we are primarily interested in accurate greenhouse gas concentrations; inaccuracies in the retrieved scattering properties are less important. For these reasons and because we already assumed that multiple scattering within the scattering layer can be neglected, we make an approximation for small extinction optical thicknesses.
Further, we assume that the spectral signal produced by absorption within the scattering layer cannot easily be disentangled from an albedo and scattering signal. For some cases, it is even identical; e.g., when the single scattering albedo \( \omega = \tau_s/\tau_e \) becomes zero, the absorption and the albedo signal become identical. Therefore, we are not aiming to explicitly retrieve the absorption within the scattering layer and approximate that \( \tau_a = 0 \) (i.e., \( \tau_e = \tau_s \)). As a result, the retrieved albedo and the amount of scattered radiation may be slightly off, which does not pose a problem as long as the retrieved greenhouse gas concentrations are not affected.

First order Taylor series approximation of Eq. 4 and Eq. 3 gives

\[
S_I(\tau_s, \zeta) \approx \zeta \tau_s \quad \text{and} \\
T_s^I(\tau_s, \zeta) \approx 1 - \zeta \tau_s. 
\]

(26)

(27)

The amount of diffuse scattered radiant flux (Eq. 12) simplifies to

\[
S_F(\tau_s, \tau_e, \tau_e) \approx \frac{E_2(\tau)}{E_3(\tau)} \tau_s. 
\]

(28)

Substituting Eq. 26–28 into Eq. 21–25 and subsequently first order Taylor series approximation of Eq. 1 at \( \tau_s = 0 \) yields:

\[
I \approx \frac{F_0}{\pi \zeta_0} T^g(\tau, \zeta_0 + \zeta) \left[ \frac{1}{4} \tau_s \zeta_0 \zeta \right] \times \\
\alpha \left( T^g_s(\tau_s, \zeta_0 + \zeta) \left[ 1 + \tau_s (\alpha E_2^2 - \zeta_0 - \zeta) \right] + \right. \\
\left. \frac{1}{2} \tau_s E_2 \left[ T^g_s(\tau_s, \zeta_0) \zeta + T^g_s(\tau_s, \zeta) \zeta_0 \right] \right] + \\
\frac{F_0 SIF}{\pi} T^g_s(\tau_s + \tau_t, \zeta) \left[ 1 - \tau_s \zeta \right]. 
\]

(29)

3.12 Pseudo-spherical geometry

Due to the spherical geometry of the Earth’s atmosphere (Fig. 2), the (solar and satellite) zenith angle changes with height \( z \).

\[
\theta(z) = \arcsin \left( \frac{r_e}{r_e + z \sin \theta} \right),
\]

with \( r_e \) being the Earth’s radius and \( \theta \) the (solar or satellite) zenith angle at the surface.
Figure 2: Spherical geometry of the Earth’s atmosphere with the Earth’s radius $r_e$, the (solar or satellite) zenith angle $\theta$ at the surface and at the heights $z_1, 2, 3, \ldots$.

Correspondingly, also the light path extensions $\zeta$ and $\zeta_0$ become height dependent. In the following, $\theta$, $\theta_0$, $\zeta$, and $\zeta_0$ shall refer to values defined at the surface. $\theta(z)$, $\theta_0(z)$, $\zeta(z)$, and $\zeta_0(z)$ shall refer to height $z$ (Eq. 30) and $\theta^s$, $\theta_0^s$, $\zeta^s$, and $\zeta_0^s$ shall refer to the scattering layer. This has implications for Eq. 2 which now becomes

$$T_i^g(K(z), \zeta(z)) = e^{-\int K(z) \, \zeta(z) \, dz}.$$  \hspace{1cm} (31)

Additionally, $\zeta$ in Eq. 3, 4, 5, 26, and 28 has to be replaced with the corresponding value at the scattering layer $\zeta^s$.

In order to keep the integrals in Eq. 7 and Eq. 15 simple, we do not account for the spherical geometry for the transmission of the diffuse fluxes contributing to multiple scattering. For this reason, we consider this approach a pseudo-spherical approximation.
4 Retrieval

In this section, the inversion algorithm is described.

The aim of the retrieval is to find the most probable atmospheric (and surface) state, especially XCO2, given an OCO-2 measurement and some a priori knowledge. According to Rodgers (2000) and as done by, e.g., Reuter et al. (2017c), this can be achieved by minimizing the cost function

$$\chi^2 = \frac{1}{m+n} \left[ (\vec{y} - \vec{F}(\vec{x}, \vec{b}))^T S^{-1}_\varepsilon (\vec{y} - \vec{F}(\vec{x}, \vec{b})) \right] + (\vec{x} - \vec{x}_a)^T S^{-1}_a (\vec{x} - \vec{x}_a) \right].$$

(32)

In this equation, m is the number of spectral pixels and n is the number of Elements in the state vector. Reuter et al. (2017c) used the Gauss-Newton method to minimize the cost function. However, due to its superior convergence stability, we here make use of a Levenberg-Marquardt method to minimize the cost function which bases on the following iteration step:

$$\vec{x}_{i+1} = \vec{x}_i + \hat{S}_i \left[ K_i^T S^{-1}_e (\vec{y} - \vec{F}(\vec{x}_i, \vec{b})) - S^{-1}_a (\vec{x}_i - \vec{x}_a) \right]$$

(33)

$$\hat{S}_i = (K_i^T S^{-1}_e K_i + (1 + \gamma) S^{-1}_a)^{-1}.$$  

(34)

All quantities used in these equations are explained and discussed in the following.

4.1 Measurement vector $\vec{y}$

The measurement vector $\vec{y}$ contains those spectral radiance data measured by the instrument from which we want to gain knowledge about the atmosphere (e.g., XCO2). Each of OCO-2’s bands consists of 1016 spectral pixels which we group into four fit windows: SIF (~758.26–759.24 nm), O$_2$ (~757.65–772.56 nm), wCO$_2$ (~1595.0–1620.6 nm), and sCO$_2$ (~2047.3–2080.9 nm). The separate SIF fit window ensures that the SIF information solely comes from free Fraunhofer lines rather than from O$_2$ absorption features which makes it much easier to avoid misinterpretations with scattering properties (Frankenberg et al., 2011). The measurement vector $\vec{y}$ is of dimension $m \times 1$ ($m \approx 2600$) and an example of a simulated and an actual measurement is illustrated in Fig. 3 and Fig. 4, respectively.

4.2 Measurement error covariance matrix $S_\varepsilon$

Strictly speaking, the measurement error covariance matrix does not only quantify the measurement errors and their correlations; it, additionally, accounts
Figure 3: SCIATRAN simulated OCO-2 measurement fitted with FOCAL. Geophysical baseline scenario and 3-Scat retrieval setup, $\theta_0 = 40^\circ$, parallel polarization (see Reuter et al. (2017c) for definitions of geophysical scenarios and retrieval setups). **Top**: Simulated and fitted radiance measurement in gray and red, respectively. **Bottom**: Measurement noise (see Sec. 5.4) and fit residual ($\vec{\epsilon} = \vec{F} - \vec{y}$) in gray/white and red, respectively. An estimate of the goodness of fit (relative to the noise) in fit window $j$ is computed by $\chi_j = (\frac{1}{m_j} \vec{\epsilon}_j^T S^{-1}_y \vec{\epsilon}_j)^{1/2}$.

Figure 4: Same as Fig. 3 but with an actual (not simulated) OCO-2 measurement of June 5, 2015, 12:01 UTC (sounding ID: 2015060512011938).
for the forward model error. The measurement noise is obtained from the OCO-2 L1b files and the forward model uncertainty is assessed as described in Sec. 5.4. We assume the total measurement uncertainty to have no cross correlations, so that $S_\epsilon$ becomes a diagonal matrix. The measurement error covariance matrix is of dimension $m \times m$ and an example is illustrated in Fig. 3 (bottom).

4.3 Forward model $\vec{F}$

The forward model $\vec{F}$ is a vector function of dimension $m \times 1$ that simulates the measurement vector, i.e., OCO-2 radiance measurements. Its inputs are the state $\vec{x}$ and parameter vector $\vec{b}$ defining the geophysical and instrumental state. Primarily, the forward model consists of the RT model described in Sec. 3. The RT computations require a discretization of the atmosphere which we split into 20 homogeneous layers, each containing the same number of dry-air particles (i.e., molecules).

Additionally to the RT calculations, the forward model simulates the instrument by convolving the RT simulations performed on a fixed high resolution wavelength grid with the ILS obtained from the OCO-2 L1b data. Furthermore, the forward model has the ability to simulate zero level offsets (i.e., additive radiance offsets), shift and squeeze the wavelength axes of the fit windows according to Eq. 35, and squeeze the ILS according to Eq. 37.

$$\lambda' = \lambda + \lambda_{sh} + \lambda_n \lambda_{sq}$$  \hspace{1cm} (35)
$$\lambda_n = 2 - 4 \frac{\lambda_1 - \lambda}{\lambda_1 - \lambda_0}$$  \hspace{1cm} (36)

Here $\lambda'$ is the modified wavelength, $\lambda$ the nominal wavelength, $\lambda_{sh}$ the wavelength shift parameter, $\lambda_n$ the normalized nominal wavelength, $\lambda_{sq}$ the wavelength squeeze parameter, and $\lambda_{0,1}$ the minimum or maximum of $\lambda$, respectively. The normalization of $\lambda$ is done in a way that the average absolute value of $\lambda_n$ is approximately one.

The squeezing of the ILS is done by:

$$\lambda'_{ILS} = \lambda_{ILS} ILS_{sq}$$  \hspace{1cm} (37)

Here $\lambda'_{ILS}$ is the modified ILS wavelength computed from the nominal ILS $\lambda_{ILS}$ wavelength and the squeeze parameter $ILS_{sq}$. 
4.4 State vector $\vec{x}$

The state vector $\vec{x}$ consists of all quantities which we retrieve from the measurement and is of dimension $n \times 1$ with $n = 37$. The dry-air mole fractions of water vapor (H$_2$O) and CO$_2$ are retrieved from both CO$_2$ fit windows within five layers splitting the atmosphere into parts containing the same number of dry-air particles. This means, each CO$_2$ and H$_2$O layer spans over four atmospheric layers used for the discretized RT calculations. The CO$_2$ and H$_2$O concentrations are homogeneous within each of the five layers. As also done by Noël et al. (2021), we further improve the H$_2$O fit quality by allowing for variations of the H$_2$O isotopologue HDO by fitting $\delta D$ defined as

$$\delta D = \frac{R_{\text{meas}}}{R_{\text{VSMOW}}} - 1.$$  \hfill (38)

Here $R_{\text{meas}}$ is the ratio of the measured HDO and H$_2$O columns, and $R_{\text{VSMOW}}$ ($3.1152 \times 10^{-4}$) is the corresponding value for Vienna Standard Mean Ocean Water (VSMOW). $\delta D$ is usually given in units of per mill.

XCO$_2$ and XH$_2$O are not part of the state vector but are calculated during the post processing from the layer concentrations.

SIF at 760 nm is derived from the SIF fit window by scaling a SIF reference spectrum $F_{SIF}^0$. The scattering parameters pressure (i.e., height) of the scattering layer $p_s$ (in units of the surface pressure $p_0$), scattering optical thickness at 760 nm $\tau_s$, and Ångström exponent $\AA$ are derived from all fit windows simultaneously.

Within the SIF fit window, FOCAL additionally fits a first order polynomial of the spectral albedo $\alpha P_{0.1}$ and shift and squeeze of the wavelength axis $\lambda_{\text{sh,sq}}$. Within the other fit windows, FOCAL additionally fits a second order polynomial of the spectral albedo $\alpha P_{0.1,2}$, shift and squeeze of the wavelength axis, and a squeeze of the instrumental line shape function $\text{ILS}_{sq}$.

We estimate the first guess zeroth order albedo polynomial coefficients $\alpha P_0$ from the continuum reflectivities $R_0 = \pi \zeta_0 I/F_0$ using up to nine spectral pixels at the fit windows’ lower wavelength length ends. The first guess profiles of H$_2$O and CO$_2$ are obtained from ECMWF (European Centre for Medium-Range Weather Forecasts) analysis fields and SLIMCO2 v2021, respectively. SLIMCO2 is the Simple cLImatological Model for atmospheric CO$_2$ which has been described by (Noël et al., 2022). Version v2021 of SLIMCO2 bases on a CO$_2$ climatology computed from data of NOAA’s (National Oceanic and Atmospheric Administration) CarbonTracker assimilation system corrected for the atmospheric annual mean growth rate obtained from NOAA.
Figure 5: CO$_2$ a priori error covariance computed from randomly chosen SLIMCO2 profiles and corresponding CarbonTracker profiles. The CO$_2$ layer variances have been up-scaled so that the a priori XCO$_2$ uncertainty becomes 7.5 ppm. **Left**: Layer-to-layer correlation matrix of the a priori uncertainty. **Right**: 1σ a priori uncertainty.

Figure 6: As Fig. 5 but for H$_2$O and estimated from day-to-day variations of ECMWF analysis profiles (without variance scaling as done for CO$_2$).

All other first guess state vector elements are scene independent and the a priori state vector $\mathbf{x}_a$ equals the first guess state vector $\mathbf{x}_0$.

Tab. 2 summarizes the state vector composition including the used fit windows, a priori $\mathbf{x}_a$ and first guess $\mathbf{x}_0$ values, a priori uncertainties $\sigma\mathbf{x}_a$, and typical values of a posteriori uncertainties $\sigma\tilde{\mathbf{x}}$ and the degrees of freedom for signal $d_s$.

### 4.5 A priori error covariance matrix $S_a$

The a priori error covariance matrix defines the uncertainties of the a priori state vector elements and their correlations. Its dimensionality is $n \times n$. Except for the CO$_2$ and H$_2$O profile layers, we assume $S_a$ to be diagonal. As described by Reuter et al. (2012), we compute the CO$_2$ layer-to-layer covariances by comparing randomly chosen SLIMCO2 profiles with corresponding CarbonTracker profiles. The CO$_2$ layer variances have been up-scaled so that the a priori XCO$_2$ uncertainty becomes 7.5 ppm. This ensures retrievals to be dominated by the measurement but not the a priori. We estimated the H$_2$O layer-to-layer covariances by analyzing H$_2$O day-to-day variations of ECMWF analysis profiles. CO$_2$ and H$_2$O a priori error covariances are shown in Fig. 5 and Fig. 6. All other (diagonal) elements of $S_a$ are listed in row $\sigma\mathbf{x}_a$ of Tab. 2.
Table 2: FOCAL’s state vector composition. From left to right, the columns represent the name of the state vector element, its sensitivity within the four fit windows, a priori $\vec{x}_a$ and first guess $\vec{x}_0$ value, and the a priori uncertainty $\sigma\vec{x}_a$.

<table>
<thead>
<tr>
<th>State vector element</th>
<th>Fit window sensitivity</th>
<th>SIF</th>
<th>O$_2$</th>
<th>wCO$_2$</th>
<th>sCO$_2$</th>
<th>$\vec{x}_0$</th>
<th>$\vec{x}_a$</th>
<th>$\sigma\vec{x}_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{P_0}^{SIF}$</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_1}^{SIF}$</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_{P_0}^{O_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_1}^{O_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_0}^{wCO_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_1}^{wCO_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_0}^{sCO_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\alpha_{P_1}^{sCO_2}$</td>
<td>•</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$\lambda_{SIF}^{sh}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{SIF}^{sq}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{wCO_2}^{sh}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{wCO_2}^{sq}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{sCO_2}^{sh}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{sCO_2}^{sq}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{O_2}^{sh}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{O_2}^{sq}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{CO_2}^{sh}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_{CO_2}^{sq}$</td>
<td>[nm]</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.01</td>
</tr>
<tr>
<td>ILS$_{O_2}^{sh}$</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.01</td>
</tr>
<tr>
<td>ILS$_{O_2}^{sq}$</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.01</td>
</tr>
<tr>
<td>SIF [mW/m$^2$/sr/nm]</td>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>10.0</td>
</tr>
<tr>
<td>$\rho_p$</td>
<td>[$\lambda_0$]</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>$\tau_s$</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>0.01</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>$\lambda$</td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>4.0</td>
<td>4.0</td>
<td>2.0</td>
</tr>
<tr>
<td>H$_2$O L$_0$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>2179.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H$_2$O L$_1$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>2186.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H$_2$O L$_2$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>1066.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H$_2$O L$_3$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>205.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H$_2$O L$_4$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>2.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$D,[$\‰$]</td>
<td></td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>1000.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO$_2$ L$_0$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>16.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO$_2$ L$_1$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>11.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO$_2$ L$_2$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>8.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO$_2$ L$_3$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>7.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO$_2$ L$_4$</td>
<td>[ppm]</td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>6.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XH$_2$O [ppm]</td>
<td></td>
<td>•</td>
<td>•</td>
<td>ECMWF</td>
<td>ECMWF</td>
<td>898.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XCO$_2$ [ppm]</td>
<td></td>
<td>•</td>
<td>•</td>
<td>SLIMCO2</td>
<td>SLIMCO2</td>
<td>7.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 7: Jacobian matrix computed with FOCAL for the geophysical Rayleigh scenario and the 3-Scat retrieval setup of Reuter et al. (2017c). Within the CO₂ fit windows, an additional dashed line shows the partial derivatives according to τ_s and ρ_s scaled by a factor of 10 and 20, respectively.
4.6 Jacobian matrix $K$

The Jacobian matrix includes the first order derivatives of the forward model with respect to the state vector elements and has a dimensionality of $m \times n$. A measurement can only include information on those state vector elements which have sufficiently linearly independent derivatives. Fig. 7 illustrates the content of a typical example of a Jacobian matrix. Note that the sensitivity to SIF has artificially been set to zero in the $O_2$ fit window in order to ensure, that the SIF information solely comes from the SIF fit window and misinterpretations with scattering parameters are avoided (Frankenberg et al., 2011).

4.7 Parameter vector $\vec{b}$

The state vector includes only a small subset of geophysical and instrumental properties that influence a simulated radiance measurement. All these additional properties are assumed to be known and form the parameter vector $\vec{b}$.

The observation geometry (particularly, the solar and satellite zenith angles $\theta_0$ and $\theta$), Earth/Sun distance, Doppler shifts, ILS, measurement wavelength grid, etc. are used as provided or calculated from data in the satellite L1b orbit files. Atmospheric temperature, pressure, and dry-air sub-column profiles are obtained from ECMWF analysis data. Gaseous absorption cross sections are calculated from NASA’s (National Aeronautics and Space Administration) tabulated absorption cross section database ABSCO or HITRAN.

We use a high resolution solar irradiance spectrum ($F_0$) which we generated by fitting the solar irradiance spectrum of Kurucz (1995) with the high resolution solar transmittance spectrum used by O’Dell et al. (2012), a forth order polynomial within each fit window, and a Gaussian ILS. The used solar induced chlorophyll fluorescence irradiance spectrum ($F_{SIF}^0$) has been obtained from the publication of Rascher et al. (2009) and scaled to $1.0 \text{mW/m}^2/\text{sr}/\text{nm}$ at 760 nm. In order to account for OCO-2 measuring one polarization direction only, we divided the solar and the chlorophyll fluorescence irradiance spectrum by a factor of two.

All FOCAL RT simulations are performed at a high resolution wavelength grid (not to be confused with the measurement wavelength grid) with a sampling distance of 0.001 nm for the SIF and the $O_2$ fit window and 0.0026 nm and 0.0044 nm for the $CO_2$ fit windows.
4.8 A posteriori error covariance matrix $\mathbf{\hat{S}}$

Once convergence is achieved, the a posteriori error covariance matrix includes the a posteriori uncertainties of the retrieved state vector elements and their correlations. It has a dimensionality of $n \times n$.

4.9 Levenberg-Marquardt damping parameter $\gamma$

We implemented a Levenberg-Marquardt minimization method making use of a damping parameter $\gamma$ (Eq. 33). Compared to the conventional Gauss–Newton method, this often improves the convergence behavior in cases of non-quadratic cost minimization by choosing more “conservative” state increments at the cost of potentially more iterations.

Iterations that do not reduce the cost function ($\chi^2_{i+1} \geq \chi^2_i$) are rejected and $\gamma$ is increased. Only iterations that actually improve the cost function ($\chi^2_{i+1} < \chi^2_i$) are accepted. In these cases $\gamma$ is decreased and the iteration step approaches the Gauss-Newton process.

4.10 Convergence

We define that convergence is achieved when the state vector increment is small compared with the a posteriori error. Specifically, we stop iterating once:

$$\frac{1}{n} \left[(\mathbf{x}_i - \mathbf{x}_{i-1})^T \mathbf{\hat{S}}^{-1} (\mathbf{x}_i - \mathbf{x}_{i-1})\right] < 0.5 .$$  \hspace{1cm} (39)

Additionally, we test if $\chi^2$ is smaller than 2. The maximum number of allowed iterations is 15.
5 Preprocessing

In order to analyze actually measured data instead of simulations, pre-filtering of the OCO-2 L1b calibrated radiances, adjustments of the noise model, and accounting for potential zero level offsets is required. During preprocessing, we collect all dynamic input datasets that are needed to run the retrievals and pre-filter soundings with potentially degraded quality or potential cloud contamination.

5.1 Data collection and preparation

The preprocessing files primarily contain OCO-2 L1b radiance measurements, corresponding noise estimates and meteorological information.

We use the spike EOF analysis provided with the OCO-2 L1b data (Eldering et al., 2015) and mask spectral pixels with potentially poor quality (referred to as bad colors), so that these are not attempted to be fitted by FOCAL. This happens predominantly in soundings above South America and the South Atlantic because of contamination by cosmic rays within the SAA caused by the shape of the inner Van Allen radiation belt (Fig. 9).

Meteorological profiles come from ECMWF ERA5 and have a resolution of one hour, $0.25^\circ \times 0.25^\circ$, and 137 height layers. As part of the preprocessor, these profiles are corrected for the actual surface height of the OCO-2 soundings and split into 20 layers containing the same number of dry-air particles.

5.2 Filtering

Due to the demanding precision and accuracy requirements for XCO$_2$ retrievals (e.g., Miller et al., 2007; Chevallier et al., 2007; Bovensmann et al., 2010) and the large amount of OCO-2 data, we prioritize quality over quantity in the course of pre-filtering.

First, we reject all soundings flagged to have potentially reduced quality (quality flag ≠ 0) or failing a data integrity test (e.g., unreasonable sounding ID or time). This filter is referred to as “sounding quality” filter in Fig. 9.

After this, we filter out very dark or bright scenes, i.e., extreme detector fillings. Specifically, we ensure that the continuum radiance in each band is between 5% and 95% of the maximum band radiance as specified by Eldering et al. (2015) (“radiance level” filter in Fig. 9).

We also filter out potentially “tricky” scenes with solar or satellite zenith angles greater than 70°, latitudes beyond ±80°, or extreme surface roughnesses
The cloud filter bases on a random forest classifier (Breiman, 2001) trained to discriminate cloud free and cloudy scenes by analyzing OCO-2 L1b radiance spectra. To train the filter, we need a data set of OCO-2 measurements where we know which measurements are affected by clouds and which are not. For this purpose we use co-located MODIS Aqua (moderate-resolution imaging spectroradiometer aboard Aqua) L2 cloud mask data with a spatial resolution of about 1 km × 1 km (collection 6, MYD35, obtained from https://ladsweb.modaps.eosdis.nasa.gov, Ackerman et al., 2010). In order to generate a binary MODIS cloud mask, we considered all valid MODIS pixels classified as clear or probably clear as cloud free and the remaining valid MODIS cloud mask pixels as cloudy.

OCO-2 and the Aqua satellite are part of the A-train satellite constellation but Aqua is lagging OCO-2 by 15 minutes. Due to the parallax effect and possible cloud movements within the different overflight times, the MODIS cloud mask cannot be used to classify each individual OCO-2 measurement with a high enough degree of confidence. However, we can be relatively sure that a OCO-2 sounding is actually cloud free if MODIS classifies all pixels within a radius of at least 50 km of the OCO-2 footprint as cloud free. Likewise, it is also highly probably that a OCO-2 sounding is actually cloud contaminated if MODIS classifies all pixels within a radius of at least 50 km of the OCO-2 footprint as cloudy. All other cases belong to a third class where the state is more or less uncertain.

In this way, we analyzed 24 days of OCO-2 and MODIS measurements in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05., 23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 15.09., 16.09., 15.10., 16.10., 15.11., 17.11., 12.12., 18.12.) which are representative with respect to spatial distribution, nadir/glint observation geometry, and season. In this data set we identified 3577887 OCO-2 soundings as certainly cloudy and 774674 as certainly cloud free from which we randomly selected 125000 cloudy and 125000 cloud free scenes as training data set (see Fig. 8)) and another 500000 scenes according to their natural abundances (409419 cloudy and 88877 cloud free) as test or validation data set.

The feature set contains those parameters from which the random forest will later predict the status of cloud coverage. During training, these parameters are mapped against the training truth. In our case, the feature set consists of the following parameters which come all from the OCO-2 L1b files: The solar and satellite zenith and azimuth angles, the surface elevation, longitude, latitude, the
Figure 8: Sampling of all 250000 OCO-2 soundings with certain MODIS cloud classification used as training data set. The soundings have been randomly drawn from 24 representative days in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05., 23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 15.09., 16.09., 15.10., 16.10., 15.11., 17.11., 12.12., 18.12.)

continuum and minimum radiance in all three OCO-2 bands, all OCO-2 radiance values as individual spectral pixel and binned in intervals of eight spectral pixels, and all six possible permutations of quotients of the continuum radiances.

The training of the random forest classifier has been realized with the Python machine learning package scikit-learn v0.24.1. The number of trees has been set to 300 and the maximum depth of the trees to 30. All other settings of the random forest classifier corresponded to their default values. The training identified the 25 features listed in Tab. 3 to be the most important.

Confronting the trained random forest classifier with the 500000 soundings of the test or validation data set which has not been involved in the training results in the confusion matrix shown in Tab. 4. According to the confusion matrix, only about 1.4‰ of actually cloud free soundings are being wrongly classified as cloudy which means that only a very small portion of all soundings is being unnecessarily rejected. About 1.7% of the predicted cloud free cases are actually cloudy. These are the cases which may have a negative influence on the retrieval quality.

Fig. 9 gives an overview of the applied preprocessing filters and their throughput. The filters are applied successively in the order as described in this section and the throughput statistics provided in Fig. 9 are cumulative. In total, about
Table 3: 25 most important features of the cloud detection random forest classifier identified during training and their relative importance.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface elevation</td>
<td>0.027574</td>
</tr>
<tr>
<td>Continuum radiance strong CO$_2$ / O$_2$ band</td>
<td>0.020674</td>
</tr>
<tr>
<td>Continuum radiance weak CO$_2$ / O$_2$ band</td>
<td>0.016991</td>
</tr>
<tr>
<td>Continuum radiance O$_2$ / weak CO$_2$ band</td>
<td>0.015964</td>
</tr>
<tr>
<td>Continuum radiance O$_2$ band</td>
<td>0.014231</td>
</tr>
<tr>
<td>Continuum radiance O$_2$ / strong CO$_2$ band</td>
<td>0.012380</td>
</tr>
<tr>
<td>O$_2$ band radiance #184</td>
<td>0.010487</td>
</tr>
<tr>
<td>Solar zenith angle</td>
<td>0.009849</td>
</tr>
<tr>
<td>O$_2$ band radiance #161</td>
<td>0.009106</td>
</tr>
<tr>
<td>Continuum radiance weak CO$_2$ / strong CO$_2$ band</td>
<td>0.008873</td>
</tr>
<tr>
<td>O$_2$ band radiance #162</td>
<td>0.008716</td>
</tr>
<tr>
<td>O$_2$ band radiance #174</td>
<td>0.008509</td>
</tr>
<tr>
<td>O$_2$ band radiance #164</td>
<td>0.008376</td>
</tr>
<tr>
<td>O$_2$ band radiance #128</td>
<td>0.007545</td>
</tr>
<tr>
<td>Continuum radiance strong CO$_2$ / weak CO$_2$ band</td>
<td>0.007360</td>
</tr>
<tr>
<td>O$_2$ band radiance #163</td>
<td>0.007193</td>
</tr>
<tr>
<td>O$_2$ band radiance #176</td>
<td>0.006377</td>
</tr>
<tr>
<td>O$_2$ band radiance #172</td>
<td>0.006282</td>
</tr>
<tr>
<td>O$_2$ band radiance #123</td>
<td>0.006048</td>
</tr>
<tr>
<td>O$_2$ band 8-binned radiance #23</td>
<td>0.005989</td>
</tr>
<tr>
<td>O$_2$ band radiance #151</td>
<td>0.005683</td>
</tr>
<tr>
<td>Minimum radiance strong CO$_2$ band</td>
<td>0.005628</td>
</tr>
<tr>
<td>O$_2$ band radiance #126</td>
<td>0.005221</td>
</tr>
<tr>
<td>O$_2$ band 8-binned radiance #16</td>
<td>0.005094</td>
</tr>
<tr>
<td>O$_2$ band 8-binned radiance #20</td>
<td>0.005045</td>
</tr>
</tbody>
</table>
Table 4: Confusion matrix of the random forest cloud classifier applied to the 500000 soundings of the test or validation data set which has not been involved in the training.

<table>
<thead>
<tr>
<th></th>
<th>Clear</th>
<th>Cloudy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Clear</td>
<td>88877</td>
<td>1577</td>
<td>90454</td>
</tr>
<tr>
<td>Predicted Cloudy</td>
<td>127</td>
<td>409419</td>
<td>409546</td>
</tr>
<tr>
<td>Total</td>
<td>89004</td>
<td>410996</td>
<td>500000</td>
</tr>
</tbody>
</table>

25% of the soundings make it through all the pre-filters, with the cloud filter being the most stringent.

### 5.3 Cross-section scaling

In order to reduce potential systematic biases, we slightly scaled the H$_2$O and CO$_2$ cross-sections so that the resulting XCO2 and XH2O best agrees on average with the corresponding a priori values. For this purpose, we analyzed two month of OCO-2 L1b data (04/2015 and 08/2015) with preliminary noise model and zero level offset correction parameters (see Sec. 5.4 and 5.5). Fig. 10 shows the results of the comparison of the retrieved and the a priori XCO2 and XH2O values. We found that the average retrieved XCO2 had to be divided by a factor of 0.9965 to match the a priori XCO2. The average retrieved XH2O had to be divided by a factor of 1.0043 to match the a priori XH2O. Therefore, we decided to scale the XCO2 and XH2O cross-section data by these factors. All further analysis and data processing has been performed using these cross-section scaling factors.

### 5.4 Noise Model

The measurement error covariance matrix has to account not only for the measurement noise but for the total error including also the forward model error (Reuter et al., 2017c). The measurement noise of the instrument is well known from laboratory measurements and in-flight estimates. In theoretical studies, as those of Reuter et al. (2017c), it is often assumed for convenience, that the measurement noise dominates and that other error components can be neglected, i.e., the noise model is approximated by the measurement noise.

Especially, when analyzing measured data, unknown inaccuracies of the forward model can violate this assumption and lead to larger fit residuals and
Figure 9: Pre-filtering statistics of the 24 days data subset used for the noise model analysis (Sec. 5.4). The filters are applied in the order: Sounding quality, radiance level, LAT/SUZ/SAZ/σALT, cloud filter. The colors represent filter activity and soundings passing all filters are shown in violet. Numbers in brackets represent cumulative filter throughputs.

unrealistic results (and error estimates) because the optimal estimation retrieval puts too much trust in the measurement. This may happen, e.g., due to imperfect knowledge of the ILS, unconsidered spectroscopic effects such as Raman scattering, inaccuracies of the spectroscopic data bases, approximations of the radiative transfer model, or imperfect meteorology.

Ideally, one would reduce the fit residuals to the instrument’s noise level by improving the forward model, but this is often not possible. A potential solution is to fit parts of the residuum by empirical orthogonal functions (EOF) computed from a representative set of measurements as done by Boesch et al. (2015). Another approach is to adjust the noise model so that it accounts for measurement noise plus forward model error (e.g., O’Dell et al., 2012; Yoshida et al., 2013; Heymann et al., 2015) and a variant of this approach is also used by us.

Most forward model errors can be interpreted to result from inaccuracies of the computed (effective) atmospheric transmittance. However, the largest scene-to-scene variability of the simulated radiance is due to changes of, e.g., albedo and solar zenith angle. Therefore, it is reasonable to assume forward model errors to be approximately proportional to the continuum signal $I_{\text{cont}}$ which
Figure 10: Comparison of the retrieved (red) and a priori (gray) XCO2 and XH2O values before scaling the line intensities of the cross-section data base.
we obtain from up to nine spectral pixels at the fit windows’ lower wavelength length ends.

We model the root mean square RSR by

\[ RSR = \sqrt{NSR^2 + \delta F^2}, \quad (40) \]

where NSR represents the root mean square of the spectral 1σ radiance noise (as reported in the OCO-2 L1b data) to continuum signal ratio and \( \delta F \) the relative forward model error.

In order to estimate the free parameter \( \delta F \), we analyzed a representative set of pre-filtered soundings (Fig. 11) with a modified FOCAL setup for which we (quadratically) added 2% of the continuum radiance to the measurement noise. This overestimation of the expected total uncertainty effects that the retrieval usually converges towards values being not very far away from the a priori, i.e., values being more or less realistic. Additionally, we switched off the SIF retrieval (which is basically identical to a zero level offset in the SIF fit window) and switched on the retrieval of zero level offsets in all four fit windows.

If the instrument noise would dominate the total error, RSR and NSR would (statistically) lie on a 1:1 line. After the removal of outliers (Fig. 12, top/left,
gray dots above the blue line), this is basically the case for the SIF fit window with forward model errors estimated to be about 0.8‰ of the (continuum) signal. The forward model error within the other fit windows is estimated to be between 1.9‰ and 3.0‰ (Fig. 12). This means, the total error in dark scenes (large NSR) is still dominated by the instrumental noise but in bright scenes (small NSR), the forward model error dominates.

Outliers are removed as follows: The data set is grouped in 35 NSR bins. Only bins with more than 500 samples are further considered. Within each bin, RSR should follow a χ²-distribution with as many degrees of freedom as spectral pixels of the fit window. The number of spectral pixels is always large enough to approximate the χ²-distribution with a Gaussian distribution. Outliers represent poor fits, e.g., due to complicated atmospheric conditions which cannot be well described by the forward model. As they usually enhance the RSR, we have to approach the expectation value of RSR from the lowermost values. The 2.28th and 15.9th percentile (Fig. 12, red and orange points) of the Gaussian distribution are two and one standard deviations (2σ) smaller than the expectation value. We used this to estimate the expectation value (Fig. 12, green points) from which we determined the free fit parameter δF of Eq. 40 (numerical values are shown in Fig. 12). Note that adding 4% instead of 2% of the continuum radiance to the measurement noise gave similar results (results of an earlier study not shown here).

Soundings with a RSR being more than 2σ larger than expected from Eq. 40 are considered outliers. For this purpose, we fitted the second order polynomial

\[ 2 \sigma = a_0 + a_1 \text{NSR} + a_2 \text{NSR}^2 \]  

and use it as threshold for the maximal allowed deviation from the RSR model (Fig. 12, blue lines).

We define the noise model which modifies the reported OCO-2 L1b radiance noise \( N \) analog to Eq. 40:

\[ N' = \sqrt{N^2 + I_{\text{cont}}^2 \delta F^2}. \]  

5.5 Zero level offset correction

We define as ZLO an additive fit window-wide radiance offset. An apparent or effective ZLO can have various reasons such as residual calibration errors or unconsidered spectroscopic effects. Many of these effects can be expected to result in ZLOs being approximately proportional to the fit window’s continuum radiance. In order to study potential ZLOs, we used the same modified FOCAL setup as in
Figure 12: Root mean square noise to signal ratio NSR versus root mean square residual to signal ratio RSR for all four fit windows. **red points:** 2.28th percentile within bins with more than 500 samples (35 bins in total). **orange points:** 15.9th percentile. **green points:** expectation value estimated from the 2.28th and 15.9th percentile. **solid green line:** RSR as computed from the RSR model (Eq. 40). **blue points:** RSR model plus 2σ estimated from the 2.28th and 15.9th percentile. **blue line:** outlier threshold. **dashed green line:** one-to-one line.
the last section but with the just defined noise model. The simultaneous retrieval
of ZLOs reduce the uncertainty reduction for XCO₂ and renders the SIF retrieval
impossible. Therefore, we aimed at a ZLO correction rather than a ZLO retrieval
per sounding. We analyzed the same 24 days of OCO-2 data as in the last section
but filtered for potential contamination with chlorophyll fluorescence because in
the SIF fit window it is not possible to disentangle ZLO and SIF (Fig. 13). For
this purpose, we used monthly L3 MODIS Aqua chlorophyll-a data (obtained
from https://modis.gsfc.nasa.gov/data/dataprod/chlor_a.php, Hu et al., 2012)
over ocean and normalized difference vegetation index (NDVI) data over land

Fig. 14 shows that we find a reasonably linear relationship (with correlations
ranging from 0.75 to 0.95) between the retrieved ZLO and the continuum
radiance within the SIF and both CO₂ fit windows hinting at ZLOs in the range
of 0.7%-1.9% of the continuum radiance. Here the linear fit has been performed
by first computing averages in 20 bins (Fig. 14, green dots) and weighting them
according to the inverse of the inner-bin standard deviation. Potential outliers,
i.e., non-converging soundings, χ²>2, or RSR exceeding the threshold of the
noise model (Sec. 5.4)) have been removed beforehand. In the following, we

Figure 13: Sampling of all pre-filtered soundings analyzed in order to determine the
ZLO correction. The data set consists of all pre-filtered OCO-2 soundings of 24
days in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05.,
23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 15.09., 16.09., 15.10., 16.10.,
15.11., 17.11., 12.12., 18.12.) additionally filtered for potential contamination with
chlorophyll fluorescence (see main text).
use the fitted linear relationship as ZLO correction for these three fit windows. In the O$_2$ fit window, the correlation between ZLO and continuum radiance is poor and the linear fit suggests a small positive slope. Therefore, we decided to not apply a ZLO correction for this fit window.

Figure 14: Retrieved zero level offset (ZLO) versus continuum radiance ($I_{\text{cont}}$) for all four fit windows. Soundings without convergence, $\chi^2 > 2$, or RSR exceeding the threshold of the noise model (Sec. 5.4) have been removed beforehand. **green dots:** Binned averages. **green line:** Linear fit through the binned averages weighted by the inverse of the inner-bin standard deviation.
6 Postprocessing

This section describes all postprocessing steps performed by FOCAL.

6.1 Filtering

First of all, we check for convergence, i.e., the state vector increment has to be small compared to the a posteriori uncertainty, the maximum number of iterations must not exceed 15, and $\chi^2$ must not exceed 2 (Sec. 4.10). Convergence is achieved in about 88% of all pre-filtered OCO-2 soundings. Globally there is no extended region where convergence problems are the predominate reason for soundings to be rejected by post-filtering except for a region in/near Ethiopia (Fig. 16).

In the next step, we check for each fit window if the RSR is smaller than the threshold for potential outliers defined in Sec. 5.4. The throughput of this filter, which is most active in the tropics and in high latitudes (Fig. 16), is about 51%.

Additionally, we filter for potential outliers by parameters that have an unexpectedly large influence on the retrieved local XCO$_2$ variability. For the example data shown in Fig. 16, this filter is most active in the SAA and in high latitudes. It has a throughput of about 79%.

This filter bases on the idea that XCO$_2$ outliers increase the local retrieved XCO$_2$ variability and are likely correlated with extreme values of some candidate parameters such as the non CO$_2$ and H$_2$O state vector elements or the continuum radiance in one of the fit windows ($I_{cont}^{O_2}$, $I_{cont}^{wCO_2}$, $I_{cont}^{sCO_2}$, see also Reuter et al., 2017b).

For a representative two months data set (April and August 2015, Fig. 15), we estimated the local retrieved XCO$_2$ variability $\text{VAR}(\Delta \text{XCO}_2)$ as follows: For each sounding, we computed the difference $\Delta \text{XCO}_2$ between XCO$_2$ and its 5°×5° daily median and subsequently, we computed the variance of all $\Delta \text{XCO}_2$ values falling in grid boxes with more than 100 samples. Now we searched for an upper or lower threshold for that candidate parameter which reduces $\text{VAR}(\Delta \text{XCO}_2)$ most when removing 1‰ of all data points. We repeated this until 20% of all data points were removed. In order to reduce the complexity of the postprocessing filter procedure, we now identified the 10 most promising candidate parameters separately for land and ocean and repeated the whole exercise to find filter thresholds for these 10 parameters.

As the sounding density for large solar zenith angles is comparably low, this filter bears the risk of removing large parts of these measurements. We mitigate this risk by defining ten bins of the solar zenith angle and weighting
Figure 15: Sampling of all soundings of April and August 2015 used to compute the post-filtering parameters (Tab. 5).

the computed difference $\Delta XCO_2$ with the number of soundings $N$ falling in the corresponding bin, i.e., we replace the variance $\text{VAR}(\Delta XCO_2)$ by the weighted variance $\text{wVAR}(\Delta XCO_2) = \text{VAR}(N \Delta XCO_2)$. In this way, it is easier for the algorithm to reduce the variance by removing soundings in well populated bins, or in other words, it introduces a penalty for removing soundings in poorly populated bins.

Fig. 17 shows that the decrease in variance somewhat reduces after the removal of the first 10%-15%. A potential interpretation is that in this range indeed primarily outliers are removed. After the removal of approximately 20% the decrease in variability is relatively constant over a larger range before it drops to zero when the last data points are removed. As the curves do not show a distinct kink, the choice to remove 20% of all data points is a bit arbitrary but seemed to be a good compromise.

Above land (Fig. 17, left), the potential outliers filter reduces the variance of $\Delta XCO_2$ from about 3.9 ppm$^2$ to 2.2 ppm$^2$. The Ångström exponent $\AA$ is the dominant parameter, contributing 45% to the variance reduction. All parameter thresholds found for the potential outliers filter above land are listed in Tab. 5 (top).

Above sea (Fig. 17, right), this filter reduces the variance of $\Delta XCO_2$ from about 2.8 ppm$^2$ to about 1.2 ppm$^2$. In glint geometry, scattering is less important and the dominant parameter is albedo polynomial parameter $\alpha P_{\text{SCO}_2}$, contributing 53% to the variance reduction. All parameter thresholds found for
Figure 16: Post-filtering statistics for April and August 2015. The filters are applied in the order: convergence, residual, and potential outliers (see main text for a description). The colors represent filter activity and soundings passing all filters are shown in red. Numbers in brackets represent filter throughputs.

The potential outliers filter above sea are listed in Tab. 5 (bottom).

The combined throughput of all three post-filters (convergence, residual, and potential outliers) is about 35% (Fig. 16)).

6.2 Bias correction

FOCAL’s bias correction scheme has been adapted from the approach of Noël et al. (2021, 2022) who use a random forest regressor (Breiman, 2001; Geurts et al., 2006) to correct for biases in FOCAL retrievals from GOSAT and GOSAT-II.

Our training truth, in the following referred to as true $db$, has been constructed from SLIMCO2 climatological model data of the years 2014–2019. The true $db$ includes only data which have been verified by TCCON. Specifically, we use only those SLIMCO2 data points which agree within 0.1 ppm with TCCON. As this would limit true $db$ values to exist only at TCCON sites, we extend the verified data points to contiguous regions where SLIMCO2’s XCO2 deviates by less than 0.1 ppm from the SLIMXCO2 values at the verified data point (see also Noël et al. (2021) for a description of the true $db$).

This allowed us to find 4007536 OCO-2 soundings of the years 2014–2019.
Table 5: Thresholds and variance reduction of the 10 parameters of the potential outliers filter for soundings above land (top) and sea (bottom). In total, the variance of $\Delta XCO_2$ is reduced from about 3.9 ppm$^2$ to 2.2 ppm$^2$ above land and from about 2.8 ppm$^2$ to 1.2 ppm$^2$ above sea. See Reuter et al. (2017b) and the main text for a description of the individual parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lower threshold</th>
<th>Upper threshold</th>
<th>Variance reduction [ppm$^2$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>À</td>
<td>1.1373</td>
<td>5.7167</td>
<td>9.3343·10$^{-1}$</td>
</tr>
<tr>
<td>ILS$^{wCO_2}_{sq}$</td>
<td>9.9368·10$^{-1}$</td>
<td>1.0011</td>
<td>3.7107·10$^{-1}$</td>
</tr>
<tr>
<td>ILS$^{O_2}_{sq}$</td>
<td>1.0039</td>
<td>1.0219</td>
<td>9.6385·10$^{-2}$</td>
</tr>
<tr>
<td>ILS$^{scO_2}_{sq}$</td>
<td>9.9305·10$^{-1}$</td>
<td>1.0064</td>
<td>8.8602·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{O_2}_2$</td>
<td>-1.5433·10$^{-3}$</td>
<td>1.6265·10$^{-4}$</td>
<td>7.8222·10$^{-2}$</td>
</tr>
<tr>
<td>$\tau_0$</td>
<td>-</td>
<td>1.4814·10$^{-1}$</td>
<td>5.8100·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{SIF}_1$</td>
<td>-4.4338·10$^{-3}$</td>
<td>6.2912·10$^{-3}$</td>
<td>3.4975·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{wCO_2}_2$</td>
<td>-2.2400·10$^{-3}$</td>
<td>-</td>
<td>2.7516·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{scO_2}_2$</td>
<td>-2.4493·10$^{-3}$</td>
<td>6.1081·10$^{-4}$</td>
<td>2.6524·10$^{-2}$</td>
</tr>
<tr>
<td>$\lambda_{sCO_2}$</td>
<td>-</td>
<td>5.7684·10$^{-1}$</td>
<td>1.7935·10$^{-2}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Land</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha P^{scO_2}_1$</td>
<td>-</td>
<td>1.6189·10$^{-5}$</td>
<td>7.5229·10$^{-1}$</td>
</tr>
<tr>
<td>$\tau_0$</td>
<td>-</td>
<td>3.6108·10$^{-2}$</td>
<td>2.8847·10$^{-1}$</td>
</tr>
<tr>
<td>$\rho_s$</td>
<td>-</td>
<td>3.6004·10$^{-1}$</td>
<td>1.6624·10$^{-1}$</td>
</tr>
<tr>
<td>ILS$^{O_2}_{sq}$</td>
<td>1.0150</td>
<td>1.0680·10$^{-1}$</td>
<td>1.0680·10$^{-1}$</td>
</tr>
<tr>
<td>ILS$^{wCO_2}_{sq}$</td>
<td>9.9355·10$^{-1}$</td>
<td>1.0035</td>
<td>7.9369·10$^{-2}$</td>
</tr>
<tr>
<td>À</td>
<td>1.1949</td>
<td>-</td>
<td>6.5081·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{SIF}$</td>
<td>-2.8179·10$^{-3}$</td>
<td>3.0376·10$^{-3}$</td>
<td>5.3789·10$^{-2}$</td>
</tr>
<tr>
<td>$\alpha P^{scO_2}_2$</td>
<td>-1.3674·10$^{-3}$</td>
<td>-3.6583·10$^{-5}$</td>
<td>5.0012·10$^{-2}$</td>
</tr>
<tr>
<td>ILS$^{scO_2}_{sq}$</td>
<td>9.9205·10$^{-1}$</td>
<td>-</td>
<td>3.6732·10$^{-2}$</td>
</tr>
<tr>
<td>$\lambda_{sCO_2}$</td>
<td>-7.6605·10$^{-4}$</td>
<td>-</td>
<td>3.6152·10$^{-2}$</td>
</tr>
</tbody>
</table>
meeting all post-filtering criteria and having co-locations with the true \( \text{db} \). We constructed a training and two test data sets, each of equal size, by randomly drawing about 1.26 million soundings for each data set. In order to assure that the three data sets have a representative and similar spatial and temporal sampling, all soundings were first sorted into a \( 5^\circ \times 5^\circ \) monthly grid. Then, three soundings from each grid box were randomly distributed among the three data sets until all grid boxes contained fewer than three soundings.

Similar to the post-filtering, we wanted to ensure that measurements with large solar zenith angles are not underrepresented in the bias correction. For this reason, we computed the number of soundings for ten solar zenith angle bins and set the training sample weights to the inverse of the number of soundings in the corresponding bins. The sampling of the training data set is illustrated in Fig. 18.

The feature set contains those parameters from which the random forest will later predict the bias. During training, these parameters are mapped against the training truth. Our bias correction feature set consists of the same parameters used as candidate parameters for the outlier detection filter (Sec. 6.1) but with the following additions: OCO-2 footprint ID (1-8), land/sea fraction, XCO2 and XH2O a posteriori uncertainty, observation mode, number of bad colors in each OCO-2 band, XCO2 column averaging kernel in the lowermost layer. Tab. 6
lists the 25 most important features identified during training. As also found by Reuter et al. (2017b), the most important feature is the OCO-2 footprint ID (i.e., the across track sounding ID [1-8]). It is followed by the retrieved height of the scattering layer and the land/sea fraction. The relative importance drops rapidly, and only 10 features have a relative importance greater than 1%.

As for the cloud filtering random forest classifier, we realized the training of the bias correction random forest regressor also with the Python machine learning package scikit-learn v0.24.1. The number of trees has been set to 300 and the maximum depth of the trees to 5. All other settings of the random forest classifier corresponded to their default values.

Before the bias correction, the root mean square (RMS) difference between the retrieved post filtered (raw) XCO2 and the \textit{true db} is 1.848 ppm for the training data set. Fig. 19 illustrates that the largest XCO2 difference to the \textit{true db} can be found, e.g., in northern Africa, the Arabian peninsula, and India. Training the bias model considerably reduces this pattern and the root mean square difference between the bias corrected XCO2 and the \textit{true db} becomes 1.381 ppm for the training data set. The corresponding RMS values for both test data sets are basically identical.

Using the trained random forest regressor to predict the bias for all soundings in April and August 2015 results in Fig. 20 showing a pattern expected from the difference of the raw XCO2 to the \textit{true db} (Fig. 19, left). Because of less noise in this figure, one can also recognize a land/sea bias in addition to the
Table 6: 25 most important features of the bias correction random forest regressor identified during training and their relative importance. See Reuter et al. (2017b) and the main text for a description of the individual parameters.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Footprint ID</td>
<td>3.4359·10^{-1}</td>
</tr>
<tr>
<td>$\rho_s$</td>
<td>2.6965·10^{-1}</td>
</tr>
<tr>
<td>Land/sea fraction</td>
<td>1.5085·10^{-1}</td>
</tr>
<tr>
<td>$\sigma_{\text{XH}_2\text{O}}$</td>
<td>5.5431·10^{-2}</td>
</tr>
<tr>
<td>$\sigma_{\text{XCO}_2}$</td>
<td>4.9654·10^{-2}</td>
</tr>
<tr>
<td>$\theta$</td>
<td>3.5094·10^{-2}</td>
</tr>
<tr>
<td>$\alpha P^2_{\text{CO}_2}$</td>
<td>1.9015·10^{-2}</td>
</tr>
<tr>
<td>$\tau_0$</td>
<td>1.8744·10^{-2}</td>
</tr>
<tr>
<td>$\text{ILS}_{\text{O}_2}^\text{sq}$</td>
<td>1.6530·10^{-2}</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>1.6076·10^{-2}</td>
</tr>
<tr>
<td>$\text{Å}$</td>
<td>6.7967·10^{-3}</td>
</tr>
<tr>
<td>$\alpha P^0_{\text{SIF}}$</td>
<td>5.7567·10^{-4}</td>
</tr>
<tr>
<td>Surface elevation</td>
<td>5.4877·10^{-4}</td>
</tr>
<tr>
<td>$\text{BG}_{\text{WCO}_2}$</td>
<td>4.9943·10^{-4}</td>
</tr>
<tr>
<td>$\alpha P^0_{\text{O}_2}$</td>
<td>2.8850·10^{-4}</td>
</tr>
<tr>
<td>$\alpha P^1_{\text{O}_2}$</td>
<td>2.6743·10^{-4}</td>
</tr>
<tr>
<td>$\text{BG}_{\text{1CO}_2}$</td>
<td>2.5795·10^{-4}</td>
</tr>
<tr>
<td>$\alpha P^2_{\text{CO}_2}$</td>
<td>8.9723·10^{-5}</td>
</tr>
<tr>
<td>$\text{ILS}_{\text{CO}_2}^\text{sq}$</td>
<td>5.8538·10^{-5}</td>
</tr>
<tr>
<td>$\alpha P^2_{\text{WCO}_2}$</td>
<td>2.6632·10^{-5}</td>
</tr>
<tr>
<td>$\text{BG}_{\text{O}_2}$</td>
<td>2.5997·10^{-5}</td>
</tr>
<tr>
<td>SIF</td>
<td>1.2224·10^{-5}</td>
</tr>
<tr>
<td>$\alpha P^0_{\text{NCO}_2}$</td>
<td>9.8717·10^{-6}</td>
</tr>
<tr>
<td>$\lambda^0_{\text{O}_2}$</td>
<td>6.2205·10^{-6}</td>
</tr>
<tr>
<td>$\lambda^{\text{sh}}_{\text{O}_2}$</td>
<td>1.7541·10^{-6}</td>
</tr>
</tbody>
</table>
expected large biases in northern Africa, the Arabian peninsula, and India. The reason for these biases is unclear, but it is reasonable to assume that aerosol scattering in combination with albedo features may cause them which may also explain the observed seasonality of the bias pattern.
7 Version History

v10

Generation of a global 8-years data set with improved coverage. Changes over v09:

- Relaxation of the latitude preprocessor filter from ±70° to ±80°.
- Removal of the OMI L3 based aerosol preprocessor filter.
- Replacement of the MODIS based preprocessor cloud filter by a random forest classifier which is trained with the MODIS MYD35 cloud mask and which analyzes OCO-2 L1b radiances.
- Removal of the preprocessor filter for the number of spectral spikes.
- Spectral spikes are now handled equally to bad detector pixels by masking affected parts of the spectra in the retrieval.
- Improvement of the RT model by assuming an isotropic instead of a Lambertian scattering phase function at the scattering layer.
- Introduction of HDO in the RT model and usage of δD as state vector element.
- Update of the CO₂, O₂, and H₂O cross section tables to ABSCO v5.1 and adding H₂O as absorber in the O₂ fit window.
- Update of the CO₂ a priori to the Simple cLImatological Model for atmospheric CO₂ (SLIMCO2) v2021 which bases on a climatology data base constructed from 16 years of NOAA CarbonTracker data (Noël et al., 2022).
- Update of the CO₂ a priori error covariance matrix corresponding to SLIMCO2 scaled to an XCO₂ a priori uncertainty of 7.5 ppm.
- Replacement of the postprocessing bias correction based on the small area assumption by a random forest regressor trained with a reference data base consisting of NOAA CarbonTracker model data in regions justified with TCCON (Noël et al., 2021).
- Generation of a global 8-years data set (09/2014-02/2022) based on OCO-2 v10 L1b data.
• Improvement of the computational efficiency.
• Bug fixes.

**v09**

Migration of L2 processor from IDL to Python.

Changes over v08:
• Migration of L2 processor from IDL to Python.
• Generation of a global 5-years data set (2015-2019) based on OCO-2 v8 L1b data.
• Extension of the 5-years data set till 05/2020 based on OCO-2 v10 L1b data.
• Usage of previous results as first guess state vector (except for albedo) in order to improve convergence behavior. This acceleration is only applied for soundings of the same orbit having distances below 25km. Additionally, the maximum number of successive accelerated soundings is limited to 25.
• Bug fixes.

**v08**

Generation of a global 4-years data set.

Changes over v06:
• Improved cross section data bases with finer temperature, pressure, and wavelength grid in the wCO$_2$ (0.0026nm) and sCO$_2$ (0.0044nm) band.
• Quadratic wavelength and linear pressure interpolation of the cross section data base.
• Usage of HITRAN2016 as H$_2$O spectroscopy.
• Allowing negative values of $p_s$ for improved convergence behavior.
• Widened limits for improved convergence behavior.
• Improved smoothing and noise error diagnostics.
• Usage of ECMWF ERA5 meteorological data.
• Bug fixes.
v06

First application and validation of the FOCAL OCO-2 XCO$_2$ algorithm to a larger global dataset of actually measured OCO-2 data as described by Reuter et al. (2017b).

Changes over v01:

- Development of a preprocessor including filtering, adaptation of the noise model, and zero level offset correction.
- Development of a postprocessor including filtering and bias correction.
- Implementation of the Levenberg-Marquardt minimizer.
- Bug fixes.

v01

The initial version of the FOCAL OCO-2 XCO$_2$ algorithm as described by Reuter et al. (2017c). This version has been used to analyzed simulated OCO-2 measurements.
References


O’Dell, C. W., Connor, B., Bösch, H., O’Brien, D., Frankenberg, C., Castano,


Reuter, M., Buchwitz, M., Schneising, O., Noël, S., Rozanov, V., Bovensmann, H., and Burrows, J. P.: A fast atmospheric trace gas retrieval for hyperspectral instruments approximating multiple scattering - Part 1: radiative transfer and


