ESA Climate Change Initiative "Plus" (CCI+)

Algorithm Theoretical Basis Document (ATBD) Version 1.5

–

For the RemoTeC XCH4 GOSAT-2

SRON Proxy Product (CH4_GO2_SRPR) Version 2.0.3

for the Essential Climate Variable (ECV)

Greenhouse Gases (GHG)

Written by:

GHG-CCI+ project team

Lead author:

Andrew Gerald Barr

(SRON Netherlands Institute for Space Research, Leiden, The Netherlands)

Co-authors:

Tobias Borsdorff, Jochen Landgraf

(SRON Netherlands Institute for Space Research, Leiden, The Netherlands)

Change log

Version Nr.	Date	Status	Reason for change
Version 1 draft	24. Feb. 2020	Draft	New document
Version 1	1. Dec. 2020	As submitted	Update after ESA review: - Definition of SRPR - Update equation and figure numbers - Remove typos
Version 1.1	3. Dec. 2020	As submitted	- Add list of acronyms and abbreviations - Correct terms - Change LMD inversion to CO_2 correction
Version 1.2	3. Dec. 2020	Approved	Minor editorial changes
Version 1.3	31. July 2021	As submitted	- Update radiance definition - Update ECMWF data - Update solar irradiance data - Update cloud filtering - Update reference
Version 1.4	10. Feb. 2023	As submitted	- New version (v2.0.2) - Update of XCO2 model used in the proxy calculation - Post-processing filtering using random forest model
Version 1.5	7. Feb. 2024	As submitted	- New version (v2.0.3) - Re-processing of Feb 2019 to Aug 2020

Table of Contents

List of acronyms and abbreviations

Acronym	Definition
AOT	Aerosol Optical Thickness
ATBD	Algorithm Theoretical Basis Document
CAMS	Copernicus Atmosphere Monitoring Service
CCI+	Climate Change Initiative Plus
CH_4	Methane
CO	Carbon Monoxide
CO_2	Carbon Dioxide
CPU	Core Processing Unit
DEM	Digital Elevation Map
DFS	Degrees of Freedom for Signal
ECMWF	European Centre for Medium Range Weather Forecasting
ESA	European Space Agency
FP	Full Physics retrieval method
FTIR	Fourier Transform InfraRed
FTS	Fourier Transform Spectrometer
GHG	GreenHouse Gas
GOSAT	Greenhouse Gases Observing Satellite
HSRS	Hybrid Solar Reference Spectrum
ISRF	Instrument Spectral Response Function
JAXA	Japan Aerospace Exploration Agency
L1	Level 1
L2	Level 2
LMD	Laboratoire de Météorologie Dynamique
NA	Not applicable
NASA	National Aeronautics and Space Administration
NetCDF	Network Common Data Format
NIR	Near Infrared
O_2	Oxygen
ppb	Parts per billion
ppm	Parts per million
PUG	Product User Guideline
RAA	Relative Azimuth Angle
RemoTeC	Remote Sensing of Greenhouse Gases for Carbon Cycle Modeling
SCIAMACHY	Scanning Imaging Absorption Spectrometer for Atmospheric CHartographY
SRPR	SRON Proxy

SRON	SRON Netherlands Institute for Space Research
SRTM	Shuttle Radar Topography Mission
SWIR	Short Wave Infrared
SZA	Solar Zenith Angle
TANSO	Thermal And Near infrared Sensor for carbon Observation
TCCON	Total Carbon Column Observing Network
TIR	Thermal Infrared
TM	Transport Model
TSIS	Total and Spectral Solar Irradiance Sensor
VZA	Viewing Zenith Angle
XCO_2	Column-averaged dry-air mixing ratios (mole fractions) of CO2
XCH_4	Column-averaged dry-air mixing ratios (mole fractions) of CH_4

Scope of document

This Algorithm Theoretical Basis Document (ATBD) describes the SRON-RemoTeC algorithm used to generate the CH4_GO2_SRPR product. GO2 stands for GOSAT-2 and SRPR is an abbreviation for the "SRON Proxy" product.

The RemoTeC algorithm is used to retrieve column averaged dry air mole fraction of methane (CH_4), denoted as XCH_4, as well as other parameters included in the Level 2 product CH4_GO2_SRPR generated from GOSAT-2 Level 1b spectra.

This document details various input data required for retrievals, physical theory and mathematical background underlying retrieval assumptions, and outlines retrieval implementation and limitations of the approach used.

Executive summary

This document describes the RemoTeC algorithm for GHG retrieval from the GOSAT-2 instrument. The algorithm is based on the paper of *Butz et al., 2009*. Tests of the retrieval algorithm have been performed on synthetic GOSAT data (*Butz et al., 2010*), and applied to real GOSAT data (*Butz et al., 2011; Schepers et al., 2012; Guerlet et al., 2013*). Here we apply the retrieval for the first time to GOSAT-2 data.

In order to account for the effect of aerosols and cirrus, the proxy method retrieves both the CH_4 and CO_2 column under the assumption of a non-scattering atmosphere assuming that scattering effects cancel by taking the ratio of the 2 is calculated. XCH_4 is then calculated by multiplying the ratio with XCO_2 obtained from a model. Additional fit parameters are the surface albedo and its 1st order spectral dependence in all bands, and the total column of water vapor, respectively.

In order to obtain a proper characterization of the retrieved XCH_4, it is important to first retrieve a vertical profile (layer averaged number density in different layers of the model atmosphere) and use this retrieved vertical profile to calculate the vertical column. Here, we choose to provide the vertical column as a product, and not the full profile, because the Degrees of Freedom for Signal (DFS) of the retrieved CH_4 and CO_2 profile is about 1. The inversion is performed using Phillips-Tikhonov regularization in combination with a reduced step size Gauss-Newton iteration scheme.

1 Data product overview

1.1 Column-averaged mixing ratios of CH_4 (XCH_4) and CO_2 (XCO_2)

In this section an overview of the data product (specified in terms of variable, its property, processing level(s) and instrument(s)) is given.

The data products:

- Column-averaged dry-air mixing ratios (mole fractions) of CH_4, denoted XCH_4 (in parts per billion, ppb).

In the following, several satellite instruments are shortly described which are used or can be used to generate the XCH_4 data product CH4_GO2_SRPR.

TANSO-FTS-2 is a Fourier-Transform-Spectrometer (FTS) onboard the Japanese GOSAT-2 satellite (*Nakajima et al., 2017*). The Japanese Greenhouse gases Observing SATellite-2 (GOSAT-2) was launched on 29th October 2018 and started operational observations form February 2019. GOSAT-2 provides dedicated global measurements of total column CO_2 and CH_4 from its SWIR bands. GOSAT-2 covers the relevant CO_2, CH_4 and O_2 absorption bands in the NIR and SWIR spectral region as needed for accurate XCO_2 and XCH_4 retrieval (in addition GOSAT-2 also covers a large part of the Thermal Infrared (TIR) spectral region). The spectral resolution of TANSO-FTS-2 is much higher compared to e.g. SCIAMACHY and also the ground pixels are smaller (9.7 km compared to several 10 km for SCIAMACHY). However, in contrast to SCIAMACHY, the GOSAT-2 scan pattern consists of non-consecutive individual ground pixels, i.e., the scan pattern is not gap-free. For a good general overview about GOSAT-2 see also http://www.GOSAT-2.nies.go.jp/.

2 Input and auxiliary data

2.1 Satellite instrument

2.1.1 GOSAT-2 TANSO-FTS-2 Level 1b

Level 1b data of the TANSO-FTS-2 (Thermal And Near-infrared Sensor for carbon Observation - Fourier Transform Spectrometer-2) onboard GOSAT-2 (Greenhouse gas Observing SATellite) are needed in the project to produce the total column CO2 and CH4 products. They serve as input for the retrieval algorithms to be used in this project.

The TANSO-FTS-2 instrument (*Nakajima et al., 2017*) has five spectral bands with a high spectral resolution 0.2 cm^{-1}. Three operate in the SWIR at 0.75-0.77, 1.56-1.69 and at the extended 1.92-2.33 µm range, providing sensitivity to the near-surface absorbers. The fourth and fifth channels operating in the thermal infrared between 5.5-8.4 and 8.4-14.3 µm providing mid-tropospheric sensitivity. FTS-2 observes sunlight reflected from the earth's surface and light emitted from the atmosphere and the surface. The former is observed in the spectral bands 1 through 3 of FTS-2 in the daytime, and the latter is captured in band 4 and 5 during both the day and the night. Within this project only level 1 data from the SWIR channels 1-3 will be used. Prior to reaching the detectors of the instrument, the light in the bands 1 through 3 is split into two orthogonally-polarized components and measured independently. The intensity component of Stokes vector is approximated by the mean of parallel (P) and perpendicular (S) components (*O'Brien et al., 2013*).

The measurement strategy of TANSO-FTS-2 is optimized for the characterization of continental-scale sources and sinks. TANSO-FTS-2 utilizes a pointing mirror to perform off-nadir measurements at the same location on each 6-day repeat cycle. The pointing mirror allows TANSO-FTS-2 to observe up to ±35° across track and ±40° along-track. These measurements nominally consist of 5 across track points spaced ~160km apart with a ground footprint diameter of approximately 9.7 km and a 4 second exposure duration. The satellite has an Intelligent pointing Monitor camera which makes it possible to adjust the line of sight of the FTS-2 to steer away from cloud contaminated areas. Whilst the majority of data is limited to measurements over land where the surface reflectance is high, TANSO-FTS-2 also observes in sunglint mode over the ocean.

Table 1: TANSO-FTS-2 bands.

Channel	Wavelength range [nm]	Resolution [cm^{-1}]
1	758-775	0.2
2	1460-1720	0.2
3	1920-2330*	0.2
4	5560-8400	0.2
5	8400-14300	0.2

***GOSAT-1 only had a spectral range up to 2080nm.**

Data availability & coverage: TANSO-FTS-2 Level-1b data are available since February 2019.
Source data product name & reference to product technical specification documents: GOSAT FTS-2 L1A/L1B Product Description Document, Japan Aerospace Exploration Agency, July 2019, GST-18005, available through https://prdct.gosat-2.nies.go.jp/en/documents/JAXA_GOSAT-2_FTS-2_L1_Data_Description_Document_101101_en.pdf
Data quality and reliability: The quality of the retrieved CO_2 and CH_4 columns has been tested against ground-based observations (i.e. the TCCON network) and has shown to be of good quality. The spectral bands showed some irregularities which required a shortened retrieval window (O_2 A-band) and spectral intensity offsets for each of the bands. This in effect worsened the cost function χ^2 of the fits compared to similar GOSAT-1 retrievals and the estimated uncertainties of the species. However, comparison to TCCON observations showed that the retrieval products of both the GOSAT-1 and GOSAT-2 are of similar quality.

2.2 Other

2.2.1 ECMWF model data

The retrieval algorithms to produce vertical columns of CO_2 and CH_4 need as input for each scene the temperature vertical profile, pressure vertical profile, specific humidity vertical profile, and wind speed. Here, temperature and pressure are needed to calculate absorption cross sections, the specific humidity vertical profile is needed to account for water vapor absorption, and the wind speed is needed to calculate the Fresnel reflection contribution on a rough ocean surface. The meteorological data mentioned above will be taken from the ECMWF model.

Originating system: ECMWF has developed one of the most comprehensive earth-system models available anywhere. The ECMWF model is uses the '4D-Var' data assimilation approach, which provides a physically consistent best fit to observations. For this project the ERA-5 data are used.

Data class: Model
Required ECMWF data:
Class: ERA-5
Stream: Atmospheric model
Type: Analysis
Dates: 02/05/2019 to – 31/12/2021
Time: 00:00:00, 03:00:00, 06:00:00, 09:00:00, 12:00:00, 15:00:00, 18:00:00, 21:00:00
Spatial grid: T639 Quasi-uniform Gaussian grid (~0.28°, 31km)
Parameters at model levels:

- temperature, specific humidity (all levels)
- logarithm of surface pressure, geopotential (lowest level)

Parameters at surface:

- 10 meter U wind component
- 10 meter V wind component

Data availability & coverage: All data are required on a global scale, with a typical delay of three months.
Data quality and reliability: The ECMWF model data sets are widely considered to be among the best available data sets for meteorological parameters.

2.2.2 CO_2 correction

The retrieval algorithm for CH_4 columns that are based on the "proxy approach" retrieve the ratio of the CH_4 and CO_2 columns, where the CO_2 column serves as a proxy for the light path. In order to obtain the CH_4 column, the retrieved ratio needs to be multiplied by the best estimate of the CO_2 column. For this purpose, we use CO2 from the Copernicus Atmosphere Monitoring Service (CAMS).

Originating system: The inversion products used here are the official CAMS v19r1, v20r2 and v21r1 products that exclusively assimilates about 130 sites of surface air sample measurements from the Global Atmosphere Watch programme.
Data class: Model
Data availability & coverage: CAMS provides global 3-hourly data at a spatial resolution of 3.75 by 1.89 degree. Data are typically available within one year. Product versions v19r1, v20r2 and v21r1 cover time periods of 2019, 2020 and 2021, respectively.
Source data product name & reference to product technical specification documents: *Chevallier, 2019, Chevallier, 2020 and Chevallier, 2021*.
Data quality and reliability: The CAMS v19r1, v20r2 and v21r1 inversion products have been validated with many independent measurements and are considered to provide accurate global CO_2 data sets.

2.2.3 SRTM DEM

The RemoTeC retrieval algorithm for CO_2 and CH_4 columns from GOSAT-2 use information about the surface elevation from an extended SRTM digital elevation map.

Originating system: The original Shuttle Radar Telemetry Mission (SRTM) was provided by the United States National Aeronautics and Space Administration (NASA). The dataset used (DEM3) is based on the SRTM dataset and includes extrapolation and gap filling from various sources.
Data class: Model
Sensor type and key technical characteristics: n/a

Data availability & coverage: The original SRTM dataset provides elevation data ranging from 56 degrees south to 60 degrees north at a 90 meter resolution. The adjusted DEM3 dataset extends the coverage, while keeping the 90 meter resolution.
Source data product name & reference to product technical specification documents: http://www.viewfinderpanoramas.org/dem3.html.

2.2.4 TCCON FTS CO_2 and CH_4 data

TCCON data for CO_2 and CH_4 is available for public for all TCCON stations (https://tccon-wiki.caltech.edu/). We use the GGG2020 official release of the data product.

Originating system: Ground based
Data class: Ground based
Sensor type and key technical characteristics: The measurements are performed using the solar absorption spectroscopy in the near infrared using an FTS.
Data availability & coverage: Coverage is limited to the locations of the TCCON stations themselves. Depending on the instrument setting (land or sunglint), we use different TCCON stations for validation. Data is typically available only after one year, although some stations deliver on a more regular interval (3 months).
Source data product name & reference to product technical specification documents: *Wunch et al., 2015, Laughner et al., 2021*
Data quantity: Individual measurements can be taken in intervals of about 20 min. The observations can only be taken with the direct sunlight. This limits the amount of data, which is different from site to site.
Data quality and reliability: For XCO_2 the precision is 0.25% (1ppm) and the systematic error (bias) is 0.2% (0.8 ppm). For XCH_4 the precision is 0.40% (7ppb) and the systematic error (bias) is also 0.40% (7 ppb).

2.2.5 Additional input data

- Absorption cross sections: For the retrieval lookup-tables are used with pre-calculated absorption cross sections of the species of interest (O_2, CO_2, CH_4, and H_2O) as a function of wavenumber, temperature, and pressure. One lookup-table per species and per spectral band is being used. At the start of processing at a given CPU (Figure 3) the cross-section lookup table is read into memory.
- Aerosol optical properties: A lookup table is being used with pre-calculated aerosol optical properties (Mie and t-Matrix theory) as a function of size parameter and refractive index.
- Solar source: The solar source for the forward simulation uses data of extraterrestrial irradiance from the Total and Spectral Solar Irradiance Sensor-1 Hybrid Solar Reference Spectrum (TSIS-1 HSRS). It covers a temporally constant irradiance spectrum between 200 and 2700 nm with a spectral resolution of 0.001 nm

(*Coddington et al., 2021*). The TSIS-1 HSRS data are downloaded from https://lasp.colorado.edu/lisird/data/tsis1_hsrs.

- Retrieval settings: A file is read in with retrieval settings such as fit parameters, spectral range, etc.

2.3 Overview of Processing Sub-System

Figure 1 provides a schematic overview of the RemoTeC GHG-CCI+ processing sub system at SRON. The first step is to download the required data from the respective data servers to SRON. GOSAT-2 L1b and ECMWF data are dynamic datasets that are continuously updated, while the SRTM topography is a static dataset. In the next step a pre-processing program is combining all relevant information per GOSAT-2 ground pixel. This includes interpolation of ECMWF data in space and time to the coordinates of the GOSAT-2 ground pixel, calculating the average height of a GOSAT-2 ground pixel and its standard deviation from the topography database.

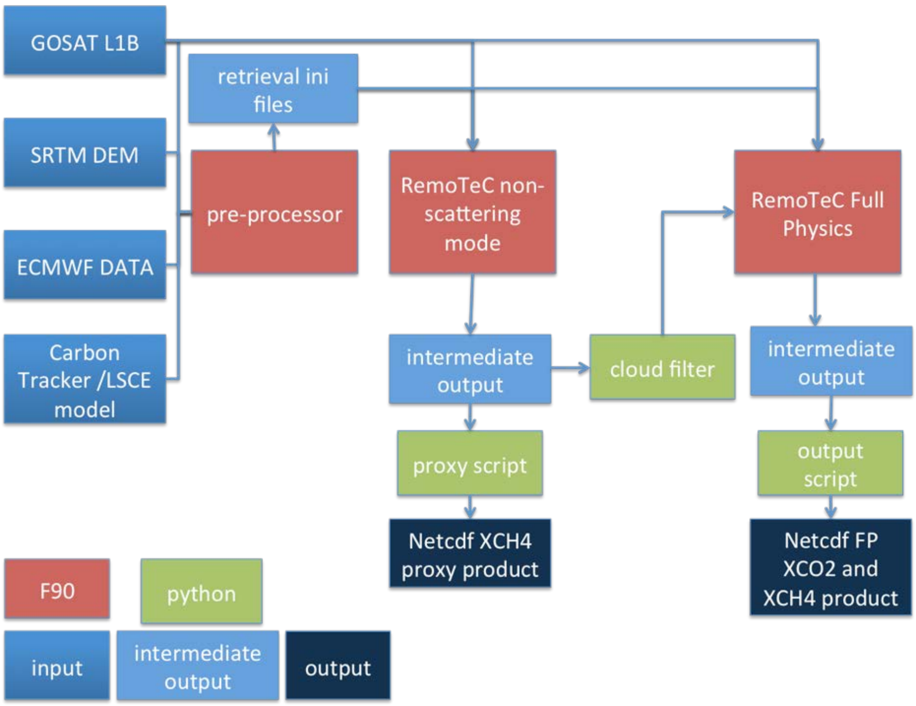

Figure 1: Schematic overview of the RemoTeC algorithm processing sub-system.

The pre-processor produces for each GOSAT-2 L1B file an auxiliary input file, hereafter referred to as the 'retrieval ini' file that contains this information. In the next step columns of CO_2, CH_4, H_2O, and O_2 are retrieved under the assumption of an atmosphere without aerosol/cirrus/cloud scattering (see 'RemoTeC non scattering mode' in Figure 1). The outcome of these retrievals is written in an ASCII file (intermediate output) and is used to generate the RemoTeC XCH_4 proxy product and for cloud filtering to select scenes to be processed by the RemoTeC Full Physics algorithm (see Section 3.5). Subsequently the intermediate output will go into *a posterior* filtering procedure, quality check (based on non-convergence, parameter boundary hits, retrieved aerosol parameters), and bias correction and finally a NetCDF output file is created (*PUG_CH4_GO2_SRPR, 2023*).

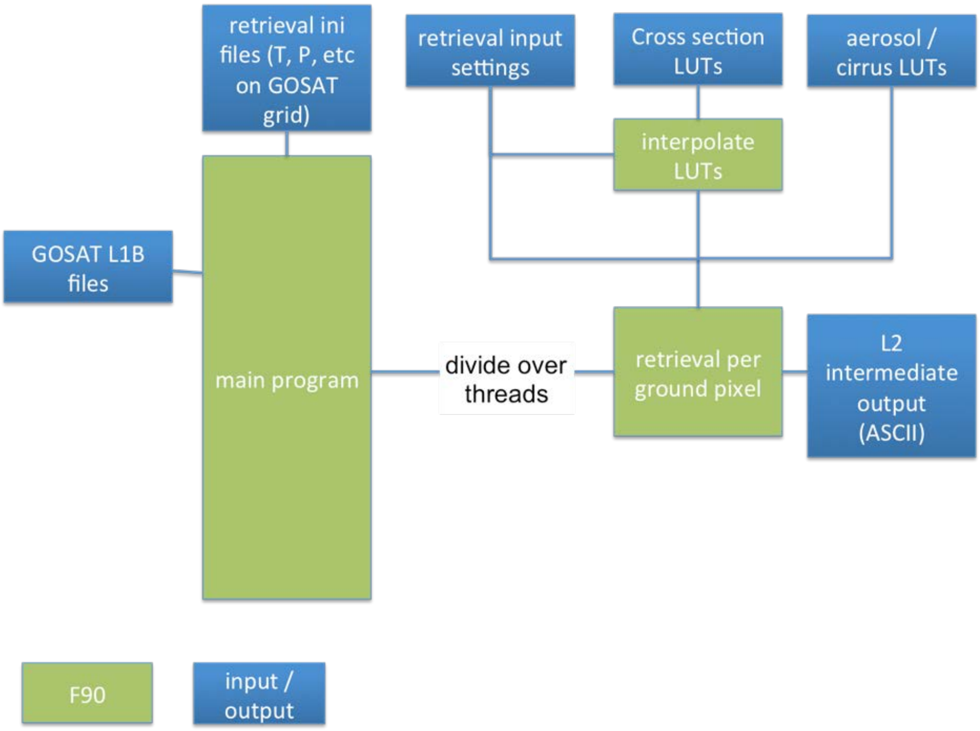

Figure 2: Schematic overview of the RemoTeC retrieval procedure including multi-threading.

Figure 1 gives a schematic overview of the core RemoTeC retrieval algorithm (same for non-scattering and Full Physics). Here, multi-threading capability is implemented using *openMP,* where different ground pixels are divided over multiple threads. Figure 2 shows the processing per ground pixel (i.e. for a single thread) in more detail.

The static input that is required is a lookup table with the relevant absorption cross sections (read into memory at beginning of processing), a lookup table with aerosol optical properties, and a file indicating the retrieval settings (i.e. fit parameters, spectral range, etc.). Further, the auxiliary retrieval input files that are produced by the pre-processor are needed for each GOSAT-2 ground pixel to be processed, together with the GOSAT-2 Level 1b data. The retrieval per pixel is then run (iterative scheme with forward model and inversion module) and after convergence an intermediate (ASCII) output file is created that is used in the a posteriori filtering and quality check (see Figure 3) and the processing of the next ground pixel starts.

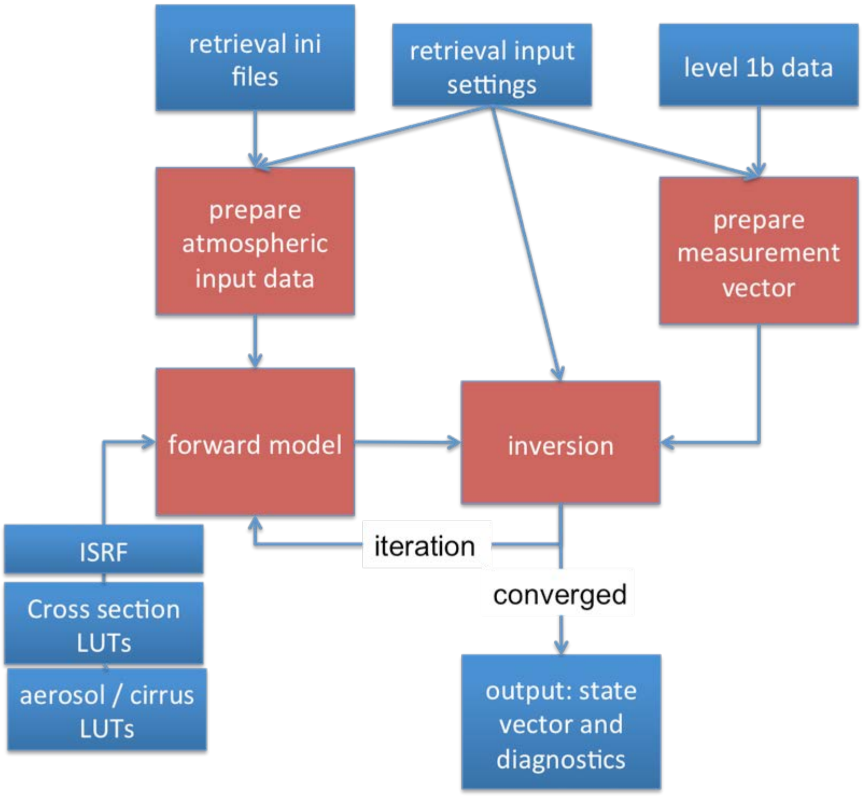

Figure 3: Overview of RemoTeC processing per ground pixel.

3 Algorithms

3.1 Algorithm description

Any retrieval algorithm aims at inferring an atmospheric state vector $\mathbf{x}$ from a measurement vector $\mathbf{y}$. The state vector is linked to the measurement vector through the true forward model $\mathbf{f}(\mathbf{x},\mathbf{b})$ that depends on the state vector $\mathbf{x}$ and the vector $\mathbf{b}$ containing ancillary parameters that are not retrieved,

$$\mathbf{y} = \mathbf{f}(\mathbf{x},\mathbf{b}) + \mathbf{e}_y \quad (1)$$

where $\mathbf{e}_y$ represents the measurement noise vector. A retrieval method approximates the true forward model $\mathbf{f}$ by a retrieval forward model $\mathbf{F}$, with a forward model error vector $\mathbf{e}_F$,

$$\mathbf{y} = \mathbf{F}(\mathbf{x},\mathbf{b}) + \mathbf{e}_y + \mathbf{e}_F \quad (2)$$

For proxy retrieval from the GOSAT-2 FTS-2 instrument the measurement vector contains the measured intensities in the SWIR (see Table 2).

Table 2: Spectral ranges from the NIR and SWIR band included in the measurement vector.

Band	Used spectral range
1 (CO_2)	6170 – 6277 cm^{-1}
2 (CH_4)	6045 – 6138 cm^{-1}

For the retrieval procedure it is needed that the non-linear forward model is linearized so that the retrieval problem can be solved iteratively. For iteration step n the forward model is approximated by

$$\mathbf{F}(\mathbf{x},\mathbf{b}) \approx \mathbf{F}(\mathbf{x}_n,\mathbf{b}) + \mathbf{K}(\mathbf{x}_n - \mathbf{x}) \quad (3)$$

where $\mathbf{x}_n$ is the state vector for the n-th iteration step and $\mathbf{K}$ is the Jacobian matrix

$$\mathbf{K} = \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \quad (4)$$

Below, we will describe the retrieval forward model, state vector, ancillary parameter vector, and the inversion method in more detail.

3.2 Forward Model

The retrieval forward model $\mathbf{F}$ simulates the measurement vector $\mathbf{y}$ for a given model atmosphere defined by the state vector **x** and the ancillary parameter vector **b**. The simulated intensity for a given spectral pixel i is given by

$$I_i = \int_{\lambda_{min}}^{\lambda_{max}} I(\lambda) S_i(\lambda) d\lambda \tag{5}$$

where $I(\lambda)$ is the modeled intensity at high spectral resolution and $S_i(\lambda)$ is the Instrument Spectral Response Function (ISRF) for spectral pixel i. In the NIR and SWIR channel $I(\lambda)$ contains many fine spectral structures due to molecular absorption, so it has to be calculated at fine spectral resolution (0.1 cm^{-1} in the NIR band and 0.02 cm^{-1} in the SWIR).

3.2.1 Model Atmosphere and Optical Properties

For the RemoTeC algorithm described here the model atmosphere is defined for $NLAY = 36$ homogeneous vertical layers that are equidistant in pressure, the lowest pressure level being defined by the surface pressure. The absorbing trace gases of interest are O_2 (in the NIR band) and CH_4, H_2O, and CO in the SWIR band. The layer sub-columns of these gases are for the first iteration step of each retrieval calculated from the input profiles of CH_4, CO (TM5) and H_2O (ECMWF) and the temperature and pressure profiles (ECMWF). They are obtained on the grid of the model atmosphere by linear interpolation. Here, first the surface pressure p_{surf} is obtained by interpolation of the input pressure profile as function of height to the surface height (input) for the corresponding ground pixel. Next the pressure values at the layer boundaries are calculated, with the pressure p_k at the lower boundary of layer k (counting from top to bottom) is given by:

$$p_{lev,k} = p_{min} + \Delta p.k \tag{6}$$

$$\Delta p = (p_{surf} - p_{min})/NLAY \tag{7}$$

where p_{min} is the pressure value of the upper boundary of the input (ECMWF) atmosphere. The different atmospheric profiles are constructed on this pressure grid. For example, the methane sub-column $\mathrm{DV\_CH}_{4k}$ for the layer bounded by pressure levels $p_{\mathrm{lev},k}$ and $p_{\mathrm{lev},k+1}$ is given by:

$$\mathrm{DV\_CH}_{4k} = \mathrm{XCH}_{4k}\ \mathrm{DV\_AIR}_k \tag{8}$$

where $\mathrm{XCH}_{4\mathrm{k}}$ is the methane dry air mixing ratio linearly interpolated from the input pressure grid to the pressure at the 'middle' of layer k defined by $(p_k + p_{k+1})/2$. $\mathrm{DV\_AIR}_\mathrm{k}$ is the sub-column of air in layer k, given by

$$\mathrm{DV\_AIR}_k = (p_{\mathrm{lev},k+1} - p_{\mathrm{lev},k})\ R\ /\ (M\ g_k \left(1 + \frac{\mathrm{XH_2O}_k}{1.60855}\right)) \quad (9)$$

where R is Avogadro's number, M is the molecular mass of air, g_k is the gravity constant in altitude layer k, and 1.60855 is the mass of air relative to the mass of water (*Wunch et al., 2010*). The sub columns of CO and H_2O are calculated in the same manner as for CH_4, and the O_2 sub-column is obtained by multiplying the air sub-column by the O_2 mixing ratio (=0.2095).

In the proxy method, scattering is neglected in the forward model and hence atmospheric scattering properties do not need to be calculated.

To summarize, the forward model needs the following inputs:

- Surface pressure to define the equidistant pressure grid
- Sub-columns of CH_4, CO, H_2O, O_2, and air for the vertical layers of the model atmosphere.
- Pressure and temperature at the middle of the model sub-layers for absorption cross-sections.
- Solar Zenith Angle (SZA).
- Viewing Zenith Angle (VZA).
- Relative Azimuth Angle (RAA).
- The aerosol complex refractive index $m = m_{\mathrm{r}} + i\ m_{\mathrm{i}}$
- A high spectral resolution solar reference spectrum.
- Lookup tables with absorption cross-sections of CH_4, CO, H_2O, and O_2 as function of pressure, temperature, and wavenumber.

Based on these inputs the optical properties can be calculated for each layer of the model atmosphere.

3.2.2 Modeling the top-of-atmosphere radiances

For the proxy method the top of the atmosphere radiance can be modeled using Lambert Beers' law and surface reflection:

$$I_{TOA} = R_{\mathrm{surf}} F_0 e^{-(\frac{\tau}{\mu_0}+\frac{\tau}{\mu_v})} \quad (10)$$

where F_0 is the incoming total flux, $\tau = \tau_{\mathrm{abs}}$, μ_0 is the cosine of the solar zenith angle, μ_v is the cosine of the viewing zenith angle, and R_{surf} is the surface reflection for the specific solar and viewing geometry under consideration.

3.3 Inverse algorithm

Definition of state vector and ancillary parameters are given as follows. The **state vector x** contains the following elements (between brackets are optional elements):

- CO_2 sub-columns in 12 vertical layers (layer interfaces coincide with $NLAY$ layers of forward model grid).
- CH_4 sub-columns in 12 vertical layers (layer interfaces coincide with $NLAY$ layers of forward model grid).
- H_2O total column.
- Lambertian surface albedo in all bands band.
- First order spectral dependence of surface albedo in all bands.
- Spectral shift of Earth radiances in all bands (higher orders optional).
- Intensity offset in all bands (in the most recent GOSAT-1 retrieval no intensity offset is used).*
- (Offset in input temperature profile.)
- (Surface pressure.)

***For convergence it was essential to include intensity offsets to each of the individual spectral windows. Potentially there are still irregularities in the quality of the individual spectral bands.**

The **ancillary parameter vector b** contains the following parameters:

- H_2O sub-columns in 36 vertical layers of forward model grid.
- Temperature vertical profile at 72 layers of cross-section vertical grid.
- Pressure vertical profile at 72 layers of cross-section vertical grid.

Table 3: A priori values for the different state vector elements.

State vector element	**A priori value**
CH_4 sub-columns	TM4
CO2 sub-columns	Carbontracker
H_2O total column	ECMWF
surface albedo (NIR + SWIR)	no prior value needed (first guess at maximum of measured reflectance)
spectral shifts	no prior needed (first guess = 0)
temperature offset	no prior needed (first guess = 0)
surface pressure	ECMWF + SRTM DEM

3.4 Inversion Procedure

The inverse method optimizes the state vector $\mathbf{x}$ with respect to the measurements $\mathbf{y}$ after applying the forward model $\mathbf{F}$ to $\mathbf{x}$. The inverse method is based by default on a Phillips-Tikhonov regularization scheme (*Phillips, 1962; Tikhonov, 1963; Hasekamp and Landgraf, 2005*). Regularization is required because the inverse problem is ill-posed, i.e., the measurements $\mathbf{y}$ typically contain insufficient information to retrieve all state vector elements independently. The inverse algorithm finds $\mathbf{x}$ by minimizing the cost function that is the sum of the least-squares cost function and a side constraint weighted by the regularization parameter γ according to

$$\hat{\boldsymbol{x}} = \min_{x}(\,||\mathbf{S}_y^{-\frac{1}{2}}(\mathbf{F}(\mathbf{x}) - \mathbf{y})||^2 + \gamma||\mathbf{W}(\mathbf{x} - \mathbf{x}_a)||^2) \tag{11}$$

where $\mathbf{S}_y$ is the diagonal measurement error covariance matrix, which contains the noise estimate. $\mathbf{x}_a$ is an a priori state vector (see Table 3), and $\boldsymbol{W}$ is a weighting matrix (see below).

For the linearized forward model for iteration step n, the equation for the updated state vector $\mathbf{x}_{n+1}$ reduces to

$$\mathbf{x}_{n+1} = \min_{x}\{\|\mathbf{K}'(\mathbf{x}' - \mathbf{x}'_n) - \mathbf{y}'\|^2 + \gamma\|\mathbf{x}' - \mathbf{x}'_a\|\} \tag{12}$$

with the weighted quantities $\mathbf{x}' = \mathbf{W}\mathbf{x}$, $\mathbf{y}' = \mathbf{S}_y^{-\frac{1}{2}}(\mathbf{y}\text{-}\mathbf{F}(\mathbf{x}_n))$, and $\mathbf{K}' = \mathbf{S}_y^{-\frac{1}{2}}\mathbf{K}\mathbf{W}^{-\mathbf{1}}$.

The solution reads

$$\mathbf{x}_{n+1} = \mathbf{G}'\mathbf{y}' + \mathbf{A}'\mathbf{x}'_n + (\mathbf{I} - \mathbf{A})\,\mathbf{x}'_{\mathrm{apr}} \tag{13}$$

with $\mathbf{A}'$ the averaging kernel matrix and $\mathbf{G}'$ the contribution function matrix given by $\mathbf{A}' = \mathbf{G}'\mathbf{K}'$ and $\mathbf{G}' = (\mathbf{K}'^T\mathbf{K}' + \gamma\mathbf{I})^{-\mathbf{1}}\mathbf{K}'^T$. If the retrieval converges after a given number of steps N (typically 7-8), the final state vector $\mathbf{x}_{\mathrm{ret}} = \mathbf{x}_N$ is related to the true state vector and to the prior via

$$\mathbf{x}_{\mathrm{ret}} = \mathbf{A}\,\mathbf{x}_{\mathrm{true}} + (\mathbf{I} - \mathbf{A})\,\mathbf{x}_{\mathrm{a}} + \mathbf{G}\mathbf{e}_y + \mathbf{G}\mathbf{e}_f \tag{14}$$

The covariance matrix $\mathbf{S}_x$ describing the retrieval noise $\mathbf{G}\mathbf{e}_y$ is given by

$$\mathbf{S}_x = \mathbf{G}\mathbf{S}_y\boldsymbol{G}^T \tag{15}$$

The target retrieval quantity is the column averaged dry air CH_4 mixing ratio, XCH_4, This quantity is obtained from the CH_4 and CO_2 entries of the retrieved state vector through

$$\mathrm{XCH_4} = \frac{V\mathrm{CH_4}}{V\mathrm{CO_2}} \times \mathrm{XCO_{2,prior}} \tag{16}$$

where $V\mathrm{CH_4}$ and $V\mathrm{CO_2}$ are the total columns of CO_2 and CH_4, respectively, obtained from the retrieval state vector by

$$V_{\mathrm{GHG}} = \boldsymbol{h}_{\mathrm{GHG}}{}^{T} \mathbf{x}_{\mathrm{ret}} \tag{17}$$

where the subscript GHG refers to either CO_2 or CH_4, $\boldsymbol{h}$ is the total column operator (summing up the partial columns in the state vector). The retrieval noise ΔXCH_4 is given by

$$\Delta\mathrm{XCH_4} = \frac{\sum_{i,j=1}^{12} S_{x,i,j}}{V_{\mathrm{air,dry}}} \tag{18}$$

This is the error estimate that will be given in the output together with XCH_4.

For validation and application purposes it is important to realize that the retrieved XCH_4 is in fact a representation of $\mathbf{ax}_{\mathrm{true}}/V_{\mathrm{air,dry}}$, where the quantity

$$\mathbf{a} = \mathbf{h}^{T}\mathbf{A} \tag{19}$$

is referred to as the column averaging kernel (*Rodgers and Connor, 2003*)

3.4.1 Regularization of state vector and iteration strategy

To retrieve a meaningful state vector $\mathbf{x}$, the side-constrain in the minimization equation should be chosen in a way that contributions from measurement noise are minimized while retaining all valuable information in the first part of the merit-function. The inverse algorithm relies on a regularized Phillips-Tikhonov scheme. The diagonal weighting matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \mathbf{LW}' \tag{20}$$

with

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & \cdots & & & & \\ -1 & 2 & -1 & & & & \\ \vdots & & \ddots & & \vdots & & \\ & & -1 & 1 & & & \\ & & \cdots & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \tag{21}$$

The upper left 12 by 12 sub matrix works on the state vector elements that contain the CH_4 sub-columns in the 12 altitude layers of the retrieval vertical grid. This sub-matrix corresponds to the matrix product

$$\mathbf{L}_1^T\mathbf{L}_1 \tag{22}$$

where $\mathbf{L}_1$ is the first derivative matrix. The $\mathbf{W}'_{jj}$ is given by $1/\mathrm{MAX}[\mathrm{K}(1{:}12,1{:}12)]$ for the state vector elements corresponding to the 12 sub-columns of methane, $1/\mathrm{MAX}[\mathrm{K}(:,j)]$ for the aerosol parameters, and 0 for all other parameters (which means they are not constrained by the side constraint and are retrieved in a least-squares sense). The value for γ is fixed such that the Degrees of Freedom for Signal (DFS) for the methane profile is in the range 1.0-1.5. This value is found empirically

3.5 Filtering

3.5.1 Cloud Filtering

The purpose of cloud filtering here is to select reasonably good pixels to be processed by the Full Physics retrieval. In this way the thresholds should not be too strict. The cloud filtering is based on retrieved columns of VO_2, VCO_2 and VH_2O retrieved independently from the 0.75 µm, 1.6 µm, and 2.0 µm bands, respectively, under the assumption of a non-scattering atmosphere:

$$0.0 < \frac{\mathrm{VO}_{2,\mathrm{re}}}{\mathrm{VO}_{2,\mathrm{ECMWF}}} < 2.0 \tag{23}$$

$$0.0 < \frac{\mathrm{VCO}_{2,1.6\mu\mathrm{m}}}{\mathrm{VCO}_{2,2.0\mu\mathrm{m}}} < 2.0 \tag{24}$$

$$0.0 < \frac{\mathrm{VH_2O}_{1.6\mu\mathrm{m}}}{\mathrm{VH_2O}_{2.0\mu\mathrm{m}}} < 2.0 \tag{25}$$

The rationale for these cloud filters is that scenes with a large light path deviation with respect to a non scattering atmosphere will result in different CO_2 and H_2O columns retrieved (without scattering) from the 1.6 and 2.0 µm band due to different light path sensitivities in the two bands. Also, the retrieved O_2 column will deviate more from the ECMWF O_2 column for large light path differences with a non-scattering atmosphere. Compared to the Full Physics, the ratio retrieved in the proxy method is less sensitive to cloud contamination.

3.5.2 Posteriori Retrieval Filtering

The identification of which retrievals are of sufficient quality to be used by users is a fundamental step. As such, after the calculation of XCH_4 via equation 16, a post-processing quality filtering step is applied to create a quality flag to guide the user in which retrievals to take with confidence.

To this end, a random forest classifier model is used to predict the quality of the retrievals based on a selection of retrieval parameters, such as the cirrus signal, intensity offset, slope of the continuum etc. The full description of the model and its use can be found in the Product User Guide (*PUG_CH4_GO2_SRPR, 2023*).

The random forest model is trained on GOSAT-2 colocations with TCCON, where retrievals are classified via the bias of XCH_4. Retrievals in the training set are flagged as good quality if the XCH_4 bias is within a certain range and those outwith this range are flagged as bad quality. The model then learns the relation between the quality of the retrieval and the selection of retrieval parameters which it uses to predict the quality of all future data.

This technique improves on applying the filtering threshold criteria used in v2.0.0, which is a more static approach and does not fully capture all inter-dependencies in the data, resulting in many retrievals being filtered out when they should be kept in. In this new version (v2.0.2) we achieve a data yield increase of 13 %, and simultaneous improvement in the single measurement precision of 2 %, with respect to v2.0.0.

The random forest model is trained on TCCON data and consequently retrievals with surface albedo $\gtrsim 0.4$ are absent from the training sample. To circumvent this we define a set of high albedo data to include in the training set following the threshold criteria of v2.0.0. Furthermore the random forest method is limited only to land retrievals due to the low number of TCCON-GOSAT-2 colocations over ocean. For ocean measurements we also apply the threshold criteria in v2.0.0 for filtering. We find an excellent correlation and low bias between v2.0.0 and v2.0.2 for land retrievals, meaning that training on TCCON data does not produce obvious biases when extrapolating the filtering to global scales.

4 Output data

The output data are stored in one NetCDF file per day. The file size varies between 1 and 5 Mb. The format of the main output data, which are the Level 2 data products, is described in the separate document (*PUG_CH4_GO2_SRPR, 2023*).

References

Butz et al., 2009: Butz, A., Hasekamp, O. P., Frankenberg, C., Aben, I.: Retrievals of atmospheric CO2 from simulated space-borne measurements of backscattered near-infrared sunlight: Accounting for aerosol effects. Appl. Opt., 48, 3322, 2009.

Butz et al., 2010: Butz, A., Hasekamp, O. P., Frankenberg, C., Vidot, J., and Aben, I.: CH4 retrievals from space-based solar backscatter measurements: Performance evaluation against simulated aerosol and cirrus loaded scenes, J. Geophys. Res., 115, D24302, doi:10.1029/2010JD014514, 2010.

Butz et al., 2011: Butz, A., Guerlet, S., Hasekamp, O., et al., Toward accurate CO_2 and CH_4 observations from GOSAT, *Geophys. Res. Lett.*, doi:10.1029/2011GL047888, 2011.

Chevallier, 2019: Chevallier, F.: Validation report for the CO2 fluxes estimated by atmospheric inversion, v19r1Version 1.0, Copernicus Atmosphere Monitoring Service, ECMWF Copernicus report, CAMS73_2018SC2_D73.1.4.1-2019_v3_202008_Validation inversion CO2 fluxes_v1, https://atmosphere.copernicus.eu/sites/default/files/2020-08/CAMS73_2018SC2_D73.1.4.1-2019-v1_202008_v3-1.pdf

Chevallier, 2020: Chevallier, F.: Validation report for the CO2 fluxes estimated by atmospheric inversion, v20r2Version 1.0, Copernicus Atmosphere Monitoring Service, ECMWF Copernicus report, CAMS73_2018SC2_D73.1.4.1-2020_v5_202109_Validation inversion CO2 fluxes_v1, https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-GHG/CAMS73_2018SC2_D73.1.4.1-2020-v5_202109_v1.pdf

Chevallier, 2021: Chevallier, F.: Validation report for the CO2 fluxes estimated by atmospheric inversion, v21r1Version 2.0, Copernicus Atmosphere Monitoring Service, ECMWF Copernicus report, CAMS255_2021SC1_D55.1.4.2-2022_202208_Validation inversion CO2 fluxes_v2, https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/CAMS255_2021SC1_D55.1.4.2-2022_202208_Validation%20inversion%20CO2%20fluxes_v2_.pdf

Coddington et al., 2021: Coddington, O. M., Richard, E. C., Harber, D., Pilewskie, P., Woods, T. N., Chance, K.: The TSIS-1 Hybrid Solar Reference Spectrum. Geophysical Research Letters, 48, e2020GL091709. https://doi.org/10.1029/2020GL091709, 2021.

Guerlet et al., 2013: Guerlet, S., S. Basu, A. Butz, M. Krol, P. Hahne, S. Houweling, O. P. Hasekamp and I. Aben, Reduced carbon uptake during the 2010 Northern Hemisphere summer from GOSAT, Geophys. Res. Lett., doi: 10.1002/grl.50402, 2013.

Hasekamp and Landgraf, 2005: Hasekamp, O. P., and Landgraf, J.: Linearization of vector radiative transfer with respect to aerosol properties and its use in satellite remote sensing, J. Geophys. Res., 110, D04203, doi:10.1029/2004JD005260, 2005.

Laughner et al., 2021: Laughner, J., Andrews, A., Roche, S., Kiel, M., and Toon, G._ginput v1.0.10: GGG2020 prior profile software (Version 1.0.10), https://doi.org/10.22002/D1.1944, 2021.

Nakajima et al., 2017: M. Nakajima, Hiroshi Suto, K. Yotsumoto, K. Shiomi, T. Hirabayashi, "Fourier transform spectrometer on GOSAT and GOSAT-2," Proc. SPIE 10563, International Conference on Space Optics — ICSO 2014, 105634O (17 November 2017); https://doi.org/10.1117/12.2304062.

O'Brien et al., 2013: O'Brien, D. M., Polonsky, I., O'Dell, C., Kuze, A., Kikuchi, N., Yoshida, Y., Natraj, V.: Testing the Polarization Model for TANSO-FTS on GOSAT Against Clear-Sky Observations of Sun Glint Over the Ocean, IEEE Trans. Geosci. Remote. Sens, vol. 51, no. 12, pp. 5199-5209, 2013, doi: 10.1109/TGRS.2012.2232673.

Phillips, 1962: Phillips, D.L.: A technique for the numerical solution of certain integral equations of the first kind. Journal of the ACM (JACM), 9(1), pp.84-97, 1962.

PUG_CH4_GO2_SRPR, 2023: Barr, A.G., and the GHG-CCI+ group at SRON: ESA Climate Change Initiative "Plus" (CCI+) Product User Guide (PUG) Version 1.0 for the RemoTeC XCH4 GOSAT-2 PROXY Product v2.0.2 (CH4_GO2_SRPR) for the Essential Climate Variable (ECV) Greenhouse Gases (GHG), ESA GHG CCI+, 2023.

Rodgers and Connor, 2003: Rodgers, C. D., and Connor, B. J.: Intercomparison of remote sounding instruments, J. Geophys. Res., 108, 4116, doi:10.1029/2002JD002299, D3, 2003.

Schepers et al., 2012: Schepers, D., Guerlet, S., Butz, A., Landgraf, J., Frankenberg, C., Hasekamp, O., Blavier, J.F., Deutscher, N.M., Griffith, D.W.T., Hase, F. and Kyro, E.: Methane retrievals from Greenhouse Gases Observing Satellite (GOSAT) shortwave infrared measurements: Performance comparison of proxy and physics retrieval algorithms, J. Geophys. Res., 117, D10307, doi:10.1029/2012JD017549, 2012.

Tikhonov, 1963: Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Math., 4, pp.1035-1038, 1963.

Wunch et al., 2010: Wunch, D., Toon, G. C., Wennberg, P. O., Wofsy, S. C., Stephens, B. B., Fischer, M. L., Uchino, O., Abshire, J. B., Bernath, P., Biraud, S. C., Blavier, J.-F. L., Boone, C., Bowman, K. P., Browell, E. V., Campos, T., Connor, B. J., Daube, B. C., Deutscher, N. M., Diao, M., Elkins, J. W., Gerbig, C., Gottlieb, E., Griffith, D. W. T., Hurst, D. F., Jiménez, R., Keppel-Aleks, G., Kort, E. A., Macatangay, R., Machida, T., Matsueda, H., Moore, F., Morino, I., Park, S., Robinson, J., Roehl, C. M., Sawa, Y., Sherlock, V., Sweeney, C., Tanaka, T., and Zondlo, M. A.: Calibration of the Total Carbon Column Observing Network using aircraft profile data, Atmospheric Measurement Techniques, 3, 1351–1362, doi:10.5194/amt-3-1351-2010, URL http://www.atmos-meas-tech.net/3/1351/2010/, 2010.

Wunch et al., 2015: Wunch, D., Toon, G.C., Sherlock, V., Deutscher, N.M., Liu, X., Feist, D.G., Wennberg, P.O., The Total Carbon Column Observing Network's GGG2014 Data Version. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA (available at: doi:10.14291/tccon.ggg2014.documentation.R0/1221662), 2015.