EOF-based regression algorithm for the fast retrieval of atmospheric CO_2 total column amount from the GOSAT observations

Andrey Bril[a,*], Shamil Maksyutov[b], Dmitry Belikov[c], Sergey Oshchepkov[a], Yukio Yoshida[b], Nicholas M. Deutscher[d,e], David Griffith[f], Frank Hase[g], Rigel Kivi[h], Isamu Morino[b], Justus Notholt[e], David F. Pollard[i], Ralf Sussmann[j], Voltaire A. Velazco[d], Thorsten Warneke[e]

[a] *Institute of Physics of National Academy of Sciences of Belarus (IPNASB), 68, Nezavisimosti Ave., 220072 Minsk, Belarus*
[b] *Center for Global Environmental Research, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan*
[c] *Tomsk State University, 36 Lenin Ave., Tomsk 634050, Russia*
[d] *Centre for Atmospheric Chemistry, School of Chemistry, University of Wollongong, Wollongong, NSW 2522, Australia*
[e] *Institute of Environmental Physics, University of Bremen, Bremen 28334, Germany*
[f] *School of Chemistry, Northfields Ave., University of Wollongong, NSW 2522, Australia*
[g] *Karlsruhe Institute of Technology, IMK-ASF, Karlsruhe 76344, Germany*
[h] *FMI-Arctic Research Center, Tähteläntie 62, FIN-99600 Sodankylä, Finland*
[i] *National Institute of Water and Atmospheric Research (NIWA), Private Bag 50061, Omakau, Central Otago, New Zealand*
[j] *Karlsruhe Institute of Technology, IMK-IFU, Garmisch-Partenkirchen 82467, Germany*

ARTICLE INFO

Keywords:
Carbon dioxide
Retrieval algorithm
Empirical orthogonal function
GOSAT
TCCON

ABSTRACT

This paper presents a novel retrieval algorithm for the rapid retrieval of the carbon dioxide total column amounts from high resolution spectra in the short wave infrared (SWIR) range observations by the Greenhouse gases Observing Satellite (GOSAT). The algorithm performs EOF (Empirical Orthogonal Function) based decomposition of the measured spectral radiance and derives the relationship of limited number of the decomposition coefficients in terms of the principal components with target gas amount and *a priori* data such as airmass, surface pressure, etc. The regression formulae for retrieving target gas amounts are derived using training sets of collocated GOSAT and ground based observations. The precision/accuracy characteristics of the algorithm are analyzed by the comparison of the retrievals with those from the Total Carbon Column Observing Network (TCCON) measurements and with the modeled data, and appear similar to those achieved by full physics retrieval algorithms.

1. Introduction

Long term experience using GOSAT (Greenhouse gases Observing Satellite) observations has shown promising prospects and benefits of carbon dioxide satellite remote sensing for estimating regional CO_2 fluxes [1,2]. An important part of the GOSAT mission is the development of the retrieval algorithms that combine measured spectral data with available a priori information to estimate column averaged dry volume CO_2 mixing ratios (XCO_2) [3 8]. These algorithms are continuously upgraded in order to improve their productivity or yield (number of valid retrievals), precision/accuracy characteristics, computation efficiency, etc. However, the quality of the satellite based atmospheric CO_2 data is still criticized [9] implying the need to continue to improve retrieval algorithms.

New greenhouse gas observing missions, such as OCO 2 (Orbiting Carbon Observatory 2) [10], and forthcoming missions, such as TanSat (Carbon Satellite: Tan means "carbon" in Chinese) [11] and GOSAT 2 [12] face new challenges in satellite based data processing including the development of very fast retrieval procedures to cope with huge data amounts.

In this paper we propose a very rapid retrieval algorithm, which is based on the decomposition of the spectral radiance of the reflected solar radiation by using empirical orthogonal functions (EOF). This algorithm has been implemented and tested employing GOSAT observations.

EOF methodology is a multipurpose tool that is known to be widely used in atmospheric science, e.g. for the extraction of the characteristic patterns from high resolution spectral data [13]. An EOF based approach was used for retrievals of the atmospheric methane profiles from the Atmospheric Infrared Sounder (AIRS)

* Corresponding author.
E-mail address: andrey.bril@gmail.com (A. Bril).

thermal infrared spectra [14]. The possibility of applying an EOF application to CO_2 retrievals from the GOSAT measurements in 1.6 μm CO_2 absorption band was demonstrated in [15]. However, information from the reflected sunlight radiance spectra in only the 1.6 μm band is generally insufficient for accurate CO_2 retrievals due to optical path modification by aerosol and clouds [4]. As a rule, to account for the optical path modifications we need additional near infrared GOSAT measurements in 2.06 μm CO_2 and in 0.76 μm O_2 absorption bands. Also, we need to find a way to include available a priori data and measurement conditions when using the EOF methodology.

The paper is organized as follows. Section 2 introduces the methodology and software for EOF decomposition of radiance spectra. In Section 3, we briefly outline the implementation of the EOF approach to GOSAT data processing using all available near infrared bands as well as *a priori* information. Section 4 describes the validation of the retrieved XCO_2 using ground based observations and the modeled data. Section 5 summarizes the results.

2. EOF-decomposition technique

Typical sets of sampled high resolution radiance spectra that serve as background data for atmospheric CO_2 retrievals are overabundant (hundreds or thousands of data points) to be used in regression based algorithms. Large number of data is beneficial for reducing random retrieval errors, but on the other hand the variability of the observed spectra is controlled to large extent by very limited number of parameters such as column abundance of major trace constituents, optical path, temperature and others. Thus, reduction of the degrees of freedom via Principal Component Analysis (PCA) is expected to be effective. To this end the spectral radiance or, for better linearity, the normalized logarithm of spectral radiance can be expressed as a linear combination of Empirical Orthogonal Functions (EOF)Ψ:

$$R_{l,\nu} = \sum \varepsilon_{l,m}\Psi_{m,\nu},\ l = 1, 2, \dots, L \text{ and } \nu = 1, 2, \dots, N \tag{1}$$

Or in matrix form

$$R = E{\cdot}\Psi, \tag{1a}$$

where l is the number of the observation, ν is the number of spectral channel, and E is matrix of weighting coefficients. The index m ranges from 1 to M, where $M = \min(L, N)$.

Standard procedures of EOF decomposition are usually implemented in tune with Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) that yields ranged weighting coefficients (first coefficient accounts for maximal R variability). This facilitates the selection of a limited set of the weighting coefficients to approximate the original function, which can be used to build the regression relations.

In this study we used subroutine LSVRR from the IMSL library (http:// www.roguewave.com/products services/imsl numerical libraries) that implements the SVD based algorithm briefly outlined below [16].

It is known that for any $L \times N$ real matrix R there exists an L x L orthogonal matrix U and a $N \times N$orthogonal matrix V such that

$$U^TRV = \Sigma \tag{2}$$

where Σ is diagonal matrix, i.e. $\Sigma = diag(\sigma_1, \dots, \sigma_m)$, and $m = \min(L, N)$. The scalars $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$ are called as the singular values of R.The columns of U are called the left singular vectors of R. The columns of V are the right singular vectors of R.

By multiplying (2) by $\boldsymbol{U}$ (left) and by V^{-1} (right) and accounting for fact that $UU^T$is a unity matrix we obtain

$$R = U\Sigma V^{-1} \tag{3}$$

By denoting $\Psi = V^{-1} = V^T$(EOF) and $E = U\Sigma$(matrix of weighting coefficients) we can rewrite Eq. (3) in the form of (1a).

3. Implementation of EOF decomposition for GOSAT data processing

The proposed algorithm for fast estimates of atmospheric XCO_2 includes the following steps:

- extraction of the compact information from the measured spectral radiance by its EOF decomposition, followed by
- combining the extracted data (weighting coefficients of the decomposition) with some available input or *a priori* information; and
- derivation of regression formulae that relate this combined information with target gas amounts using training sets of collocated GOSAT and ground based reference observations.

3.1. Reference bases for the EOF decomposition

The reference orthogonal bases $\Psi = V^T$were created for the three spectral regions that were selected for XCO_2 retrieval from GOSAT observations [8]. These regions include

(1) 6180 cm^{-1} 6270 cm^{-1} from TANSO FTS Band 2,
(2) 4815 cm^{-1} 4885 cm^{-1} from TANSO FTS Band 3,
(3) 13000 cm^{-1} 13090 cm^{-1} from TANSO FTS Band 1 (auxiliary spectral region used for the atmospheric correction).

Given the GOSAT spectral sampling interval in the near infrared, which is approximately 0.2 cm^{-1}, the number of available spectral channels is about 450 in spectral regions 1 and 3 ($N_{(1)} \approx N_{(3)} \approx 450$); and about 350 in spectral region 2 ($N_{(2)} \approx 350$).

For the construction of the "measured signal" R, we used the scalar spectral radiance S that was generated by NIES operational algorithm for CO_2 retrievals [17]. This scalar radiance was computed from P and S signal polarizations provided by the Japan Aerospace Exploration Agency (JAXA) within the L1B product [18].

For spectral region (1), we defined R as

$$R_{(1)} = \frac{-\ln S_{(1)} + \ln S_{(1)}^{\max}}{A}, \tag{4}$$

where $S_{(1)}^{\max}$ is the maximal value of scalar radiance in the spectral range (1) and airmass A is defined as

$$A = 1/\cos(\theta_0) + 1/\cos(\theta_1), \tag{5}$$

where θ_0 and θ_1 are solar and satellite zenith angles, respectively.

To construct the linear regression, we expect some advantages when using logarithm of radiance instead of absolute radiance values because the logarithm provides more linear dependence of R on optical thickness and XCO_2. In the GOSAT data processing, radiance S is obtained by transforming the interferograms measured by TANSO FTS on board GOSAT, and the apodization effect can result in non physical negative radiance values. This usually happens in case of deep absorption lines that are typical for spectral regions (2) and (3). For this reason, instead of using a logarithm transformation we defined R for the spectral regions (2) and (3) as follows

$$R_{(2,3)} = \frac{S_{(2,3)}}{S_{(2,3)}^{\max}}. \tag{6}$$

To generate a data set for creating the reference orthogonal bases, we applied Cloud and Aerosol Imager (CAI) cloud flag screening in the same way as for the NIES L2 processing [6,18]. The CAI screening procedure was designed to remove TANSO FTS observations contaminated with optically thick clouds. Typically CAI screened sets include about 40,000 to 50,000 observations per month. These sets could include observations that were taken in the presence of the sub visual cirrus clouds or optically thick aerosols, which might require further data screening. In this study, we use only over land observations (reducing the original data set by 50% or more) for four months (January, April, July, and October) representing different seasons in 2010 and 2012. Additionally, we skip "noisy" data with a Signal to Noise Ratio (SNR) below 75 in at least one of the three TANSO FTS bands.

GOSAT CAI screened observations are non uniformly distributed over latitude L. For example, in January 2010 the latitudinal zone $15^0 \le L < 30^0$includes more than 4000 over land observations, whilst the zone $45^0 \le L < 60^0$ includes only 18 similar observations. Using such distributed data for basis creation involves a risk that "sparse regions" would weakly affect the derived weighting coefficients (which in turn could result in poor XCO_2 approximations for these regions). In addition, using standard software for EOF decomposition imposes limitations on the size of the data set for the basis construction. With this in mind, we created the required data set in two steps:

- First, we divided the globe into latitudinal zones of 15^0 width and selectively eliminated part of the observations to reduce data amount and balance data distribution over latitude. This results in the data set of about 25,000 observations taken within eight months that represent four seasons of 2010 and 2012;
- Further reduction (e.g. to reduce computational costs and time consumption) was performed by simple selection of each n th observation in the chronologically ranged set. In particular, for reference orthogonal basis computation we used reduced data set of about 5000 observations (n=5). The locations of these observations are shown in Fig. 1. The spectral radiances within this data set were reduced to the unified wavenumber grids by the spline based interpolation.

With this compact data selection, we created sets of reference EOFs for each spectral band

$$\Psi_{(k)} = V^T_{(k)},\ k = 1,\ 2,\ 3, \tag{7}$$

that should be representative for XCO_2 retrievals. As a result any spectral signal can be expressed in terms of reference EOFs with weighting coefficients defined by

$$E_{(k)} = R_{(k)} \cdot \Psi^T_{(k)} = R_{(k)} \cdot V_{(k)}. \tag{7a}$$

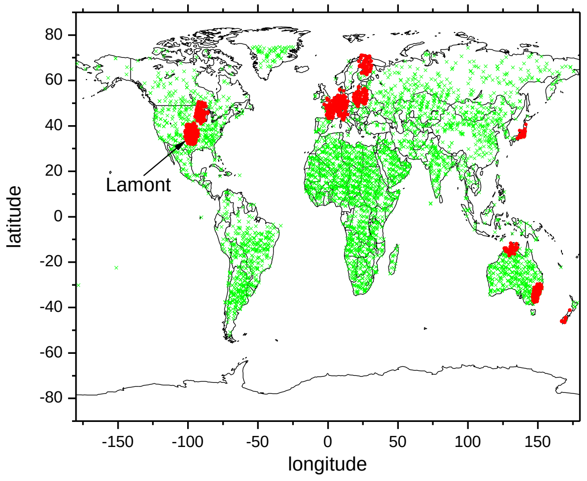

Fig. 1. Global locations of the GOSAT observations (footprints) that were chosen to create reference bases for the EOF decomposition, Section 3.1 (crosses) and the training Subsets 1 and 2, Section 3.3 (solid circles).

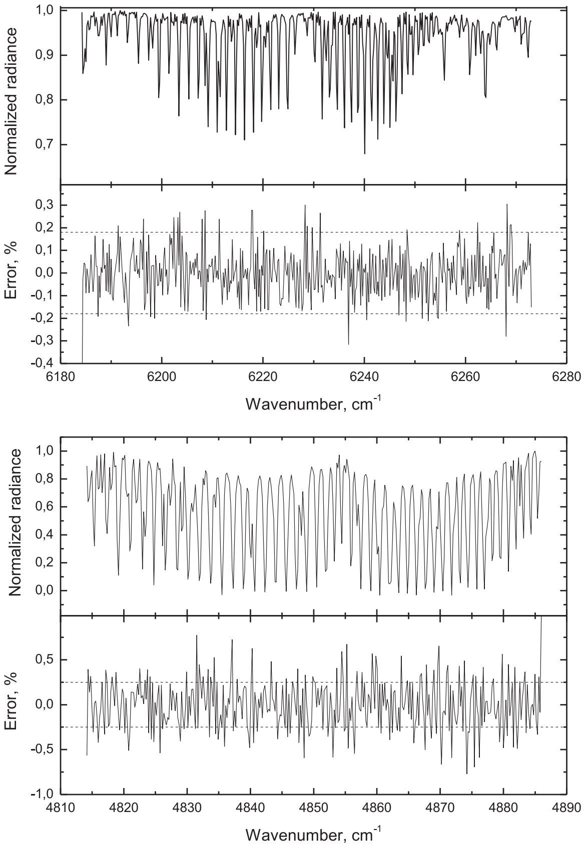

Fig. 2. Examples of normalized radiance spectra and EOF-approximation errors for spectral regions (1) (upper panels) and (2) (lower panels). Dashed lines indicate the noise levels, which are estimated as 1/SNR in the spectral regions.

The number of weighting coefficients for EOF decomposition of the individual observations is limited by the numbers of spectral channels $N_{(j)}, j=1, 2, 3$ giving a total of about 1250 for all spectral regions. Assuming this number is excessive, we use first $M_{(k)}$ ($k=1$, 2, 3) of ranged (i.e. maximal) coefficients for each k th spectral region. In this study we have empirically chosen the following values: $M_{(1)} = 35$ and $M_{(2)} = M_{(3)} = 20$.

These limited numbers of weighting coefficients provide a reasonably accurate fit of the original radiance spectra by the EOF decomposition. Fig. 2 shows typical radiance spectra (normalized to the maximal values) in spectral regions (1) and (2) as well as similarly normalized approximation errors, that are mostly below (spectral region (1)) or comparable (spectral region (2)) with the observation noise levels.

3.2. Construction of the generalized vector of weighting coefficients

For each observation the generalized vector of weighing coefficients consists of first $M_{(k)}$ weighting coefficients for three spectral regions as well as of P pieces of input or *a priori* information for this observation (e.g. observation geometry and/or meteorological conditions).

$$\tilde{E} = \left\{ E^1_{(1)}, \ldots, E^{M(1)}_{(1)}, E^1_{(2)}, \ldots, E^{M(2)}_{(2)}, E^1_{(3)}, \ldots, E^{M(3)}_{(3)}; \Pi_1, \ldots \Pi_P \right\} \tag{8}$$

This generalized vector is expected to include necessary information on XCO_2 that is extracted using "transformation vector" G

$$X_{CO2} = G{\cdot}\tilde{E} \tag{9}$$

Eq. (9) can be applied to any arbitrary number of observations. In the case of L observations, $X_{CO2}$is a vector of dimension L, G is a transposed vector of dimension $Q = M_{(1)} + M_{(2)} + M_{(3)} + P$, and $\tilde{E}$ is a matrix of dimension $Q \times L$: each l th row of this matrix is the generalized vector of weighting coefficients for the l th observation.

As a priori or input information $\Pi_1, \Pi_2, \dots, \Pi_P$ in the Eq. (8) we use airmass A (Π_1), surface pressure $P_S(\Pi_2)$, and *a priori* XCO_2 value (Π_3). To account for the non linear radiance dependence on A and P_S, we also included their squared values $A^2(\Pi_4)$ and P_S^2 (Π_5). These values (Π_1 Π_3) were used partly by analogy with some tested retrieval algorithms [3,8] that explicitly include them in the retrieval procedure to provide accurate XCO_2 estimates. For example, air mass is required to determine target gas optical depth and surface pressure is used to transform total column amount of CO_2 to dry volume mixing ratio (X_{CO2}).

A priori XCO_2 values were defined on the basis of "zonal" CO_2 volume mixing ratios. We compute zonal concentrations by longitudinal (from 0° to 360°) and latitudinal (from lower zone bound to upper zone bound) averaging of X_{CO2}, which were simulated by the NIES atmospheric transport model [19]. An equidistant latitudinal grid of 10° was used. We created zonal $\tilde{X}_{CO2}$for four months (January, April, July, and October) of 2010. Next we assume constant zonal XCO_2 within seasons (e.g. for all winter months we used January data, for all spring months we used April data, etc.). Interannual XCO_2 growth of 2 ppm per year was also included, i.e. zonal $\tilde{X}_{CO2}$ for arbitrary year YYYY was calculated as $\tilde{X}_{CO2}(YYYY) = \tilde{X}_{CO2}(2010) + 2 \times (YYYY - 2010)$.

3.3. *Training of the algorithm using reference ground based observations*

The training procedure includes

- EOF decomposition (Eq. (7a)) of the spectral radiance for all L observations within the training subset using predefined orthogonal matrices (Eq. (7)). The EOF decomposition was preceded by the spline based interpolation of the spectral radiances onto unified wavenumber grids that were used for the generations of the reference bases $\Psi_{(k)}$ (Section 3.1);
- Construction of the matrix $\tilde{E}_*$of dimension $Q \times L$ that includes L rows of the generalized vectors of weighting coefficients of length Q (Eq. (8)) for all observations of the training subset. The subscript * denotes "training subset".
- Determination of the "transformation vector" G from the condition of the best fit of X_{CO2} over the "training subset" of the observations for which XCO_2 values are somehow known

$$G = X_{*,CO2}{\cdot}\tilde{E}_*^T{\cdot}\left(\tilde{E}_*{\cdot}\tilde{E}_*^T\right)^{-1}. \tag{10}$$

For training and validation purposes we used TCCON ground based X_{CO2}observations [20 27]. TCCON XCO_2 measurements taken within ± 1 h of the GOSAT overpass time were chosen as the "known" value for the GOSAT observation if the footprint of the observation was located within a 5° latitude longitude circle around TCCON site. In this study we used data from 12 TCCON stations, Białystok (53.2 °N, 23.1 °E), Bremen (53.1 °N, 8.85 °E), Darwin (45.0 °S, 169.7 °E), Garmisch (47.5 °N, 11.1 °E), Karlsruhe (49.1 °N, 8.44 °E), Lamont (36.6 °N, 97.5 °W), Lauder (45.0 °S, 169.7 °E), Orléans (48.0 °N, 2.11 °E), Park Falls (45.9 °N, 90.3 °W), Sodankylä (67.4 °N, 26.6 °E), Tsukuba (36.0 °N, 140.2 °E), and Wollongong (34.4 °S, 150.9 °E), for the period from June 2009 to December 2012. We selected about 12,000 collocated GOSAT TCCON observations (including about 9000 over land observations) from the NIES L2 CAI screened data set [17]. These observations were non uniformly distributed among TCCON sites: the largest share of them is located around Lamont (mostly because of frequent requests for special observation mode for this site as well as of high percentage of clear sky conditions over Lamont). Two training subsets were created using only over land observation: Subset 1 included data around the Lamont site only; the second (Subset 2) was created with roughly balanced representations of different stations: for Lamont data we choose one of each five sequential observations (1 in 5 collocations); Wollongong (1:4) Garmisch, Karlsruhe, Orléans, Darwin, Park Falls, and Tsukuba (1:3); Białystok, Bremen, Lauder (1:2); Sodankylä (1:1). Both these training subsets include about 3200 scans. The global locations of these observations from both training subsets are shown in Fig. 1.

3.4. *Retrievals of XCO_2 and post screening procedure*

Provided that the transformation vector G is defined, the XCO_2 retrieval procedure for an arbitrary observation includes the construction of a generalized vector $\tilde{E}$, Eq. (8), for this observation and the application of Eq. (9) to compute XCO_2 value. The only output of the retrieval procedure is XCO_2, no information on retrieval uncertainty or averaging kernel is available.

Following tested XCO_2 retrieval algorithms [3 8], we also studied the possibility to improve retrieval quality by applying post screening procedures. In this study, the post screening was implemented by limiting the discrepancy between measured spectral radiance $S_{(1,2,3)}$ and its approximation by SVD decomposition$S^*_{(1,2,3)}$ with a limited number of the weighting coefficients, Eq. (9). The following expression for spectral region $k$was utilized to characterize the discrepancy

$$\tilde{\chi}^{(k)} = \frac{300^2}{N^{(k)}}\frac{\sum_{i=1}^{N^{(k)}}\left(S_{(k)} - S^*_{(k)}\right)^2}{\left(S^{\max}_{(k)}\right)^2}, \tag{11}$$

where $N^{(k)}$and $S^{\max}_{(k)}$ are the number of spectral channels and maximal value of the radiance, respectively. A numerical coefficient of 300 corresponds to designated signal to noise ratio for GOSAT observations [18].

4. Validation of the EOF-based XCO_2 retrievals

4.1. *Validation using TCCON data*

Fig. 3 and Table 1 show the comparison results of the GOSAT EOF retrievals and TCCON XCO_2 for 12 TCCON sites within the coincidence criterion. The figure shows the time series of the XCO_2 retrievals and the table presents key statistical characteristics of the GOSAT EOF (Y) and TCCON (X) XCO_2 relationship that include:

Bias:

$$Bias = \overline{(Y_i - X_i)}, \tag{12}$$

where the overline denotes averaging over coincident $N(i = 1, 2, \dots, N)$ observations assuming uniform errors in X and Y;

Standard deviation:

$$STD = \sqrt{\overline{(Y_i - X_i - Bias)^2}}; \tag{13}$$

Pearson's correlation coefficient

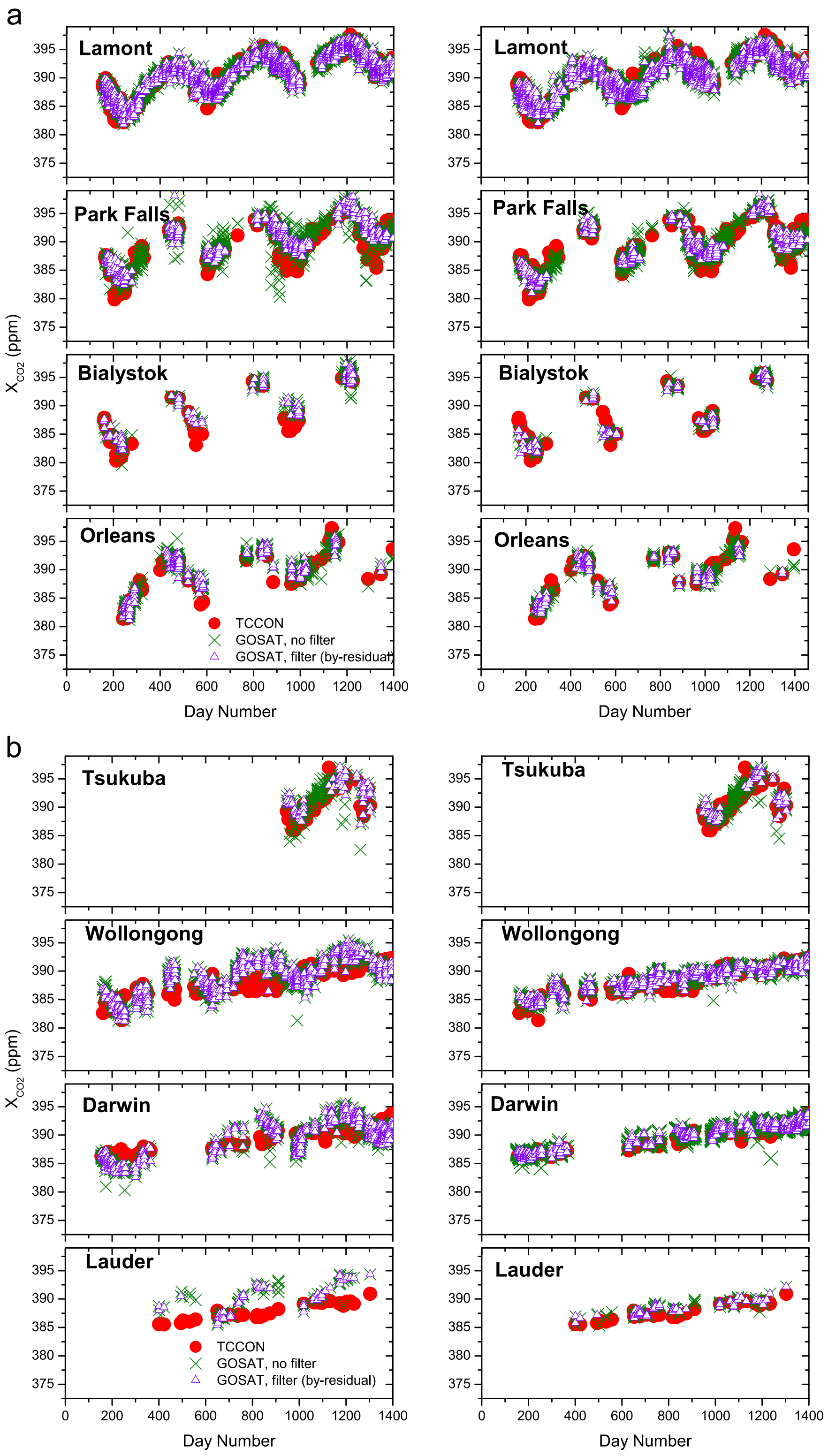

Fig. 3. a. GOSAT versus TCCON XCO_2 intercomparison results for the collocated observations around Lamont, Park Falls, Białystok, and Orléans in terms of time series. GOSAT retrievals were obtained with training Subsets 1 (left-hand panels) and 2 (right-hand panels), respectively. Both post-screened (open triangles) and non-filtered (crosses) GOSAT-EOF retrievals are shown versus TCCON data (solid circles); day number is counted from January 1, 2009. b. Same as in Fig. 3a but for Tsukuba, Wollongong, Darwin, and Lauder TCCON sites.

$$r = \frac{\overline{\left(X_i - \overline{X}_i\right)\left(Y_i - \overline{Y}_i\right)}}{\sqrt{\overline{\left(X_i - \overline{X}_i\right)^2 \left(Y_i - \overline{Y}_i\right)^2}}}, \tag{14}$$

and the linear regression slope (*Slope*). Deviations of the *Slope* from unity imply that the retrieval results fail to reproduce temporal and/or spatial variations of XCO_2 as compared to reference TCCON data.

The left hand panels of Fig. 3 present retrieval results for the training Subset 1 (using the Lamont site only) and right hand panels show the results for the training Subset 2 (selected

Table 1
Statistical characteristics of the GOSAT versus TCCON XCO_2 intercomparison. Subsets 1 and 2 corresponds to training Subsets 1 and 2 (Section 3.3). The Subset 3 includes XCO_2 retrievals by NIES operational algorithm, version v02.21; release level for General Users. N is the number of XCO_2 retrievals (yield). For Subsets 1 and 2 N is presented for the algorithm application without (no parentheses) and with (in parentheses) application of post-screening procedure, Section 3.4. Other comparable characteristics (mean bias, standard deviation *STD*, regression slope and correlation coefficient *r*) are defined in Section 4.1.

Site	Subset	N	Bias (ppm)	*STD* (ppm)	Slope	r
Białystok	1	204 (147)	0.54 (0.52)	1.40 (1.36)	0.90 (0.89)	0.96 (0.97)
	2	204 (147)	-0.30 (-0.36)	1.01 (1.01)	0.99 (0.98)	0.98 (0.98)
	3	134	-0.64	1.89	1.07	0.94
Bremen	1	111 (75)	0.27 (0.12)	1.63 (1.70)	0.99 (1.03)	0.90 (0.91)
	2	111 (75)	-0.58 (-0.67)	1.69 (1.91)	1.07 (1.12)	0.90 (0.90)
	3	68	-0.81	2.22	1.22	0.82
Darwin	1	648 (613)	-0.34 (-0.29)	2.29 (2.27)	1.66 (1.61)	0.67 (0.68)
	2	648 (613)	0.22 (0.25)	0.99 (0.97)	1.00 (0.99)	0.90 (0.90)
	3	256	-1.91	1. 60	1.35	0.84
Garmisch	1	574 (343)	1.28 (1.26)	1.43 (1.37)	1.05 (1.04)	0.92 (0.94)
	2	574 (343)	0.49 (0.60)	1.32 (1.22)	1.03 (1.00)	0.95 (0.95)
	3	313	0.08	2.35	1.26	0.82
Karlsruhe	1	569 (358)	0.28 (0.43)	1.50 (1.40)	0.85 (0.81)	0.90 (0.92)
	2	569 (358)	-0.77 (-0.63)	1.21 (1.16)	0.95 (0.92)	0.94 (0.94)
	3	345	-1.24	2.28	0.97	0.77
Lamont	1	3197 (2499)	-0.02 (-0.04)	1.06 (1.10)	0.95 (0.95)	0.95 (0.95)
	2	3197 (2499)	-0.45 (-0.45)	1.36 (1.41)	0.90 (0.87)	0.91 (0.91)
	3	2022	-1.97	1.81	1.10	0.87
Lauder	1	92 (71)	2.42 (1.97)	2.10 (2.09)	3.21 (3.17)	0.49 (0.51)
	2	92 (71)	0.64 (0.64)	0.74 (0.69)	1.00 (0.95)	0.82 (0.84)
	3	68	-0.98	1.88	2.56	0.70
Orléans	1	429 (278)	0.25 (0.41)	1.19 (1.17)	1.02 (1.02)	0.95 (0.95)
	2	429 (278)	-0.26 (-0.04)	0.98 (0.96)	0.93 (0.93)	0.96 (0.97)
	3	270	-1.40	2.18	1.12	0.84
Park Falls	1	1147 (527)	1.21 (1.64)	2.22 (1.90)	0.92 (0.85)	0.79 (0.87)
	2	1147 (527)	0.24(0.52)	1.62 (1.54)	0.91 (0.91)	0.89 (0.92)
	3	641	-0.41	2.39	1.32	0.85
Sodankylä	1	334 (43)	2.13 (0.79)	2.30 (1.64)	0.73 (0.90)	0.83 (0.81)
	2	334 (43)	0.18 (-1.06)	2.05 (1.72)	0.81 (1.12)	0.86 (0.82)
	3	210	-0.55	2.39	1.29	0.89
Tsukuba	1	174 (77)	0.78 (1.39)	2.23 (2.04)	1.08 (0.83)	0.64 (0.74)
	2	174 (77)	0.51 (0.98)	1.69 (1.66)	1.04 (0.99)	0.79 (0.84)
	3	102	1.52	3.17	1.96	0.56
Wollongong	1	926 (759)	0.87 (0.76)	2.49 (2.49)	1.62 (1.65)	0.58 (0.57)
	2	926 (759)	0.29 (0.31)	1.19 (1.16)	0.89 (0.89)	0.85 (0.85)
	3	707	-0.97	2.45	1.57	0.62

observations over 12 TCCON sites). As expected, with Subset 1 we have almost perfect XCO_2 retrievals over Lamont (in this case the retrieval procedure has been applied directly to the training set). However, the retrievals around other TCCON sites are much worse. In particular, for Northern Hemisphere sites such as Park Falls and Sodankylä both the bias and scatter (*STD*) of XCO_2 with respect to the "reference" TCCON data are large compared to the results of recently developed algorithms [7]. Moreover, using Subset 1 results in the transfer of Lamont like seasonal pattern to Southern Hemisphere regions (Darwin, Wollongong and Lauder sites) that produces noticeable false seasonal variations of the retrieved XCO_2. Additionally, the Southern Hemisphere retrievals are strongly biased and have rather large scatter. Unfortunately, post screening by limiting spectral discrepancy does not fix these drawbacks. Some better results hold when applying the post screening with chi squared test (Eq. 11) as follows

$$\tilde{\chi}^{(1)} \le 1;\ \tilde{\chi}^{(2)} \le 5;\ \tilde{\chi}^{(3)} \le 5. \qquad (15)$$

These limitations considerably reduced the number of "approved" observations: as seen in Table 1, we have two fold reduction for Park Falls site and about eight fold reduction for Sodankylä (statistical characteristics of post filtered results are shown in brackets). Additionally, we have some reduction of scatter. However, other statistical characteristics (bias, correlation coefficient, and slope) are not improved. Post screening does not remove the false seasonal cycles for the Southern Hemisphere.

Significant improvement of the retrieval results was achieved when using training Subset 2 (Table 1 and right hand panels in Figs. 3a and b).). In this case, application of the retrieval procedure to the training set directly leads to the following precision/accuracy characteristics: mean bias of 0.00 ppm, standard deviation of 1.49 ppm, correlation coefficient of 0.91, and regression slope of 0.91. As expected, we have a small degradation of the results for Lamont site as compared with training Subset 1. At the same time, we have noticeable improvement for almost all Northern Hemisphere sites and significant improvements for Southern Hemisphere: as seen in Fig. 3, the XCO_2 retrievals now more accurately reproduce smooth TCCON like inter annual growth with no "false" seasonal cycles. As well as for Subset 1, the application of post screening by limiting spectral discrepancy does not result in much improvement in the retrieval results. A small improvement of scatter does not justify the considerable reduction of observation data output.

For comparison purposes, we have also included in the Table 1 XCO_2 retrievals by the NIES operational algorithm, version v02.21; release level for General Users. A considerable number of observation points from Subsets 1 and 2 are excluded from the operational Subset 3, mostly at the stage of post screening [6]. The accuracy and precision of EOF based algorithm are generally comparable to the operational algorithm, with similar characteristics while providing a noticeably higher yield (N) of retrievals.

As mentioned above, collocated GOSAT TCCON observations summarized in the Table 1 were selected from the NIES L2 CAI screened data set. The CAI based pre screening removes GOSAT observations taken in presence of optically thick/visible clouds.

However, the remaining data could be still affected by aerosols and/or optically thin (sub visual) cirrus clouds. NIES L2 operational algorithm is designed to correct these light scattering effects by simultaneous retrievals of both gas concentrations and aerosol/cloud optical thickness. The proposed EOF based algorithm has been trained using the observation data that are affected by atmospheric light scattering. We expect that such training allows for optical path modification by aerosols and clouds. These expectations are generally supported by the results in the Table 1: the precision/accuracy characteristics of EOF based algorithm are comparable with the similar characteristics of the "full physics" algorithm that simultaneously retrieves target gas amount and aerosol/cloud optical thickness.

We also performed independent XCO_2 retrievals for the GOSAT observations over TCCON site at Park Falls using the simplified algorithm (IMAP DOAS [30]) that ignores light scattering effects. The precision/accuracy of these retrievals proved to be very poor: mean bias of 8.9 ppm, standard deviation of 22.9 ppm and correlation coefficient r = 0.19. These data are further evidence that 1) we processed GOSAT observations affected by aerosols and/or optically thin clouds; and 2) EOF based algorithm does account for optical path modification by aerosols and clouds.

The presented results demonstrate that EOF based algorithm successfully reproduces dissimilar XCO_2 seasonal cycles for individual TCCON sites. Note also that for the validation purpose we used all available TCCON data, while for training we selected about 30% of these data. However, to overcome a certain circularity of the approach (i. e., the use of similar data for training and validation), additional tests are required.

4.2. *Additional tests using model simulations*

To additionally test the EOF based retrieval algorithm we select about 25,000 observations taken all over the globe within eight months that represent four seasons of 2010 and 2012, Fig. 1. (Recall that we used a reduced 1:5 version of this set to create the reference orthogonal basis). As reference XCO_2 data we use the original output of NIES (National Institute for Environmental studies) atmospheric tracer transport model, version 08.1i [19].

The application of the EOF based algorithm to these global observations gave strongly underestimated XCO_2 for the low surface pressure P_S values that were beyond the range of P_S variations over the TCCON sites (Fig. 4). These discrepancies are quite explainable: a decrease in gaseous optical thickness due to the drop of P_S is interpreted as low XCO_2 values. A clearly expressed dependence of the discrepancies on P_S enables one to derive a simple correction formula. However, such corrections are beyond the purposes of this study and instead we just limit ourselves with observations for P_S values that do not exceed the training set limits. Namely, we discard observations with $P_S < 880$ hPa (there are about 11% of such observations in the extended test set). The remaining ~90% data show rather good agreement with the reference model data except several strongly underestimated XCO_2 values (Fig. 4), all of which were taken over polar region of Eastern Hemisphere under low Sun conditions (i.e. again under conditions that are not covered by the training set) (Fig. 4).

Table 2 summarizes key statistical characteristics of the EOF model XCO_2 intercomparison. As seen from the table, the worst characteristics (i.e. maximal discrepancies) are seen for the tropics, which can be partially explained by the small number of tropics observations in the training set. Nevertheless, statistical characteristics are comparable with similar characteristics of recently developed algorithms [7] with a significant benefit in the amount of the available data (yield) and computation time.

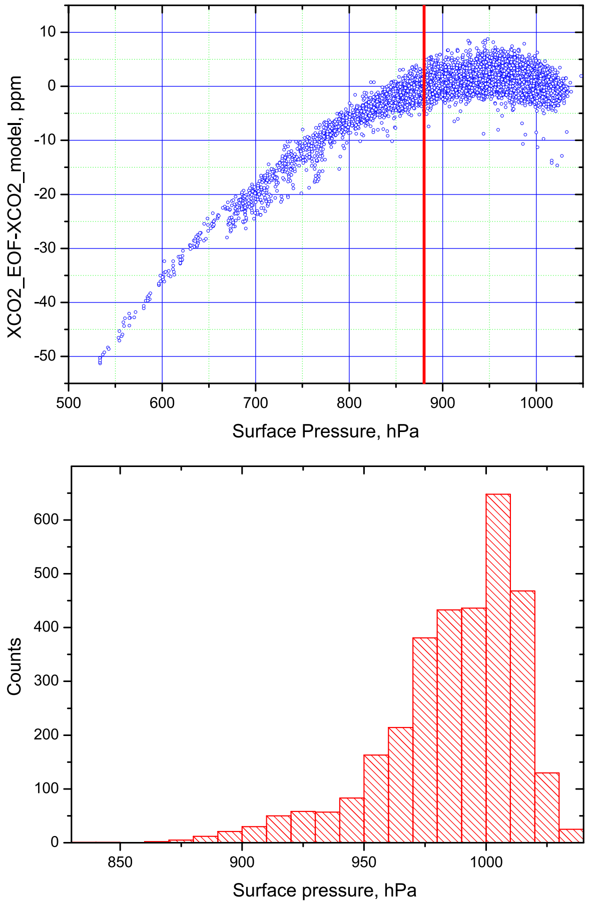

Fig. 4. The difference between GOSAT-EOF XCO_2 retrievals and NIES-TM model data as a function of surface pressure for the test set of about 25 000 cloud-free GOSAT observations taken within eight months that represent four seasons of 2010 and 2012 (upper panel). Lower panel show the distribution of the surface pressure values within test set of GOSAT observations around 12 TCCON sites. The vertical line in the upper panel indicates the value of the surface pressure, below which the current algorithm version is not valid.

Table 2
Statistical characteristics of the GOSAT-EOF versus model XCO_2 intercomparison.

	N	Bias (ppm)	σ (ppm)	Slope	r
All observations	22602	0.93	1.48	1.00	0.86
North, latitude > 23.5°	8940	0.59	1.45	1.05	0.90
South, latitude < −23.5°	3436	0.74	0.96	0.87	0.91
Tropics, −23.5° < latitude < 23.5°	10226	1.29	1.56	0.94	0.81

5. Discussion and conclusions

Development of very fast XCO_2 retrieval algorithms to process the huge amounts of ongoing (e. g. from GOSAT and OCO 2) and future (e.g. TanSat, GOSAT 2, etc.) satellite observation data is still of interest.

We propose a novel retrieval algorithm for rapid retrieval of carbon dioxide total column amounts from the Greenhouse gases Observing Satellite (GOSAT) observations. The algorithm performs EOF decomposition of the measured spectral radiance and combines a limited number of the decomposition coefficients in terms of principal components with *a priori* data such as airmass, surface pressure, etc. The regression formulae for retrieving target gas amounts are derived using training sets of collocated GOSAT and ground based observations.

This regression like algorithm proves to be a promising option with very low computational costs and a rather encouraging quality of retrieval results: the algorithm provides the XCO_2 precision/accuracy that is comparable with similar characteristics of current operational data [3 8]. Additionally, this algorithm provides an impressive yield (number of the retrievals in the final product).

The precision/ accuracy of the algorithm were shown to depend dramatically on the selection of the training set that must span the variability of XCO_2 and observation conditions (e. g. airmass, surface pressure, etc.). To create a training set we used reference observation data from twelve TCCON sites and rather simple criteria to select collocated GOSAT TCCON observations. Further improvement of the global algorithm precision/ accuracy is expected from extension of the training set by 1)including additional TCCON sites (e. g. Caltech, Eureka, and Edwards, Northern America; Ny Alesund and Paris, Europe); and 2) by using more advanced collocation criteria, such as the T700 colocation method [20] or the model based methods [7,28,29]. These advanced criteria enable us to expand areas of GOSAT TCCON collocated data providing higher variability of meteorological and geo locational conditions within the training set.

Funding sources and acknowledgments

Andrey Bril and Sergey Oshchepkov were supported by Belarusian Republican Foundation for Fundamental Research (BRFFR), Grant F15CO 023.

Dmitry Belikov was supported by The Tomsk State University Academic D.I. Mendeleev Fund Program in 2014 2015 under the Grant of the Ministry for Education and Science of the Russian Federation No. 5.628.2014/K.

TCCON data were obtained from the TCCON Data Archive, hosted by the Carbon Dioxide Information Analysis Center (CDIAC; tccon.ornl.gov).

Observations at Lamont and Park Falls are funded by NASA grants NNX14AI60G, NNX11AG01G, NAG5 12247, NNG05 GD07G, and NASA Orbiting Carbon Observatory Program.

Observations at Darwin and Wollongong are funded by NASA grants NAG5 12247 and NNG05 GD07G and the Australian Research Council grants DP140101552, DP110103118, DP0879468 and LP0562346. We are grateful to the DOE ARM program for technical support in Darwin. Nicholas Deutscher is supported by an Australian Research Council Discovery Early Career Researcher Award, DE140100178. Observations at Bremen, Bialystok and Orleans are funded by the EU projects InGOS (Grant agreement No 284274) and ICOS INWIRE (Grant agreement no 313169), and by the Senate of Bremen, and supported at Orleans by the RAMCES team at LSCE.

The Lauder TCCON programme is core funded by NIWA through New Zealand's Ministry of Business, Innovation and Employment.

Research at the FMI was supported by the Academy of Finland under Grant no. 140408 and by EU under project GAIA CLIM. (Grant agreement No 640276).

Garmisch TCCON work has been supported by the EU within INGOS and by the ESA ghg cci project.

References

[1] Takagi H, Saeki T, Oda T, Saito M, Valsala V, Belikov D, et al. On the benefit of GOSAT observations to the estimation of regional CO_2 fluxes. SOLA 2011;7:161.

[2] Maksyutov S, Takagi H, Valsala VK, Saito M, Oda T, Saeki T, et al. Regional CO_2 flux estimates for 2009–2010 based on GOSAT and ground-based CO_2 observations. Atmos Chem Phys 2013;13:9351. http://dx.doi.org/10.5194/acp-13-9351-2013.

[3] Butz A, Guerlet S, Hasekamp O, Schepers D, Galli A, Aben I, et al. Toward accurate CO_2 and CH_4 observations from GOSAT. Geophys Res Lett 2011;38:L14812. http://dx.doi.org/10.1029/2011GL047888.

[4] O'Dell CW, Connor B, Bosch H, O'Brien D, Frankenberg C, Castano R, et al. The ACOS CO_2 retrieval algorithm – Part 1: description and validation against synthetic observations. Atmos Meas Tech 2012;5:99. http://dx.doi.org/10.5194/amt-5-99-2012.

[5] Cogan AJ, Boesch H, Parker RJ, Feng L, Palmer PI, Blavier J-FL, et al. Atmospheric carbon dioxide retrieved from the Greenhouse gases Observing Satellite (GOSAT): comparison with ground-based TCCON observations and GEOS-Chem model calculations. J Geophys Res 2012;117(D21). http://dx.doi.org/10.1029/2012JD018087.

[6] Yoshida Y, Kikuchi N, Morino I, Uchino O, Oshchepkov S, Bril A, et al. Improvement of the retrieval algorithm for GOSAT SWIR XCO_2 and XCH_4 and their validation using TCCON data. Atmos Meas Tech 2013;6:1533.

[7] Oshchepkov S, Bril A, Yokota T, Wennberg PO, Deutscher NM, Wunch D, et al. Effects of atmospheric light scattering on spectroscopic observations of greenhouse gases from space. Part 2: algorithm intercomparison in the GOSAT data processing for CO_2 over TCCON sites. J Geophys Res 2013;118:1. http://dx.doi.org/10.1002/jgrd.50146.

[8] Oshchepkov S, Bril A, Yokota T, Yoshida Y, Blumenstock T, Deutscher NM, et al. Simultaneous retrieval of atmospheric CO_2 and light path modification from space-based spectroscopic observations of greenhouse gases: methodology and application to GOSAT measurements over TCCON sites. Appl Opt 2013;52:1339.

[9] Chevallier F, Palmer PI, Feng L, Boesch H, O'Dell CW, Bousquet P. Toward robust and consistent regional CO2 flux estimates from in situ and spaceborne measurements of atmospheric CO_2. Geophys Res Lett 2014;41:1065. http://dx.doi.org/10.1002/2013GL058772.

[10] Frankenberg C, Pollock R, Lee RAM, Rosenberg R, Blavier J-F, Crisp D, et al. The Orbiting Carbon Observatory (OCO-2): spectrometer performance evaluation using pre-launch direct sun measurements. Atmos Meas Tech 2015;8:301.

[11] Yi Liu, Cai Zhaonan, Yang Dongxu, Duan Minzheng, Lv Daren, Yin Zengshan, et al. Development of Chinese Carbon Dioxide Satellite (TanSat). Geophys Res Abstr 2013;15 [EGU2013-2524].

[12] ⟨www.gosat.nies.go.jp⟩.

[13] Hannachi A, Jolliffe IT, Stephenson DB. Empirical orthogonal functions and related techniques in atmospheric science: a review. Int J Climatol 2007;27:1119.

[14] Zhang Y, Xiong X, Tao J, Yu C, Zou M, Chen L Su. Methane retrieval from Atmospheric Infrared Sounder using EOF-based regression algorithm and its validation. Chin Sci Bull 2014;59(14):150.

[15] Kataev M, Kataev S, Maksyutov S, Andreev A, Bazelyuk S, Lukianov A. Mathematical algorithms for processing and analysis of near-infrared data from a satellite-borne Fourier transform spectrometer. Russ Phys J 2012;55(3):330–335. http://dx.doi.org/10.1007/s11182-012-9816-3.

[16] Golub GH, van Loan CF. Matrix computation. Baltimore, MD: John Hopkins University Press; 1996. p. 664.

[17] Yoshida Y, Ota Y, Eguchi N, Kikuchi N, Nobuta K, Tran H, et al. Retrieval algorithm for CO_2 and CH_4 column abundances from short-wavelength infrared spectral observations by the Greenhouse gases observing satellite. Atmos Meas Tech 2011;4:717.

[18] Kuze A, Suto H, Nakajima M, Hamazaki T. Thermal and near infrared sensor for carbon observation Fourier-transform spectrometer on the Greenhouse Gases Observing Satellite for greenhouse gases monitoring. Appl Opt 2009;48:6716–6733.

[19] Belikov D, Maksyutov S, Sherlock V, Aoki S, Deutscher NM, Dohe S, et al. Simulations of column-average CO_2 and CH_4 using the NIES TM with a hybrid sigma isentropic (σ-θ) vertical coordinate. Atmos Chem Phys 2013;13:1713.

[20] Wunch D, Toon GC, Blavier J-FL, Washenfelder RA, Notholt J, Connor BJ, et al. The total carbon column observing network (TCCON). Philos Trans Roy Soc A 2011;369:2087.

[21] Wunch D, Wennberg PO, Toon GC, Keppel-Aleks G, Yavin YG. Emissions of greenhouse gases from a North American megacity. Geophys Res Lett 2009;36:L15810. http://dx.doi.org/10.1029/2009GL039825.

[22] Deutscher NM, Griffith DWT, Bryant GW, Wennberg PO, Toon GC, Washenfelder RA, et al. Total column CO_2 measurements at Darwin, Australia-site description and calibration against in situ aircraft profiles. Atmos Meas Tech 2010;3:947–958. http://dx.doi.org/10.5194/amt-3-947-2010.

[23] Messerschmidt J, Macatangay R, Notholt J, Petri C, Warneke T, Weinzierl C. Side by side measurements of CO2 by ground-based Fourier transform spectrometry (FTS). Tellus B 2010;62:749–758. http://dx.doi.org/10.1111/j.1600-0889.2010.00491.x [2010].

[24] Messerschmidt J, Chen H, Deutscher NM, Gerbig C, Grupe P, Katrynski K, et al. Automated ground-based remote sensing measurements of greenhouse gases at the Białystok site in comparison with collocated in-situ measurements and model data. Atmos Chem Phys Discuss 2011;11:32245–32282. http://dx.doi.org/10.5194/acpd-11-32245-2011.

[25] Ohyama H, Morino I, Nagahama T, Machida T, Suto H, Oguma H, et al. Column-averaged volume mixing ratio of CO2 measured with ground-based Fourier transform spectrometer at Tsukuba. J Geophys Res 2009;114:D18303. http://dx.doi.org/10.1029/2008JD011465.

[26] Hausmann P, Sussmann R, Smale D. Contribution of oil and natural gas production to renewed increase in atmospheric methane (2007–2014): top–down estimate from ethane and methane column observations. Atmos Chem Phys 2016;16:3227–3244. http://dx.doi.org/10.5194/acp-16-3227-2016.

[27] Kivi R, Heikkinen P. Fourier transform spectrometer measurements of column CO2 at Sodankylä, Finland. Geosci Instrum Method Data Syst 2016;5:271–279. http://dx.doi.org/10.5194/gi-5-271.

[28] Guerlet S, Butz A, Schepers D, Basu S, Hasekamp OP, Kuze A, et al. Impact of aerosol and thin cirrus on retrieving and validating XCO_2 from GOSAT shortwave infrared measurements. J Geophys Res Atmos 2013;118:4887.

[29] Belikov DA, Maksyutov S, Ganshin A, Zhuravlev R, Deutscher NM, Wunch D, et al. Study of the footprints of short-term variation in XCO_2 observed by TCCON sites using NIES and FLEXPART atmospheric transport models. Atmos Chem Phys Discuss 2016. http://dx.doi.org/10.5194/acp-2016-201.

[30] Frankenberg C, Platt U, Wagner T. Iterative maximum a posteriori (IMAP)-DOAS for retrieval of strongly absorbing trace gases: model studies for CH_4 and CO_2 retrieval from near infrared spectra of SCIAMACHYonboard ENVISAT. Atmos Chem Phys 2005;5:9–22.