 GHG-CCI+ project	ESA Climate Change Initiative “Plus” (CCI+) Algorithm Theoretical Basis Document Version 5 (ATBDv5) for the FOCAL XCO₂ OCO-2 Data Product CO₂_OC₂_FOCA (v11) for the Essential Climate Variable (ECV) Greenhouse Gases (GHG)	Page 1
		Version 5–Final
		22-February-2024

ESA Climate Change Initiative “Plus” (CCI+)

Algorithm Theoretical Basis Document Version 5 (ATBDv5)

-

for the FOCAL XCO₂ OCO-2 Data Product CO₂_OC₂_FOCA (v11)

for the Essential Climate Variable (ECV)
Greenhouse Gases (GHG)

Authors:
 M.Reuter (mreuter@iup.physik.uni-bremen.de),
 M.Hilker, S.Noël, M.Buchwitz, O.Schneising, 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



 GHG-CCI+ project	ESA Climate Change Initiative “Plus” (CCI+) Algorithm Theoretical Basis Document Version 5 (ATBDv5) for the FOCAL XCO2 OCO-2 Data Product CO2_OC2_FOCA (v11) for the Essential Climate Variable (ECV) Greenhouse Gases (GHG)	Page 2
		Version 5–Final
		22-February-2024

Change log:

Version Nr.	Date	Status	Reason for change
Version 1 – Final 1	23. Aug. 2019	Final Version	New document.
Version 2 – Final 1	24. Aug. 2020	Final Version	Updated document.
Version 3 – Final 1	15. Oct. 2021	Final Version	Updated document.
Version 4 – Final 1	06. Mar. 2023	Final Version	Updated document.
Version 4.1 – Final 1	21. Apr. 2023	Final Version	Updated document.
Version 5 – Final 1	22. Feb. 2024	Final Version	Updated document.

Algorithm Theoretical Basis Document Version 5 (ATBDv5)

-

Retrieval of XCO₂ 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

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	4
---------------------	---	---	----------

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 \vec{y}	22
4.2	Measurement error covariance matrix \mathbf{S}_ϵ	22
4.3	Forward model \vec{F}	24
4.4	State vector \vec{x}	25
4.5	A priori error covariance matrix \mathbf{S}_a	26
4.6	Jacobian matrix \mathbf{K}	29
4.7	Parameter vector \vec{b}	29
4.8	A posteriori error covariance matrix $\hat{\mathbf{S}}$	30
4.9	Levenberg-Marquardt damping parameter γ	30
4.10	Convergence	30

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	5
---------------------	---	---	----------

5	Preprocessing	31
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	41
6	Postprocessing	44
6.1	Filtering	44
6.2	Bias correction	46
6.3	Uncertainty estimates	51
7	Version History	53
	References	56

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	6
---------------------	---	---	----------

1 Introduction

Satellite retrievals of the atmospheric column-average dry-air mole fraction of CO₂ (XCO₂) based on hyper-spectral measurements in appropriate near (NIR) and short wave infrared (SWIR) O₂ and CO₂ 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 XCO₂ 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 XCO₂ 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 O₂-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 CO₂ 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 CO₂ 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 XCO₂ 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

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	7
---------------------	---	---	----------

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 atmOspheric traCe gAs retrieval 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 bases 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 v11. A version history itemizing the main changes from version to version can be found in Section 7.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	8
---------------------	---	---	----------

2 Algorithm Overview

2.1 Physical Basis

The FOCAL OCO-2 XCO₂ algorithm described in this ATBD fits the OCO-2 measured radiance simultaneously in four fit windows: SIF (~ 758.26 – 759.24 nm), O₂ (~ 757.65 – 772.56 nm), wCO₂ (~ 1595.0 – 1620.6 nm), and sCO₂ (~ 2047.3 – 2080.9 nm). This is achieved by iteratively optimizing the state vector including, among others, the following geophysical parameters: five layered CO₂ and H₂O 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)

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	9
---------------------	---	---	----------

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 CO₂ 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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	10
---------------------	---	---	-----------

2.2 Input Data

OCO-2 v11 L1b data (Eldering et al., 2015; Crisp et al., 2017) obtained from <https://daac.gsfc.nasa.gov> are the main input for the FOCAL v11 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. The observation's surface elevation is computed from the high-resolution Copernicus global digital elevation model GLO-30 (European Space Agency, Sinergise, 2021). Gaseous absorption cross sections are calculated from NASA's (National Aeronautics and Space Administration) tabulated absorption cross section database ABSCO v5.2 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 ISS solar reference spectrum of Meftah et al. (2018) with a polynomial multiplied 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) and the product user guide (PUGv5, Reuter et al., 2024a). The final L2 database has a data volume of about 3GB 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 preprocessed L1 data in about one week making it about 52 times faster than real-time.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	11
---------------------	---	---	-----------

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).

Parameter	Type	Dimension	Unit	Description
solar_zenith_angle	Float	n	Degrees	Solar zenith angle (0°=zenith)
sensor_zenith_angle	Float	n	Degrees	Satellite zenith angle (0°=nadir)
time	Double	n	Seconds	Seconds since 01.01.1970 00:00 UTC
longitude	Float	n	Degrees	Longitude of pixel centre
latitude	Float	n	Degrees	Latitude of pixel centre
pressure_levels	Float	n×(m+1)	hPa	Retrieval pressure levels
pressure_weight	Float	n×m	-	Pressure weights
sif_760nm	Float	n	mW/m ² /sr/nm	Solar-induced chlorophyll fluorescence at 760nm
xh2o	Float	n	ppm	Retrieved XH2O
xh2o_uncertainty	Float	n	ppm	Uncertainty in retrieved XH2O
xh2o_averaging_kernel	Float	n×m	-	Normalized column averaging kernel for XH2O
h2o_profile_apriori	Float	n×m	ppm	A priori H ₂ O profile
xh2o_quality_flag	Float	n	-	Quality flag for XH2O retrieval (0=good)
xco2	Float	n	ppm	Retrieved XCO ₂
xco2_uncertainty	Float	n	ppm	Uncertainty in retrieved XCO ₂
xco2_averaging_kernel	Float	n×m	-	Normalized column averaging kernel for XCO ₂
co2_profile_apriori	Float	n×m	ppm	A priori CO ₂ profile
xco2_quality_flag	Float	n	-	Quality flag for XCO ₂ retrieval (0=good)

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	12
---------------------	---	---	-----------

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} :

$$I = I_C + I_{SD} + I_{CD} + I_{SI} + I_{CI} + I_{SIF} \quad (1)$$

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}). 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}). 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}). 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}). 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, τ 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_I^g represents a transmittance of intensity through a gaseous absorber.

3.1 Radiance transmission

The transmittance T_I^g along a slant light path through a plane parallel atmospheric layer with gaseous absorption can be computed with Beer-Lambert's

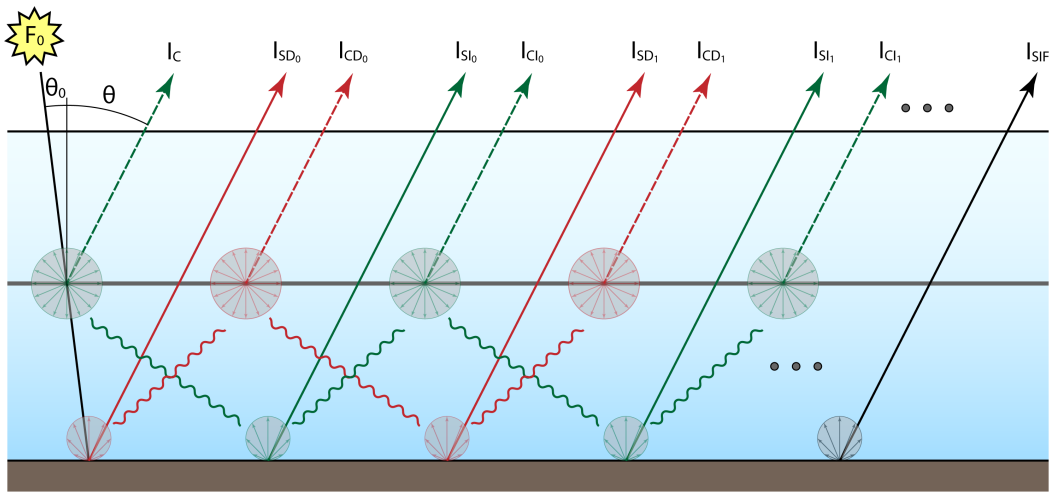


Figure 1: Schematic of the FOCAL radiative transfer forward model with an absorbing atmosphere, a surface with Lambertian reflectance, and an optically thin semi-transparent layer which can partly transmit, absorb, or scatter light in an isotropic way. F_0 is the solar incoming flux, θ_0 and θ are the 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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	14
---------------------	---	---	-----------

law

$$\begin{aligned} T_I^g(\tau_g, \zeta) &= e^{-\zeta \int K(z) dz} \\ &= e^{-\zeta \tau_g} \end{aligned} \quad (2)$$

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

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

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

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:

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

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

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)):

$$T_{F_{lam}}^g(\tau_g) = \int_0^{2\pi} \int_0^{\pi/2} e^{-\frac{\tau_g}{\cos \theta}} B_L(\theta) \sin \theta d\theta d\varphi. \quad (7)$$

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

$$T_{F_{lam}}^g(\tau_g) = 2 \int_1^\infty \frac{e^{-\tau_g \zeta}}{\zeta^3} d\zeta, \quad (8)$$

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	15
---------------------	---	---	-----------

which is basically the definition of the third exponential integral E_3

$$T_{F_{lam}}^g(\tau_g) = 2 E_3(\tau_g). \quad (9)$$

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

$$T_F^{gs}(\tau_\downarrow + \tau_e) = 2 E_3(\tau_\downarrow + \tau_e) \quad (10)$$

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

$$E_F(\tau_e, \tau_\downarrow) = 1 - \frac{E_3(\tau_\downarrow + \tau_e)}{E_3(\tau_\downarrow)}. \quad (11)$$

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

$$S_F(\tau_s, \tau_e, \tau_\downarrow) = \frac{\tau_s}{\tau_e} E_F(\tau_e, \tau_\downarrow) \quad (12)$$

$$A_F(\tau_a, \tau_e, \tau_\downarrow) = \frac{\tau_a}{\tau_e} E_F(\tau_e, \tau_\downarrow) \quad (13)$$

3.3 Downward irradiance (diffuse) transmission

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

$$P_S = \frac{1}{4\pi}. \quad (14)$$

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).

$$T_{F_{iso}}^g(\tau_g) = \int_0^{2\pi} \int_0^{\pi/2} e^{-\frac{\tau_g}{\cos \theta}} 2 P_S \sin \theta d\theta d\varphi. \quad (15)$$

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	16
---------------------	---	---	-----------

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

$$T_{F_{iso}}^g(\tau_g) = \int_1^\infty \frac{e^{-\tau_g \zeta}}{\zeta^3} d\zeta, \quad (16)$$

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

$$F = \frac{F_0}{\zeta_0} T_I^g(\tau_\uparrow, \zeta_0). \quad (18)$$

Here τ_\uparrow is the gaseous optical thickness above the scattering layer and ζ_0 the light path extension due to the solar zenith angle θ_0 . $T_I^g(\tau_\uparrow, \zeta_0)$ 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

$$I_0 = \frac{F_0}{\zeta_0} T_I^g(\tau_\uparrow, \zeta_0) T_I^g(\tau_\uparrow, \zeta) = \frac{F_0}{\zeta_0} T_I^g(\tau_\uparrow, \zeta_0 + \zeta) \quad (19)$$

with ζ being the light path extension due to the satellite zenith angle θ .

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_i}). This means, solar radiation transmits directly through the scattering layer ($T_I^g(\tau_e, \zeta_0)$) and the atmosphere below ($T_I^g(\tau_\downarrow, \zeta_0)$) and illuminates the surface with the albedo α . This produces a Lambertian

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	17
---------------------	---	---	-----------

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 ($T_l^g(\tau_\downarrow, \zeta)$), the scattering layer ($T_l^s(\tau_e, \zeta)$), and the upper atmosphere after i diffuse reflections between surface and scattering layer ($\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)$).

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

$$\begin{aligned}
I_{SD} &= \frac{I_0 \alpha}{\pi} T_l^s(\tau_e, \zeta_0) T_l^s(\tau_e, \zeta) T_l^g(\tau_\downarrow, \zeta_0) T_l^g(\tau_\downarrow, \zeta) \\
&\quad \sum_{i=0}^{\infty} \left(\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g) \right)^i \\
&= \frac{I_0 \alpha}{\pi} T_l^s(\tau_e, \zeta_0) T_l^s(\tau_e, \zeta) T_l^g(\tau_\downarrow, \zeta_0) T_l^g(\tau_\downarrow, \zeta) \\
&\quad \frac{1}{1 - \frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)} \tag{21}
\end{aligned}$$

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 ($T_l^s(\tau_e, \zeta_0)$) and the atmosphere below ($T_l^g(\tau_\downarrow, \zeta_0)$) and illuminates the surface with the albedo α . 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 ($T_{F_{lam}}^g(\tau_g)$) and getting scattered into the upper hemisphere ($\frac{1}{2} S_F(\tau_s, \tau_e, \tau_\downarrow)$) after i diffuse reflections between surface and scattering layer ($\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)$).

Summing up all individual radiance components I_{CD_i} results in the following

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	18
---------------------	---	---	-----------

geometric series:

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

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 α 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_i} corresponds to the directly transmitted radiance from the surface through the lower atmosphere ($T_I^g(\tau_\downarrow, \zeta)$), the scattering layer ($T_I^s(\tau_e, \zeta)$), and the upper atmosphere after i diffuse reflections between surface and scattering layer ($\frac{\alpha}{2} S_F(\tau_s, \tau_e, \tau_\downarrow) T_{F_{lam}}^g(\tau_g) T_{F_{iso}}^g(\tau_g)$).

Summing up all individual radiance components I_{SI_i} results in the following geometric series:

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

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 α and produces a Lambertian

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	19
---------------------	---	---	-----------

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

Summing up all individual radiance components I_{Cl_i} results in the following geometric series:

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

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 ($T_I^g(\tau_\downarrow + \tau_\uparrow, \zeta)$) and the scattering layer ($T_I^s(\tau_e, \zeta)$) but ignoring multiple scattering because of the weak signal.

$$I_{SIF} = \frac{F_{SIF}^0}{\pi} T_I^s(\tau_e, \zeta) T_I^g(\tau_\downarrow + \tau_\uparrow, \zeta) \quad (25)$$

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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	20
---------------------	---	---	-----------

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

$$S_I(\tau_s, \zeta) \approx \zeta \tau_s \text{ and} \quad (26)$$

$$T_I^s(\tau_s, \zeta) \approx 1 - \zeta \tau_s. \quad (27)$$

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

$$S_F(\tau_s, \tau_e, \tau_\downarrow) \approx \frac{E_2(\tau_\downarrow)}{E_3(\tau_\downarrow)} \tau_s. \quad (28)$$

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

$$\begin{aligned} I \approx & \frac{F_0}{\pi \zeta_0} T_I^g(\tau_\uparrow, \zeta_0 + \zeta) \left[\frac{1}{4} \tau_s \zeta_0 \zeta + \right. \\ & \alpha \left(T_I^g(\tau_\downarrow, \zeta_0 + \zeta) \left[1 + \tau_s (\alpha E_2^2 - \zeta_0 - \zeta) \right] + \right. \\ & \left. \frac{1}{2} \tau_s E_2 \left[T_I^g(\tau_\downarrow, \zeta_0) \zeta + T_I^g(\tau_\downarrow, \zeta) \zeta_0 \right] \right] + \\ & \left. \frac{F_{SIF}^0}{\pi} T_I^g(\tau_\downarrow + \tau_\uparrow, \zeta) [1 - \tau_s \zeta] \right]. \end{aligned} \quad (29)$$

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 .

$$\theta(z) = \arcsin\left(\frac{r_e}{r_e + z} \sin \theta\right), \quad (30)$$

with r_e being the Earth's radius and θ the (solar or satellite) zenith angle at the surface.

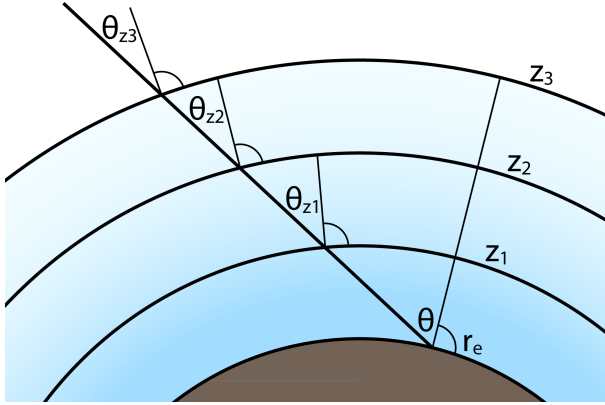


Figure 2: Spherical geometry of the Earth's atmosphere with the Earth's radius r_e , the (solar or satellite) zenith angle θ at the surface and at the heights $z_{1,2,3,\dots}$.

Correspondingly, also the light path extensions ζ and ζ_0 become height dependent. In the following, θ , θ_0 , ζ , and ζ_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 θ^s , θ_0^s , ζ^s , and ζ_0^s shall refer to the scattering layer. This has implications for Eq. 2 which now becomes

$$T_l^g(K(z), \zeta(z)) = e^{-\int K(z) \zeta(z) dz}. \quad (31)$$

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

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	22
---------------------	---	---	-----------

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

$$\chi^2 = \frac{1}{m+n} [(\vec{y} - \vec{F}(\vec{x}, \vec{b}))^T \mathbf{S}_\epsilon^{-1} (\vec{y} - \vec{F}(\vec{x}, \vec{b})) + (\vec{x} - \vec{x}_a)^T \mathbf{S}_a^{-1} (\vec{x} - \vec{x}_a)]. \quad (32)$$

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:

$$\vec{x}_{i+1} = \vec{x}_i + \hat{\mathbf{S}}_i [\mathbf{K}_i^T \mathbf{S}_\epsilon^{-1} (\vec{y} - \vec{F}(\vec{x}_i, \vec{b})) - \mathbf{S}_a^{-1} (\vec{x}_i - \vec{x}_a)] \quad (33)$$

$$\hat{\mathbf{S}}_i = (\mathbf{K}_i^T \mathbf{S}_\epsilon^{-1} \mathbf{K}_i + (1 + \gamma) \mathbf{S}_a^{-1})^{-1}. \quad (34)$$

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), O₂ (~ 757.65 – 772.56 nm), wCO₂ (~ 1595.0 – 1620.6 nm), and sCO₂ (~ 2047.3 – 2080.9 nm). The separate SIF fit window ensures that the SIF information solely comes from free Fraunhofer lines rather than from O₂ 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 \mathbf{S}_ϵ

Strictly speaking, the measurement error covariance matrix does not only quantify the measurement errors and their correlations; it, additionally, accounts

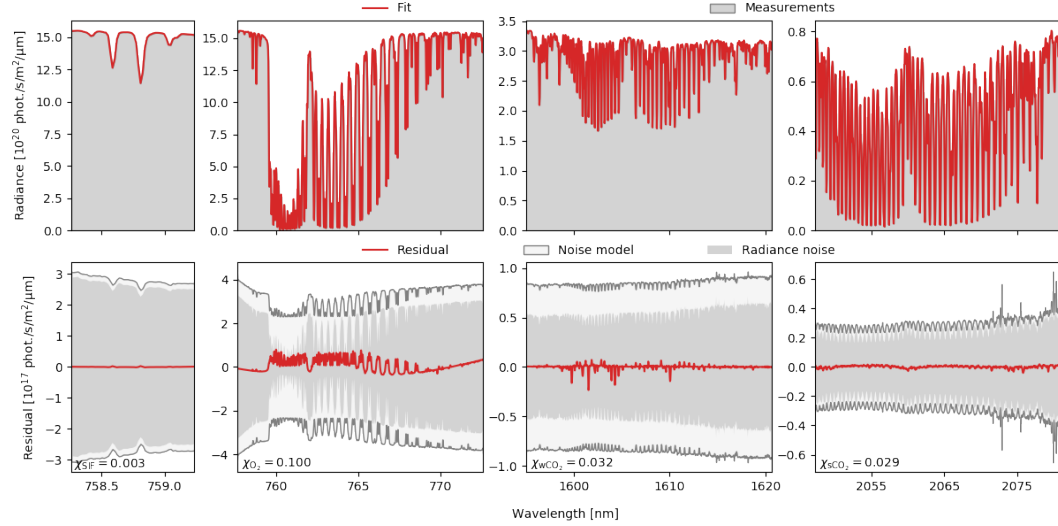


Figure 3: SCIATRAN simulated OCO-2 measurement fitted with FOCAL. Geophysical *baseline* scenario and *3-Scat* retrieval setup, $\theta_0 = 40^\circ$, parallel polarization (see Reuter et al. (2017c) for definitions of geophysical scenarios and retrieval setups). **Top:** Simulated and fitted radiance 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 $\chi_j = (\frac{1}{m_j} \vec{\epsilon}_j^T \mathbf{S}_{\epsilon_j}^{-1} \vec{\epsilon}_j)^{1/2}$.

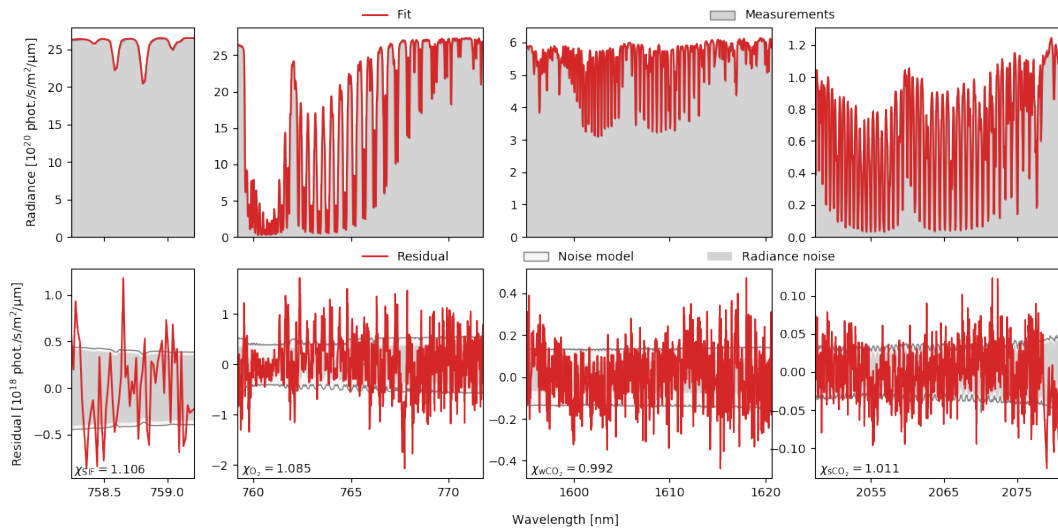


Figure 4: Same as Fig. 3 but with an actual (not simulated) OCO-2 measurement of June 5, 2015, 12:01 UTC (sounding ID: 2015060512011938).

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	24
---------------------	---	---	-----------

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 \mathbf{S}_ϵ 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.

$$\lambda' = \lambda + \lambda_{sh} + \lambda_n \lambda_{sq} \quad (35)$$

$$\lambda_n = 2 - 4 \frac{\lambda_1 - \lambda}{\lambda_1 - \lambda_0} \quad (36)$$

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

The squeezing of the ILS is done by:

$$\lambda'_{ILS} = \lambda_{ILS} ILS_{sq} \quad (37)$$

Here λ'_{ILS} is the modified ILS wavelength computed from the nominal ILS λ_{ILS} wavelength and the squeeze parameter ILS_{sq} .

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	25
---------------------	---	---	-----------

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_2O) 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_2O layer spans over four atmospheric layers used for the discretized RT calculations. The CO_2 and H_2O concentrations are homogeneous within each of the five layers. As also done by Noël et al. (2021), we further improve the H_2O fit quality by allowing for variations of the H_2O isotopologue HDO by fitting δD defined as

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

Here R_{meas} is the ratio of the measured HDO and H_2O columns, and R_{VSMOW} (3.1152×10^{-4}) is the corresponding value for Vienna Standard Mean Ocean Water (VSMOW). δD is usually given in units of per mill.

XCO₂ and XH₂O are not part of the state vector but are calculated during the postprocessing 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 τ_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 ILS_{sq} .

We estimate the first guess zeroth order albedo polynomial coefficients αP_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_2O 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.

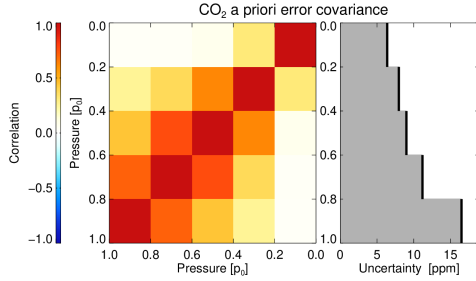


Figure 5: CO₂ a priori error covariance computed from randomly chosen SLIMCO2 profiles and corresponding CarbonTracker profiles. The CO₂ layer variances have been up-scaled so that the a priori XCO₂ uncertainty becomes 7.5 ppm. **Left:** Layer-to-layer correlation matrix of the a priori uncertainty. **Right:** 1 σ a priori uncertainty.

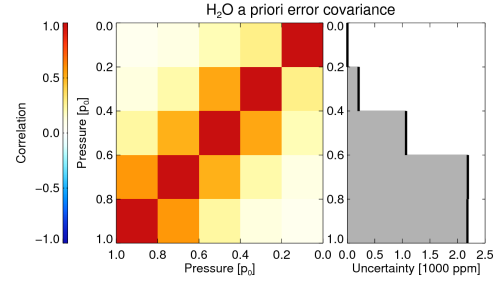


Figure 6: As Fig. 5 but for H₂O and estimated from day-to-day variations of ECMWF analysis profiles (without variance scaling as done for CO₂).

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

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

4.5 A priori error covariance matrix \mathbf{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 CO₂ and H₂O profile layers, we assume \mathbf{S}_a to be diagonal. As described by Reuter et al. (2012), we compute the CO₂ layer-to-layer covariances by comparing randomly chosen SLIMCO2 profiles with corresponding CarbonTracker profiles. The CO₂ layer variances have been up-scaled so that the a priori XCO₂ uncertainty becomes 7.5 ppm. This ensures retrievals to be dominated by the measurement but not the a priori. We estimated the H₂O layer-to-layer covariances by analyzing H₂O day-to-day variations of ECMWF analysis profiles. CO₂ and H₂O a priori error covariances are shown in Fig. 5 and Fig. 6. All other (diagonal) elements of \mathbf{S}_a are listed in row $\sigma\vec{x}_a$ 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 \bar{x}_a and first guess \bar{x}_0 value, and the a priori uncertainty $\sigma\bar{x}_a$.

State vector element	Fit window sensitivity				\bar{x}_0	\bar{x}_a	$\sigma\bar{x}_a$
	SIF	O ₂	wCO ₂	sCO ₂			
αP_0^{SIF}	•				R_0^{SIF}	R_0^{SIF}	0.1
αP_1^{SIF}	•				0.0	0.0	0.01
$\alpha P_0^{\text{O}_2}$		•			$R_0^{\text{O}_2}$	$R_0^{\text{O}_2}$	0.1
$\alpha P_1^{\text{O}_2}$		•			0.0	0.0	0.01
$\alpha P_2^{\text{O}_2}$		•			0.0	0.0	0.01
$\alpha P_0^{\text{wCO}_2}$			•		$R_0^{\text{wCO}_2}$	$R_0^{\text{wCO}_2}$	0.1
$\alpha P_1^{\text{wCO}_2}$			•		0.0	0.0	0.01
$\alpha P_2^{\text{wCO}_2}$			•		0.0	0.0	0.01
$\alpha P_0^{\text{sCO}_2}$				•	$R_0^{\text{sCO}_2}$	$R_0^{\text{sCO}_2}$	0.1
$\alpha P_1^{\text{sCO}_2}$				•	0.0	0.0	0.01
$\alpha P_2^{\text{sCO}_2}$				•	0.0	0.0	0.01
$\lambda_{\text{sh}}^{\text{SIF}}$ [nm]	•				0.0	0.0	0.01
$\lambda_{\text{sq}}^{\text{SIF}}$ [nm]	•				0.0	0.0	0.01
$\lambda_{\text{sh}}^{\text{O}_2}$ [nm]		•			0.0	0.0	0.01
$\lambda_{\text{sq}}^{\text{O}_2}$ [nm]		•			0.0	0.0	0.01
$\text{ILS}_{\text{sq}}^{\text{O}_2}$		•			1.0	1.0	0.01
$\lambda_{\text{sh}}^{\text{wCO}_2}$ [nm]			•		0.0	0.0	0.01
$\lambda_{\text{sq}}^{\text{wCO}_2}$ [nm]			•		0.0	0.0	0.01
$\text{ILS}_{\text{sq}}^{\text{wCO}_2}$			•		1.0	1.0	0.01
$\lambda_{\text{sh}}^{\text{sCO}_2}$ [nm]				•	0.0	0.0	0.01
$\lambda_{\text{sq}}^{\text{sCO}_2}$ [nm]				•	0.0	0.0	0.01
$\text{ILS}_{\text{sq}}^{\text{sCO}_2}$				•	1.0	1.0	0.01
SIF [mW/m ² /sr/nm]	•				0.0	0.0	10.0
p_s [p_0]	•	•	•	•	0.2	0.2	1.0
τ_s	•	•	•	•	0.01	0.01	0.1
\bar{A}	•	•	•	•	4.0	4.0	2.0
H ₂ O L ₀ [ppm]			•	•	ECMWF	ECMWF	2179.9
H ₂ O L ₁ [ppm]			•	•	ECMWF	ECMWF	2186.9
H ₂ O L ₂ [ppm]			•	•	ECMWF	ECMWF	1066.0
H ₂ O L ₃ [ppm]			•	•	ECMWF	ECMWF	205.4
H ₂ O L ₄ [ppm]			•	•	ECMWF	ECMWF	2.67
δD , [‰]			•		0.0	0.0	1000.0
CO ₂ L ₀ [ppm]			•	•	SLIMCO2	SLIMCO2	16.50
CO ₂ L ₁ [ppm]			•	•	SLIMCO2	SLIMCO2	11.19
CO ₂ L ₂ [ppm]			•	•	SLIMCO2	SLIMCO2	8.00
CO ₂ L ₃ [ppm]			•	•	SLIMCO2	SLIMCO2	7.97
CO ₂ L ₄ [ppm]			•	•	SLIMCO2	SLIMCO2	6.39
XH ₂ O [ppm]					ECMWF	ECMWF	898.2
XCO ₂ [ppm]					SLIMCO2	SLIMCO2	7.5

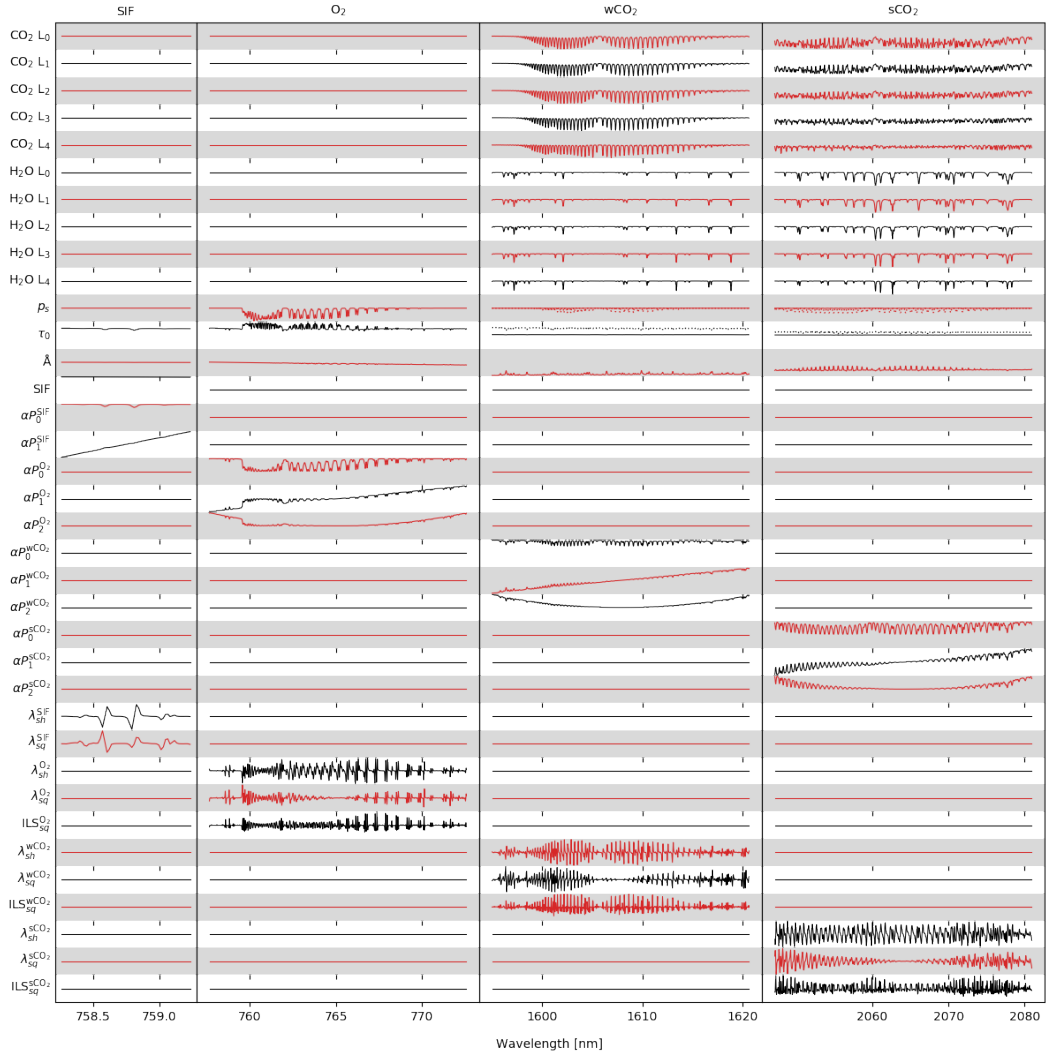


Figure 7: Jacobian matrix computed with FOCAL for the geophysical *Rayleigh* scenario and the *3-Scat* retrieval setup of Reuter et al. (2017c). Within the CO₂ fit windows, an additional dashed line shows the partial derivatives according to τ_s and p_s scaled by a factor of 10 and 20, respectively.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	29
---------------------	---	---	-----------

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 θ_0 and θ), 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 $1.0 \text{ mW/m}^2/\text{sr/nm}$ 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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	30
---------------------	---	---	-----------

4.8 A posteriori error covariance matrix $\hat{\mathbf{S}}$

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 γ

We implemented a Levenberg-Marquardt minimization method making use of a damping parameter γ (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 ($\chi_{i+1}^2 \geq \chi_i^2$) are rejected and γ is increased. Only iterations that actually improve the cost function ($\chi_{i+1}^2 < \chi_i^2$) are accepted. In these cases γ 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:

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

Additionally, we test if χ^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. All results and figures shown in this section correspond to FOCAL v10.1. However, as the preprocessing steps have not been substantially changed, but only adapted to the updated input data, they are also suitable to illustrate the preprocessing of FOCAL v11.

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 hours, $0.25^\circ \times 0.25^\circ$, 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 XCO₂ 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 \neq 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 $\pm 80^\circ$, or extreme surface roughnesses (standard deviation of the surface elevation) greater than 1000 m. In Fig. 9, this filter is referred to as LAT/SUZ/SAZ/ σ ALT.

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 $1\text{ km} \times 1\text{ km}$ (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 50km 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 50km 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

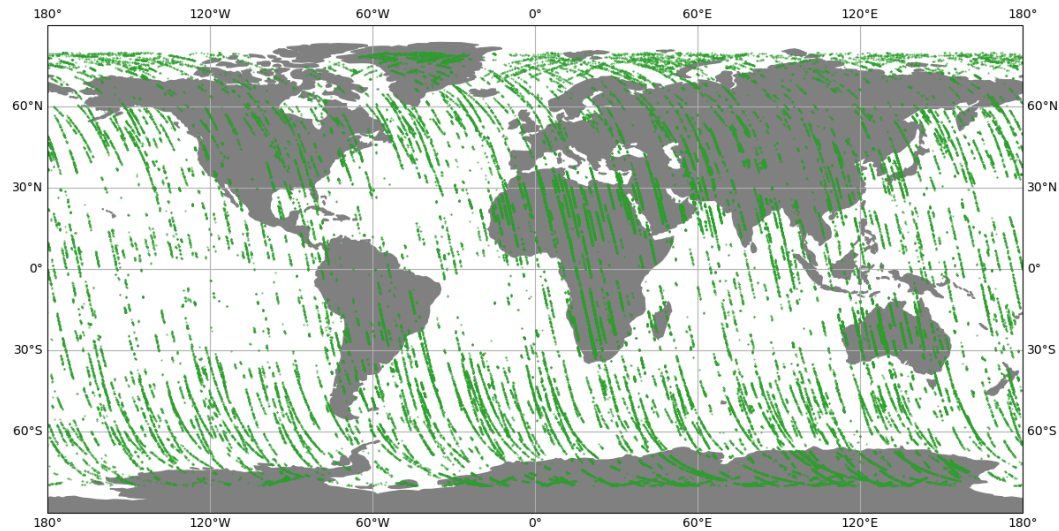


Figure 8: Sampling of all 250000 OCO-2 soundings with certain MODIS cloud classification used as training data set. The soundings have been randomly drawn from 24 representative days in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05., 23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 15.09., 16.09., 15.10., 16.10., 15.11., 17.11., 12.12., 18.12.)

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 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. 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 through-

Table 3: 25 most important features of the cloud detection random forest classifier identified during training and their relative importance.

Feature	Relative importance
Surface elevation	0.027574
Continuum radiance strong CO ₂ / O ₂ band	0.020674
Continuum radiance weak CO ₂ / O ₂ band	0.016991
Continuum radiance O ₂ / weak CO ₂ band	0.015964
Continuum radiance O ₂ band	0.014231
Continuum radiance O ₂ / strong CO ₂ band	0.012380
O ₂ band radiance #184	0.010487
Solar zenith angle	0.009849
O ₂ band radiance #161	0.009106
Continuum radiance weak CO ₂ / strong CO ₂ band	0.008873
O ₂ band radiance #162	0.008716
O ₂ band radiance #174	0.008509
O ₂ band radiance #164	0.008376
O ₂ band radiance #128	0.007545
Continuum radiance strong CO ₂ / weak CO ₂ band	0.007360
O ₂ band radiance #163	0.007193
O ₂ band radiance #176	0.006377
O ₂ band radiance #172	0.006282
O ₂ band radiance #123	0.006048
O ₂ band 8-binned radiance #23	0.005989
O ₂ band radiance #151	0.005683
Minimum radiance strong CO ₂ band	0.005628
O ₂ band radiance #126	0.005221
O ₂ band 8-binned radiance #16	0.005094
O ₂ band 8-binned radiance #20	0.005045

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.

		True		Total
		Clear	Cloudy	
Predicted	Clear	88877	1577	90454
	Cloudy	127	409419	409546
Total		89004	410996	500000

put. 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 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_2O and CO_2 cross-sections so that the resulting XCO_2 and XH_2O 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 XCO_2 and XH_2O values. We found that the average retrieved XCO_2 had to be divided by a factor of 0.9965 to match the a priori XCO_2 . The average retrieved XH_2O had to be divided by a factor of 1.0043 to match the a priori XH_2O . Therefore, we decided to scale the XCO_2 and XH_2O 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.

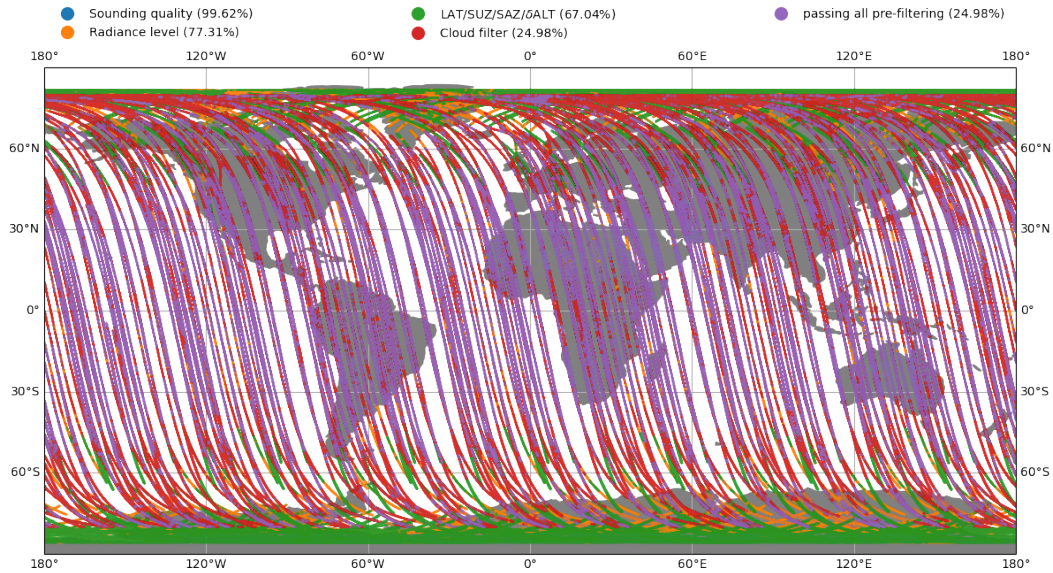


Figure 9: Pre-filtering statistics of the 24 days data subset used for the noise model analysis (Sec. 5.4). The filters are applied in the order: Sounding quality, radiance level, LAT/SUZ/SAZ/ σ ALT, cloud filter. The colors represent filter activity and soundings passing all filters are shown in violet. Numbers in brackets represent cumulative filter throughputs.

Especially, when analyzing measured data, unknown inaccuracies of the forward model can violate this assumption and lead to larger fit residuals and 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.,

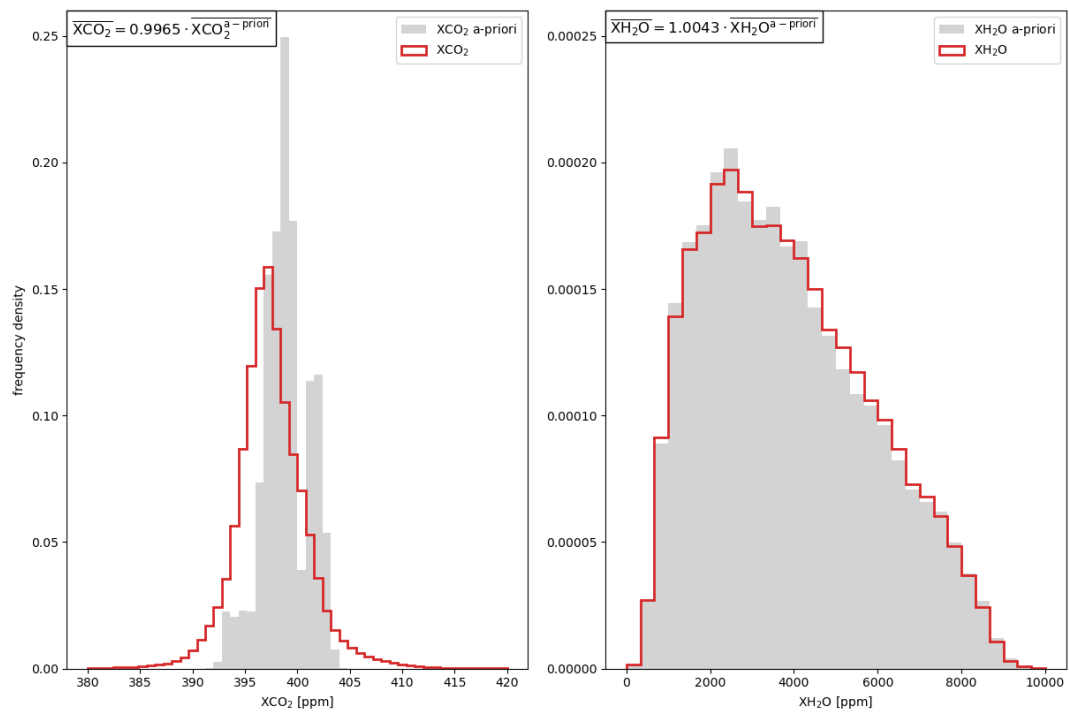


Figure 10: Comparison of the retrieved (**red**) and a priori (**gray**) XCO₂ and XH₂O values before scaling the line intensities of the cross-section data base.

albedo and solar zenith angle. Therefore, it is reasonable to assume forward model errors to be approximately proportional to the continuum signal I_{cont} which 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 δF the relative forward model error.

In order to estimate the free parameter δF , 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 χ^2 -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 χ^2 -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

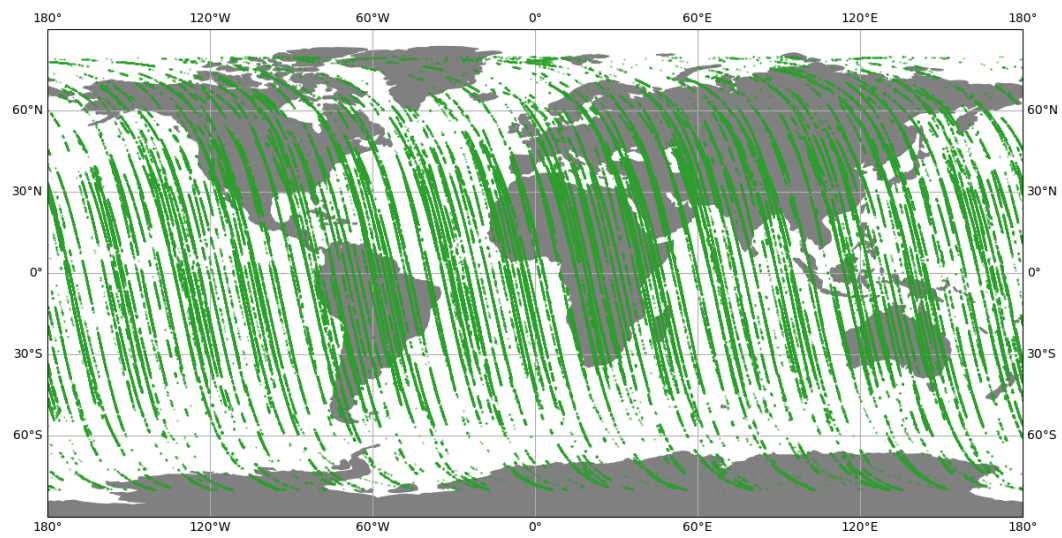


Figure 11: Sampling of all pre-filtered soundings analyzed in order to determine the noise model. The data set consists of 10% of all pre-filtered OCO-2 soundings (randomly selected) of 24 days in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05., 23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 15.09., 16.09., 15.10., 16.10., 15.11., 17.11., 12.12., 18.12.). This results in a manageable but still representative data set with respect to nadir/glint observation geometry, season, and spatial distribution.

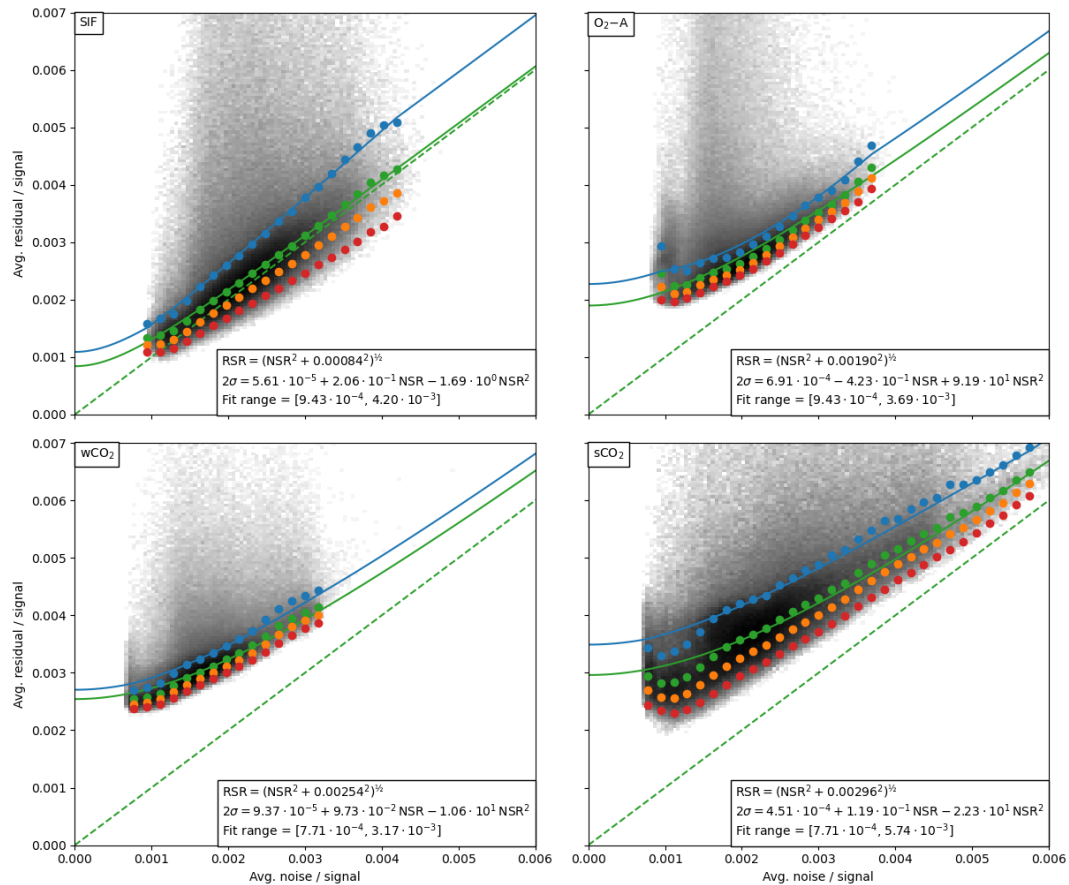


Figure 12: Root mean square noise to signal ratio NSR versus root mean square residual to signal ratio RSR for all four fit windows. **red points:** 2.28th percentile within bins with more than 500 samples (35 bins in total). **orange points:** 15.9th percentile. **green points:** expectation value estimated 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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	41
---------------------	---	---	-----------

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

$$2\sigma = a_0 + a_1 NSR + a_2 NSR^2 \quad (41)$$

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:

$$N' = \sqrt{N^2 + I_{\text{cont}}^2 \delta F^2}. \quad (42)$$

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 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/dataproduct/chlor_a.php, Hu et al., 2012) over ocean and normalized difference vegetation index (NDVI) data over land (obtained from <https://modis.gsfc.nasa.gov/data/dataproduct/mod13.php>).

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, $\chi^2 > 2$, or RSR exceeding the threshold of the noise model (Sec. 5.4)) have been removed beforehand. In the following, we use the fitted linear relationship as ZLO correction for these three fit windows.

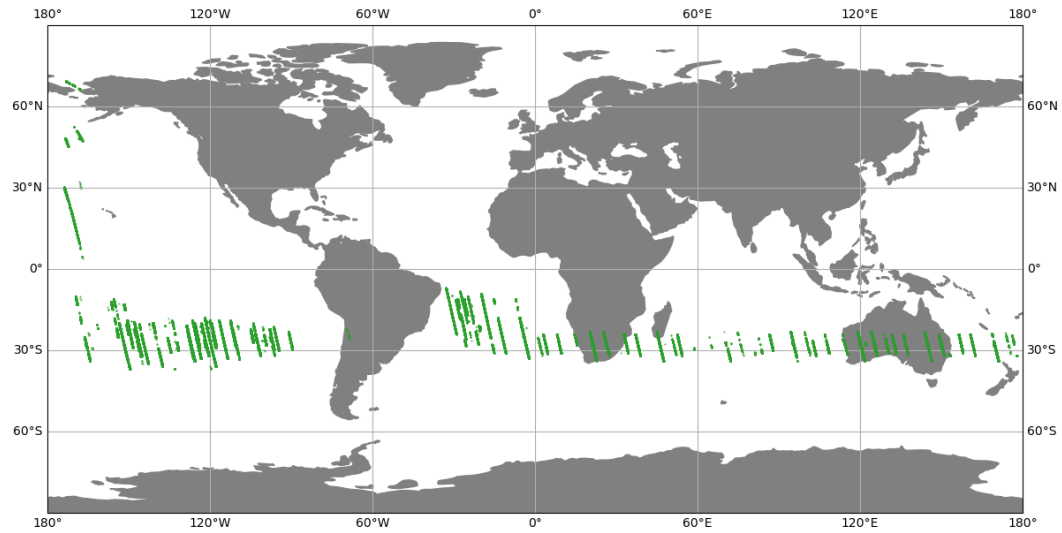


Figure 13: Sampling of all pre-filtered soundings analyzed in order to determine the ZLO correction. The data set consists of all pre-filtered OCO-2 soundings of 24 days in 2015 (13.01., 15.01., 14.02., 16.02., 10.03., 20.03., 03.04., 19.04., 08.05., 23.05., 08.06., 24.06., 15.07., 16.07., 15.08., 16.08., 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).

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.

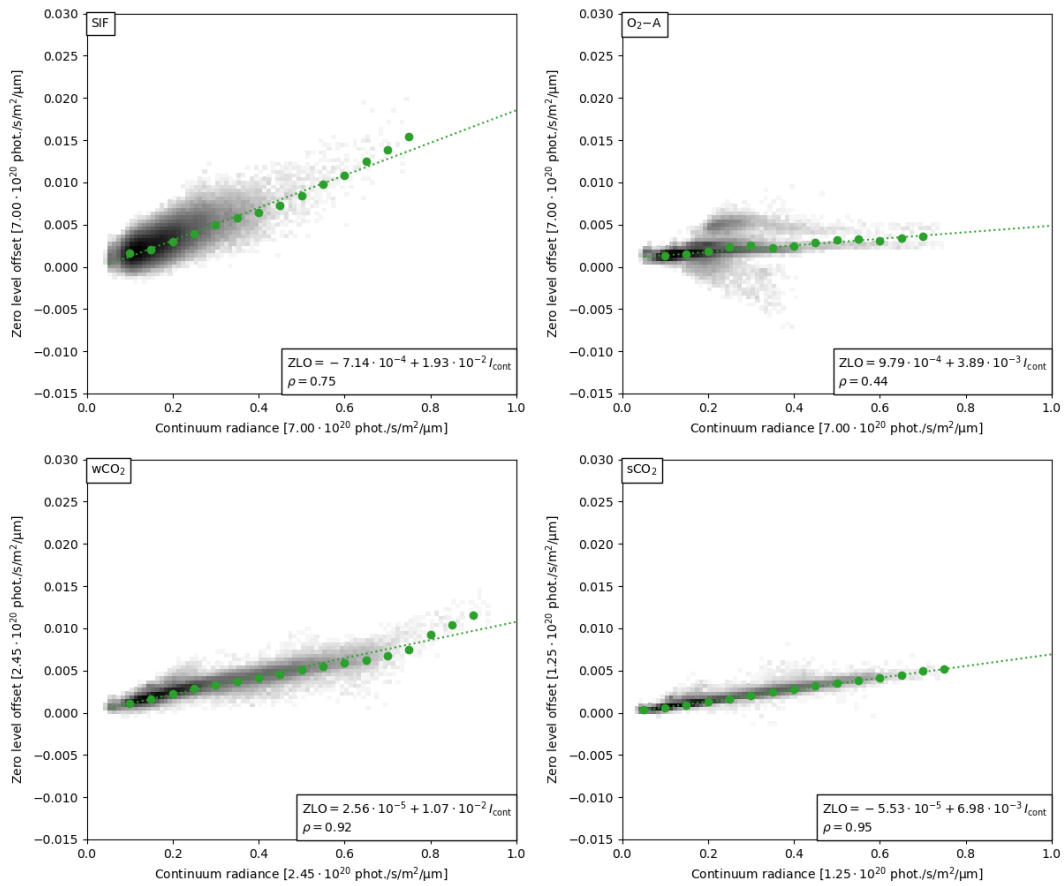


Figure 14: Retrieved zero level offset (ZLO) versus continuum radiance (I_{cont}) for all four fit windows. Soundings without convergence, $\chi^2 > 2$, or RSR exceeding the threshold of the noise model (Sec. 5.4)) have been removed beforehand. **green dots:** Binned averages. **green line:** Linear 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. All results and figures shown in this section correspond to FOCAL v10.1. However, as the postprocessing steps have not been substantially changed, but only adapted to the updated input data, they are also suitable to illustrate the postprocessing of FOCAL v11.

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 χ^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 a unexpectedly large influence on the retrieved local XCO₂ 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 XCO₂ outliers increase the local retrieved XCO₂ variability and are likely correlated with extreme values of some candidate parameters such as the non CO₂ and H₂O state vector elements or the continuum radiance in one of the fit windows ($I_{\text{cont}}^{\text{O}_2}$, $I_{\text{cont}}^{\text{wCO}_2}$, $I_{\text{cont}}^{\text{sCO}_2}$, see also Reuter et al., 2017b).

For a representative two months data set (April and August 2015, Fig. 15), we estimated the local retrieved XCO₂ variability $\text{VAR}(\Delta\text{XCO}_2)$ as follows: For each sounding, we computed the difference ΔXCO_2 between XCO₂ and its $5^\circ \times 5^\circ$ daily median and subsequently, we computed the variance of all ΔXCO_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

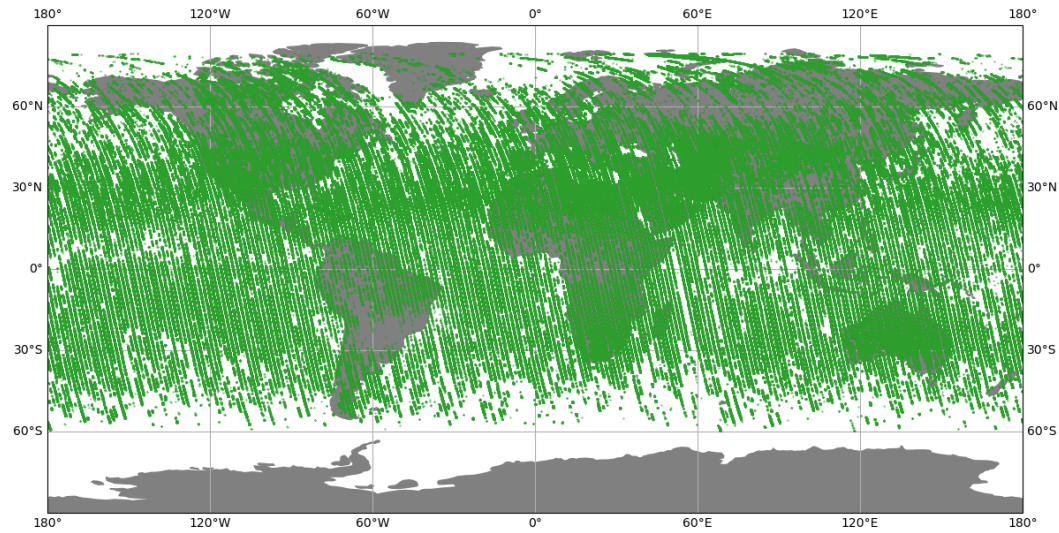


Figure 15: Sampling of all soundings of April and August 2015 used to compute the post-filtering parameters (Tab. 5).

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 the computed difference ΔXCO_2 with the number of soundings N falling in the corresponding bin, i.e., we replace the variance $VAR(\Delta XCO_2)$ by the weighted variance $wVAR(\Delta XCO_2) = VAR(N \Delta XCO_2)$. 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 ΔXCO_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).

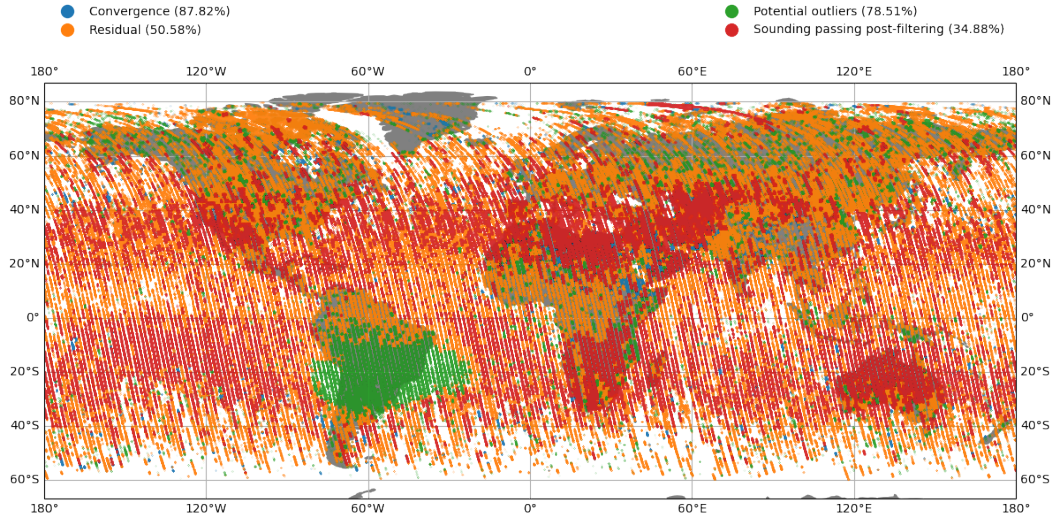


Figure 16: Post-filtering statistics for April and August 2015. The filters are applied in the order: convergence, residual, and potential outliers (see main text for a description). The colors represent filter activity and soundings passing all filters are shown in red. Numbers in brackets represent filter throughputs.

Above sea (Fig. 17, right), this filter reduces the variance of ΔXCO_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 $\alpha P_1^{sCO_2}$, contributing 53% to the variance reduction. All parameter thresholds found for 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

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 ΔXCO_2 is reduced from about 3.9 ppm² to 2.2 ppm² above land and reduced from about 2.8 ppm² to 1.2 ppm² above sea. See Reuter et al. (2017b) and the main text for a description of the individual parameters.

	Parameter	Lower threshold	Upper threshold	Variance reduction [ppm ²]
Land	\tilde{A}	1.1373	5.7167	$9.3343 \cdot 10^{-1}$
	$ILS_{sq}^{wCO_2}$	$9.9368 \cdot 10^{-1}$	1.0011	$3.7107 \cdot 10^{-1}$
	$ILS_{sq}^{O_2}$	1.0039	1.0219	$9.6385 \cdot 10^{-2}$
	$ILS_{sq}^{sCO_2}$	$9.9305 \cdot 10^{-1}$	1.0064	$8.8602 \cdot 10^{-2}$
	$\alpha P_2^{O_2}$	$-1.5433 \cdot 10^{-3}$	$1.6265 \cdot 10^{-4}$	$7.8222 \cdot 10^{-2}$
	τ_0	-	$1.4814 \cdot 10^{-1}$	$5.8100 \cdot 10^{-2}$
	αP_1^{SIF}	$-4.4338 \cdot 10^{-3}$	$6.2912 \cdot 10^{-3}$	$3.4975 \cdot 10^{-2}$
	$\alpha P_2^{wCO_2}$	$-2.2400 \cdot 10^{-3}$	-	$2.7516 \cdot 10^{-2}$
	$\alpha P_2^{sCO_2}$	$-2.4493 \cdot 10^{-3}$	$6.1081 \cdot 10^{-4}$	$2.6524 \cdot 10^{-2}$
	BG^{O_2}	-	$5.7684 \cdot 10^{-1}$	$1.7935 \cdot 10^{-2}$
Sea	$\alpha P_1^{sCO_2}$	-	$1.6189 \cdot 10^{-5}$	$7.5229 \cdot 10^{-1}$
	τ_0	-	$3.6108 \cdot 10^{-2}$	$2.8847 \cdot 10^{-1}$
	p_s	-	$3.6004 \cdot 10^{-1}$	$1.6624 \cdot 10^{-1}$
	$ILS_{sq}^{O_2}$	-	1.0150	$1.0680 \cdot 10^{-1}$
	$ILS_{sq}^{wCO_2}$	$9.9355 \cdot 10^{-1}$	1.0035	$7.9369 \cdot 10^{-2}$
	\tilde{A}	1.1949	-	$6.5081 \cdot 10^{-2}$
	αP_1^{SIF}	$-2.8179 \cdot 10^{-3}$	$3.0376 \cdot 10^{-3}$	$5.3789 \cdot 10^{-2}$
	$\alpha P_2^{sCO_2}$	$-1.3674 \cdot 10^{-3}$	$-3.6583 \cdot 10^{-5}$	$5.0012 \cdot 10^{-2}$
	$ILS_{sq}^{sCO_2}$	$9.9205 \cdot 10^{-1}$	-	$3.6732 \cdot 10^{-2}$
	$\lambda_{sh}^{sCO_2}$	$-7.6605 \cdot 10^{-4}$	-	$3.6152 \cdot 10^{-2}$

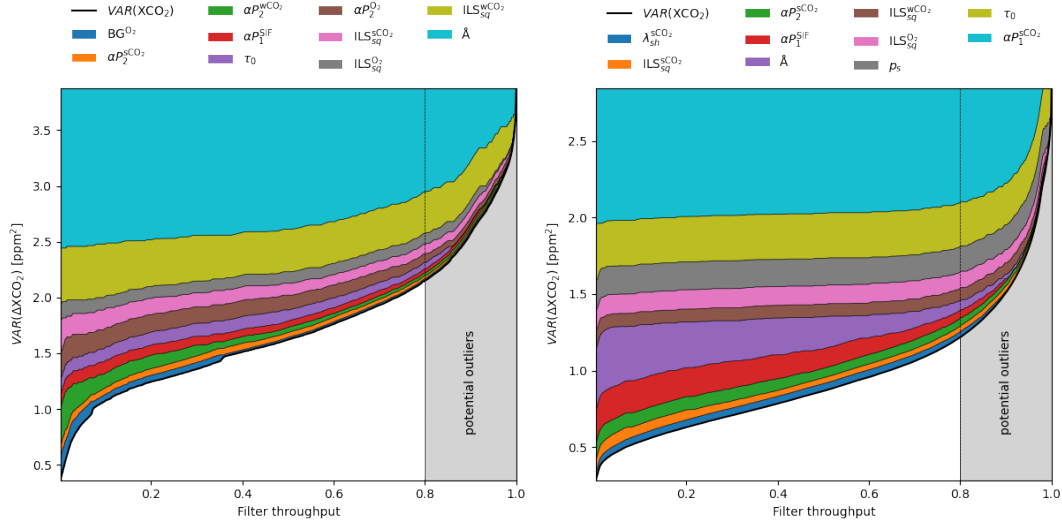


Figure 17: Variance versus filter throughput for the 10 most promising parameters identified for the potential outliers filter. The colors represent the prorated variance reduction of the individual parameters. See Reuter et al. (2017b) and the main text for a description of the individual parameters. **left:** Land. **right:** Sea.

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 meeting all post-filtering criteria and having co-locations with the *true 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

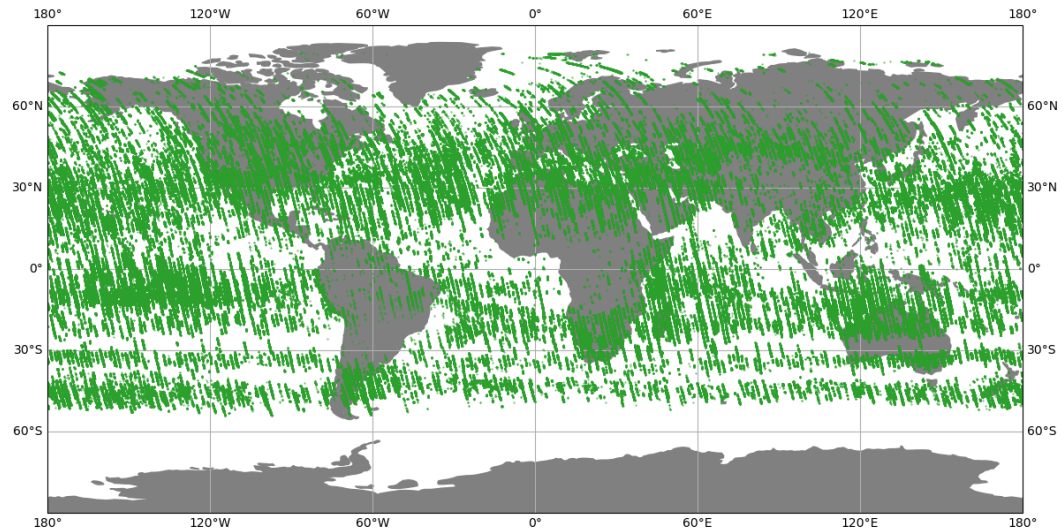


Figure 18: Sampling of all soundings within the time period 2014–2019 used to train the random forest bias correction.

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. 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 *true db* is 1.848 ppm for the training data set. Fig. 19 illustrates that the largest XCO₂ difference to the *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 *true db* becomes 1.381 ppm for the training data set. The corresponding RMS values for both test data sets are basically identical.

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.

Feature	Relative importance
Footprint ID	$3.4359 \cdot 10^{-1}$
p_s	$2.6965 \cdot 10^{-1}$
Land/sea fraction	$1.5085 \cdot 10^{-1}$
$\sigma_{\text{XH}_2\text{O}}$	$5.5431 \cdot 10^{-2}$
σ_{XCO_2}	$4.9654 \cdot 10^{-2}$
θ	$3.5094 \cdot 10^{-2}$
$\alpha P_2^{\text{sCO}_2}$	$1.9015 \cdot 10^{-2}$
τ_0	$1.8744 \cdot 10^{-2}$
$\text{ILS}_{sq}^{\text{O}_2}$	$1.6530 \cdot 10^{-2}$
θ_0	$1.6076 \cdot 10^{-2}$
\AA	$6.7967 \cdot 10^{-3}$
αP_0^{SIF}	$5.7567 \cdot 10^{-4}$
Surface elevation	$5.4877 \cdot 10^{-4}$
BG^{wCO_2}	$4.9943 \cdot 10^{-4}$
$\alpha P_0^{\text{O}_2}$	$2.8850 \cdot 10^{-4}$
$\alpha P_1^{\text{O}_2}$	$2.6743 \cdot 10^{-4}$
BG^{sCO_2}	$2.5795 \cdot 10^{-4}$
$\alpha P_1^{\text{sCO}_2}$	$8.9723 \cdot 10^{-5}$
$\text{ILS}_{sq}^{\text{sCO}_2}$	$5.8538 \cdot 10^{-5}$
$\alpha P_2^{\text{wCO}_2}$	$2.6632 \cdot 10^{-5}$
BG^{O_2}	$2.5997 \cdot 10^{-5}$
SIF	$1.2224 \cdot 10^{-5}$
$\alpha P_0^{\text{wCO}_2}$	$9.8717 \cdot 10^{-6}$
$\lambda_{\text{O}_2}^{\text{sh}}$	$6.2205 \cdot 10^{-6}$
$\lambda_{sq}^{\text{sCO}_2}$	$1.7541 \cdot 10^{-6}$

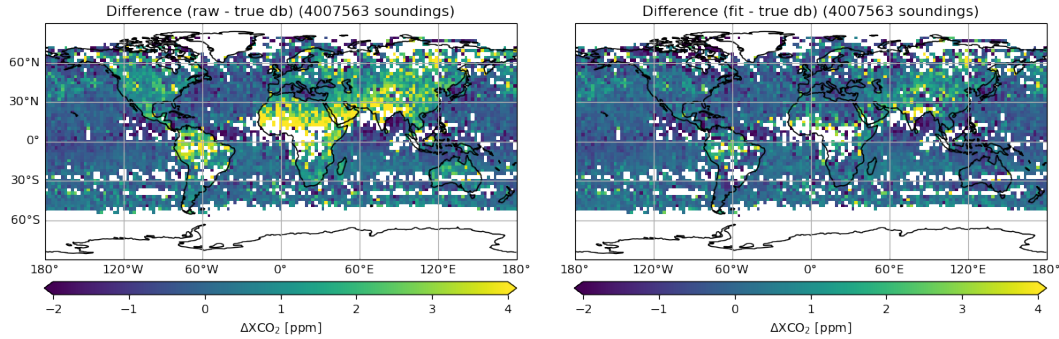


Figure 19: **Left:** Difference of the non-bias corrected retrieved post filtered (raw) XCO₂ and the *true db*. **Right:** Difference of the bias corrected retrieved post filtered (fit) XCO₂ and the *true db*.

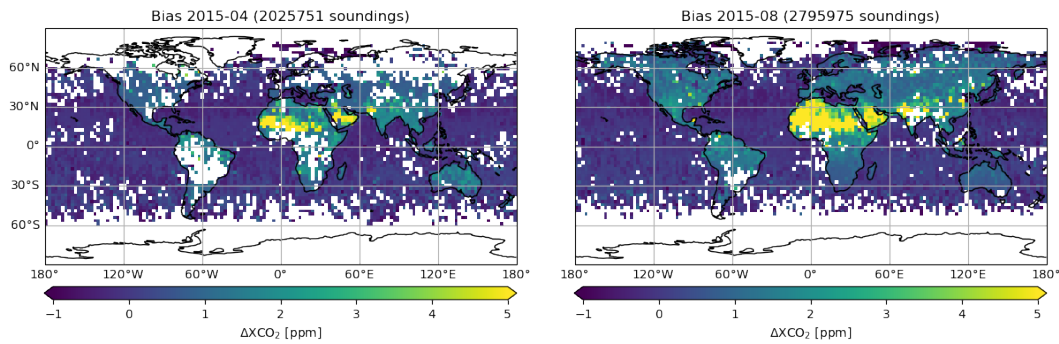


Figure 20: FOCAL's bias pattern predicted by the random forest regressor at the example of April 2015 (**left**) and August 2015 (**right**).

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 *true db* (Fig. 19, left). Because of less noise in this figure, one can also recognize a land/sea bias in addition to the 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.

6.3 Uncertainty estimates

Especially for the application of flux inversion, reliable information on the uncertainty of each individual sounding is necessary. For this purpose, we analyzed the validation dataset of co-located FOCAL and TCCON measurements

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	52
---------------------	---	---	-----------

described in the end-to-end ECV uncertainty budget document (E3UBv5, Reuter et al., 2024b). Specifically, we compared FOCAL's XCO₂ a posteriori uncertainty reported by the optimal estimation retrieval σ_{XCO_2} (Sec. refsec:math:l2:sh) with the actual stochastic error of the FOCAL measurements (that component of the difference between FOCAL and TCCON which cannot be explained by systematic errors). For this purpose we pooled the entire data set of all co-locations into 20 bins with increasing reported uncertainty in a way that each bin included the same number of co-locations. For each bin, we computed the (quadratic) average reported uncertainty which we compared with the standard deviation of the actual stochastic error, i.e., the actual precision. We found that both quantities are connected by a fairly linear relationship which we use to empirically correct the reported uncertainty.

$$\sigma_{\text{XCO}_2'} = 1.015 \sigma_{\text{XCO}_2} + 0.862 \text{ ppm} \quad (43)$$

The equation shows that FOCAL's originally diagnosed uncertainty σ_{XCO_2} is somewhat to optimistic. However it shall be noted that the analyzed difference to TCCON does not only include instrumental noise but also pseudo noise from representation errors. It shall also be notated, that the reported uncertainty is mainly driven by the instrumental noise which is in turn driven by the radiance so that the darkest scenes usually have the largest reported uncertainties. This means, especially the bins including the largest (or smallest) reported uncertainties may be dominated by an individual validation site with especially dark (or bright) albedo, while the other bins usually consist of data from a larger mixture of TCCON sites.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	53
---------------------	---	---	-----------

7 Version History

v11

- Extension of the data set now covering 09/2014-12/2023.
- Update of the OCO-2 L1b radiance data to v11.
- Update of the used digital elevation model to Copernicus GLO-30.
- Update of the CO₂, O₂, and H₂O cross section tables to ABSCO v5.2.
- Update of the solar irradiance spectrum to the ISS solar reference spectrum.
- Adaptation of the preprocessor to the new L1b input data including re-training of the cloud filter.
- Adaptation of the postprocessor including re-training of the bias correction.
- Minor bug fixes.
- Minor efficiency improvements.

v10.1

- Extension of the data set now covering 09/2014-02/2022.
- Improvement of bias correction.
- Improvement of uncertainty estimates.

v10

- Relaxation of the latitude preprocessor filter from $\pm 70^\circ$ to $\pm 80^\circ$.
- 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.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	54
---------------------	---	---	-----------

- 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 data set (09/2014-12/2021) with improved coverage based on OCO-2 v10 L1b data.
- Improvement of the computational efficiency.

v09

- Migration of the 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,

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	55
---------------------	---	---	-----------

the maximum number of successive accelerated soundings is limited to 25.

v08

- Generation of a global 4-years data set.
- Improved cross section data bases with finer temperature, pressure, and wavelength grid in the wCO₂ (0.0026nm) and sCO₂ (0.0044nm) band.
- Quadratic wavelength and linear pressure interpolation of the cross section data base.
- Usage of HITRAN2016 as H₂O 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.

v06

- First application and validation of the FOCAL OCO-2 XCO₂ algorithm to a larger global dataset of actually measured OCO-2 data as described by Reuter et al. (2017b).
- 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.

v01

The initial version of the FOCAL OCO-2 XCO₂ algorithm as described by Reuter et al. (2017c). This version has been used to analyzed simulated OCO-2 measurements only.

References

- Ackerman, S., Frey, R., Strabala, K., Liu, Y., Gumley, L., Baum, B., and Menzel, P.: Discriminating clear-sky from cloud with MODIS - Algorithm Theoretical Basis Document (MOD35), Version 6.1, Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin - Madison, 2010.
- Bennartz, R. and Preusker, R.: Representation of the photon pathlength distribution in a cloudy atmosphere using finite elements, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 98, 202–219, 2006.
- Boesch, H., Brown, L., Castano, R., Christi, M., Connor, B., Crisp, D., Eldering, A., Fisher, B., Frankenberg, C., Gunson, M., Granat, R., McDuffie, J., Miller, C., Natraj, V., O'Brien, D., O'Dell, C., Osterman, G., Oyafuso, F., Payne, V., Polonski, I., Smyth, M., Spurr, R., Thompson, D., and Toon, G.: Orbiting Carbon Observatory-2 (OCO-2) - Level 2 Full Physics Retrieval - Algorithm Theoretical Basis, Version 2.0 Rev 2, National Aeronautics and Space Administration, Jet Propulsion Laboratory, California Institute of Technology, URL https://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/OC02_L2_ATBD.V6.pdf, 2015.
- Bovensmann, H., Burrows, J. P., Buchwitz, M., Frerick, J., Noël, S., Rozanov, V. V., Chance, K. V., and Goede, A.: SCIAMACHY – Mission Objectives and Measurement Modes, *Journal of the Atmospheric Sciences*, 56, 127–150, URL [http://dx.doi.org/10.1175/1520-0469\(1999\)056<0127:SMOAMM>2.0.CO;2](http://dx.doi.org/10.1175/1520-0469(1999)056<0127:SMOAMM>2.0.CO;2), 1999.
- Bovensmann, H., Buchwitz, M., Burrows, J. P., Reuter, M., Krings, T., Gerilowski, K., Schneising, O., Heymann, J., Tretner, A., and Erzinger, J.: A remote sensing technique for global monitoring of power plant CO₂ emissions from space and related applications, *Atmospheric Measurement Techniques*, 3, 781–811, doi:10.5194/amt-3-781-2010, URL <http://www.atmos-meas-tech.net/3/781/2010/>, 2010.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, doi:10.1023/a:1010933404324, URL <https://doi.org/10.1023/a:1010933404324>, 2001.
- Bril, A., Oshchepkov, S., and Yokota, T.: Application of a probability density function-based atmospheric light-scattering correction to carbon dioxide retrievals from GOSAT over-sea observations, *Remote Sensing of the Environment*, 117, 301–306, 2012.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	57
---------------------	---	---	-----------

- Bril, A., Oshchepkov, S., Yokota, T., and Inoue, G.: Parameterization of aerosol and cirrus cloud effects on reflected sunlight spectra measured from space: application of the equivalence theorem, *Applied Optics*, 46, 2460–2470, 2007.
- Buchwitz, M., Detmers, R., and GHG-CCI project team with contributions from NASA/ACOS/OCO-2 team members: ESA Climate Change Initiative (CCI) Product Specification Document (PSD) for the Essential Climate Variable (ECV) Greenhouse Gases (GHG) - Description of Common Parameters for core (ECA) products, URL https://www.iup.uni-bremen.de/carbon_ghg/docs/GHG-CCIplus/PSD/PSDv3_GHG-CCI_final.pdf, 2014.
- Burrows, J. P., Hölzle, E., Goede, A. P. H., Visser, H., and Fricke, W.: SCIAMACHY – Scanning Imaging Absorption Spectrometer for Atmospheric Cartography, *Acta Astronautica*, 35, 445–451, 1995.
- Chevallier, F., Bréon, F.-M., and Rayner, P. J.: Contribution of the Orbiting Carbon Observatory to the estimation of CO₂ sources and sinks: Theoretical study in a variational data assimilation framework, *Journal of Geophysical Research*, 112, D09307, doi:10.1029/2006JD007375, URL <http://dx.doi.org/10.1029/2006JD007375>, 2007.
- Crisp, D., Atlas, R. M., Bréon, F.-M., Brown, L. R., Burrows, J. P., Ciais, P., Connor, B. J., Doney, S. C., Fung, I. Y., Jacob, D. J., Miller, C. E., O'Brien, D., Pawson, S., Randerson, J. T., Rayner, P., Salawitch, R. S., Sander, S. P., Sen, B., Stephens, G. L., Tans, P. P., Toon, G. C., Wennberg, P. O., Wofsy, S. C., Yung, Y. L., Kuang, Z., Chudasama, B., Sprague, G., Weiss, P., Pollock, R., Kenyon, D., and Schroll, S.: The Orbiting Carbon Observatory (OCO) mission, *Advances in Space Research*, 34, 700–709, 2004.
- Crisp, D., Pollock, H. R., Rosenberg, R., Chapsky, L., Lee, R. A. M., Oyafuso, F. A., Frankenberg, C., O'Dell, C. W., Bruegge, C. J., Doran, G. B., Eldering, A., Fisher, B. M., Fu, D., Gunson, M. R., Mandrake, L., Osterman, G. B., Schwandner, F. M., Sun, K., Taylor, T. E., Wennberg, P. O., and Wunch, D.: The on-orbit performance of the Orbiting Carbon Observatory-2 (OCO-2) instrument and its radiometrically calibrated products, *Atmospheric Measurement Techniques*, 10, 59–81, doi:10.5194/amt-10-59-2017, URL <https://www.atmos-meas-tech.net/10/59/2017/>, 2017.
- Eldering, A., Pollock, R., Lee, R., Rosenberg, R., Oyafuso, F., Crisp, D., Chapsky, L., and Granat, R.: Orbiting Carbon Observatory-2 (OCO-2) - LEVEL 1B - Algorithm Theoretical Basis, Version 1.2 Rev 1, National

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	58
---------------------	---	---	-----------

Aeronautics and Space Administration, Jet Propulsion Laboratory, California Institute of Technology, URL http://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/OCO2_L1B_ATBD.V7.pdf, 2015.

European Space Agency, Sinergise: Copernicus Global Digital Elevation Model. Distributed by OpenTopography, doi:<https://doi.org/10.5069/G9028PQB>, 2021.

Frankenberg, C., Butz, A., and Toon, G. C.: Disentangling chlorophyll fluorescence from atmospheric scattering effects in O₂ A-band spectra of reflected sun-light, *Geophysical Research Letters*, 38, n/a–n/a, doi:10.1029/2010GL045896, URL <http://dx.doi.org/10.1029/2010GL045896>, 103801, 2011.

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, *Machine Learning*, 63, 3–42, doi:10.1007/s10994-006-6226-1, URL <https://doi.org/10.1007/s10994-006-6226-1>, 2006.

Gordon, I., Rothman, L., Hill, C., Kochanov, R., Tan, Y., Bernath, P., Birk, M., Boudon, V., Campargue, A., Chance, K., Drouin, B., Flaud, J.-M., Gamache, R., Hodges, J., Jacquemart, D., Perevalov, V., Perrin, A., Shine, K., Smith, M.-A., Tennyson, J., Toon, G., Tran, H., Tyuterev, V., Barbe, A., Császár, A., Devi, V., Furtenbacher, T., Harrison, J., Hartmann, J.-M., Jolly, A., Johnson, T., Karman, T., Kleiner, I., Kyuberis, A., Loos, J., Lyulin, O., Massie, S., Mikhailenko, S., Moazzen-Ahmadi, N., Müller, H., Naumenko, O., Nikitin, A., Polyansky, O., Rey, M., Rotger, M., Sharpe, S., Sung, K., Starikova, E., Tashkun, S., Auwera, J. V., Wagner, G., Wilzewski, J., Wcisło, P., Yu, S., and Zak, E.: The HITRAN2016 molecular spectroscopic database, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 203, 3–69, doi:<https://doi.org/10.1016/j.jqsrt.2017.06.038>, URL <https://www.sciencedirect.com/science/article/pii/S0022407317301073>, HITRAN2016 Special Issue, 2017.

Heymann, J., Reuter, M., Hilker, M., Buchwitz, M., Schneising, O., Bovensmann, H., Burrows, J. P., Kuze, A., Suto, H., Deutscher, N. M., Dubey, M. K., Griffith, D. W. T., Hase, F., Kawakami, S., Kivi, R., Morino, I., Petri, C., Roehl, C., Schneider, M., Sherlock, V., Sussmann, R., Velasco, V. A., Warneke, T., and Wunch, D.: Consistent satellite XCO₂ retrievals from SCIAMACHY and GOSAT using the BESD algorithm, *Atmospheric Measurement Techniques*, 8, 2961–2980, doi:10.5194/amt-8-2961-2015, URL <http://www.atmos-meas-tech.net/8/2961/2015/>, 2015.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	59
---------------------	---	---	-----------

- Hu, C., Lee, Z., and Franz, B.: Chlorophyll-a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference, *Journal of Geophysical Research: Oceans*, 117, 2012.
- Kurucz, H. L.: The solar spectrum: atlases and line identifications, workshop on laboratory and astronomical high resolution spectra. astronomical society of the pacific conference series, *Proceedings of ASP Conference No.81* Held in Brussels, Belgium 29 August-2 September 1994, pp. 17–31, URL www.scopus.com, 1995.
- Kuze, A., Suto, H., Nakajima, M., and Hamazaki, T.: Thermal and near infrared sensor for carbon observation Fourier-transform spectrometer on the Greenhouse Gases Observing Satellite for greenhouse gases monitoring, *Applied Optics*, 48, 6716, doi:10.1364/AO.48.006716, URL <http://dx.doi.org/10.1364/AO.48.006716>, 2009.
- Meftah, M., Damé, L., Bolsée, D., Hauchecorne, A., Pereira, N., Sluse, D., Cessateur, G., Irbah, A., Bureau, J., Weber, M., Bramstedt, K., Hilbig, T., Thiéblemont, R., Marchand, M., Lefèvre, F., Sarkissian, A., and Bekki, S.: SOLAR-ISS: A new reference spectrum based on SOLAR/SOLSPEC observations, *Astronomy & Astrophysics*, 611, A1, doi:10.1051/0004-6361/201731316, 2018.
- Miller, C. E., Crisp, D., DeCola, P. L., Olsen, S. C., Randerson, J. T., Michalak, A. M., Alkhaled, A., Rayner, P., Jacob, D. J., Suntharalingam, P., Jones, D. B. A., Denning, A. S., Nicholls, M. E., Doney, S. C., Pawson, S., Bösch, H., Connor, B. J., Fung, I. Y., O'Brien, D., Salawitch, R. J., Sander, S. P., Sen, B., Tans, P., Toon, G. C., Wennberg, P. O., Wofsy, S. C., Yung, Y. L., and Law, R. M.: Precision requirements for space-based X_{CO_2} data, *Journal of Geophysical Research*, 112, D10314, doi:10.1029/2006JD007659, 2007.
- Noël, S., Reuter, M., Buchwitz, M., Borchardt, J., Hilker, M., Bovensmann, H., Burrows, J. P., Di Noia, A., Suto, H., Yoshida, Y., Buschmann, M., Deutscher, N. M., Feist, D. G., Griffith, D. W. T., Hase, F., Kivi, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Podolske, J. R., Pollard, D. F., Sha, M. K., Shiomi, K., Sussmann, R., Té, Y., Velazco, V. A., and Warneke, T.: X_{CO_2} retrieval for GOSAT and GOSAT-2 based on the FOCAL algorithm, *Atmospheric Measurement Techniques*, 14, 3837–3869, doi:10.5194/amt-14-3837-2021, URL <https://amt.copernicus.org/articles/14/3837/2021/>, 2021.
- Noël, S., Reuter, M., Buchwitz, M., Borchardt, J., Hilker, M., Schneising, O., Bovensmann, H., Burrows, J. P., Di Noia, A., Parker, R. J., Suto, H., Yoshida,

- Y., Buschmann, M., Deutscher, N. M., Feist, D. G., Griffith, D. W. T., Hase, F., Kivi, R., Liu, C., Morino, I., Notholt, J., Oh, Y.-S., Ohyama, H., Petri, C., Pollard, D. F., Rettinger, M., Roehl, C., Rousogenuous, C., Sha, M. K., Shiomi, K., Strong, K., Sussmann, R., T  , Y., Velazco, V. A., Vrekoussis, M., and Warneke, T.: Retrieval of greenhouse gases from GOSAT and GOSAT-2 using the FOCAL algorithm, *Atmospheric Measurement Techniques*, 15, 3401–3437, doi:10.5194/amt-15-3401-2022, URL <https://amt.copernicus.org/articles/15/3401/2022/>, 2022.
- O'Dell, C. W., Connor, B., B  sch, H., O'Brien, D., Frankenberg, C., Castano, R., Christi, M., Eldering, D., Fisher, B., Gunson, M., McDuffie, J., Miller, C. E., Natraj, V., Oyafuso, F., Polonsky, I., Smyth, M., Taylor, T., Toon, G. C., Wennberg, P. O., and Wunch, D.: The ACOS CO₂ retrieval algorithm - Part 1: Description and validation against synthetic observations, *Atmospheric Measurement Techniques*, 5, 99–121, doi:10.5194/amt-5-99-2012, URL <http://www.atmos-meas-tech.net/5/99/2012/>, 2012.
- Rascher, U., Agati, G., Alonso, L., Cecchi, G., Champagne, S., Colombo, R., Damm, A., Daumard, F., de Miguel, E., Fernandez, G., Franch, B., Franke, J., Gerbig, C., Gioli, B., G  mez, J. A., Goulas, Y., Guanter, L., Guti  rrez-de-la C  mara, O., Hamdi, K., Hostert, P., Jim  nez, M., Kosvanova, M., Lognoli, D., Meroni, M., Miglietta, F., Moersch, A., Moreno, J., Moya, I., Neininger, B., Okujeni, A., Ounis, A., Palombi, L., Raimondi, V., Schickling, A., Sobrino, J. A., Stellmes, M., Toci, G., Toscano, P., Udelhoven, T., van der Linden, S., and Zaldei, A.: CEFLES2: the remote sensing component to quantify photosynthetic efficiency from the leaf to the region by measuring sun-induced fluorescence in the oxygen absorption bands, *Biogeosciences*, 6, 1181–1198, doi:10.5194/bg-6-1181-2009, URL <http://www.biogeosciences.net/6/1181/2009/>, 2009.
- Reuter, M., Buchwitz, M., Schneising, O., Hase, F., Heymann, J., Guerlet, S., Cogan, A. J., Bovensmann, H., and Burrows, J. P.: A simple empirical model estimating atmospheric CO₂ background concentrations, *Atmospheric Measurement Techniques*, 5, 1349–1357, doi:10.5194/amt-5-1349-2012, URL <http://dx.doi.org/10.5194/amt-5-1349-2012>, 2012.
- Reuter, M., Buchwitz, M., Hilker, M., Heymann, J., Bovensmann, H., Burrows, J. P., Houweling, S., Liu, Y. Y., Nassar, R., Chevallier, F., Ciais, P., Marshall, J., and Reichstein, M.: How Much CO₂ Is Taken Up by the European Terrestrial Biosphere?, *Bulletin of the American Meteorological Society*, 98, 665–671, doi:10.1175/BAMS-D-15-00310.1, 2017a.

ESA CCI+ ECV GHG	ATBD FOCAL OCO-2 Version 5 February 2024	Institute of Env. Physics, University of Bremen	61
---------------------	---	---	-----------

- Reuter, M., Buchwitz, M., Schneising, O., Noël, S., Bovensmann, H., and Burrows, J. P.: A fast atmospheric trace gas retrieval for hyperspectral instruments approximating multiple scattering - Part 2: application to XCO₂ retrievals from OCO-2, Remote Sensing, 9, doi:10.3390/rs9111102, URL <http://www.mdpi.com/2072-4292/9/11/1102>, 2017b.
- 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 a potential OCO-2 XCO₂ retrieval setup, Remote Sensing, 9, doi:10.3390/rs9111159, URL <http://www.mdpi.com/2072-4292/9/11/1159>, 2017c.
- Reuter, M., Hilker, M., and Buchwitz, M.: Product User Guide Version 5 (PUGv5) for the FOCAL XCO₂ OCO-2 Data Product CO₂_OC₂_FOCA v11 for the Essential Climate Variable (ECV) Greenhouse Gases (GHG), ESA Climate Change Initiative Plus ESA GHG CCI+, 2024a.
- Reuter, M., Hilker, M., and GHG-CCI group at IUP: End-to-End ECV Uncertainty Budget Version 5 (E3UBv5) for the FOCAL XCO₂ OCO-2 Data Product CO₂_OC₂_FOCA v11 for the Essential Climate Variable (ECV) Greenhouse Gases (GHG), ESA Climate Change Initiative Plus ESA GHG CCI+, 2024b.
- Rodgers, C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice, World Scientific Publishing, Singapore, 2000.
- Roedel, W. and Wagner, T.: Physik unserer Umwelt: Die Atmosphäre, Springer, 2011.
- Thompson, D. R., Chris Benner, D., Brown, L. R., Crisp, D., Malathy Devi, V., Jiang, Y., Natraj, V., Oyafuso, F., Sung, K., Wunch, D., and et al.: Atmospheric validation of high accuracy CO₂ absorption coefficients for the OCO-2 mission, J Quant Spectrosc Radiat Transfer, 113, 2265–2276, doi: 10.1016/j.jqsrt.2012.05.021, URL <http://dx.doi.org/10.1016/j.jqsrt.2012.05.021>, 2012.
- Yoshida, Y., Kikuchi, N., Morino, I., Uchino, O., Oshchepkov, S., Bril, A., Saeki, T., Schutgens, N., Toon, G. C., Wunch, D., and et al.: Improvement of the retrieval algorithm for GOSAT SWIR XCO₂ and XCH₄ and their validation using TCCON data, Atmospheric Measurement Techniques, 6, 1533–1547, doi:10.5194/amt-6-1533-2013, URL <http://dx.doi.org/10.5194/amt-6-1533-2013>, 2013.