Reply to Maximilian Reuter

Dear Referee,

We thank you for the encouraging general comment on our manuscript amtd-7-12173-2014 “A linear method for the retrieval of sun-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY data” and for the constructive comments, which helped to improve the manuscript. We did our best to address all your comments in the revised version of the manuscript. Please find below an item-by-item response to your comments and concerns (comments are listed in black and responses in blue).
Please note that we have additionally listed the most significant revisions in the general final comment.

Sincerely,

Philipp Köhler on behalf of the authors

Orthogonality: You implicitly assume that the SIF spectrum hf · T↑ is orthogonal to all used PCs. Any correlation will result in the retrieval being not perfectly able to disentangle the amount of SIF (Fs) from the other fitted coefficients. This should at least be discussed. Additionally, you could enhance the confidence in the method by showing that there are no such spectral correlations.

From our point of view, it is not necessary to assume a complete orthogonality. Nevertheless, a significant correlation of the reference fluorescence emission spectrum (hf) with any of the provided PCs would lead to higher errors. Thus, we tested the correlation in the simulated data set and provide a correlation error matrix for a sample fit in order to address the raised concern. Results from this analysis are presented in the new Fig. 8 of the revised manuscript. The correlation error matrix shows that spectral correlations are not perfectly zero, but absolute values do not exceed 0.3. In addition, we tested the correlation of the PCs obtained by SCIAMACHY and the assumed hf spectrum with a similar finding (mean correlation coefficient is 0.2).

Heterogenity: You emphasise that you need to include soundings above ice in order to construct a training data set that covers enough heterogeneity to be useful for all scenes (including vegetation). What is the justification to assume that vegetated surfaces can be excluded from the training? An indication for this would be if the residuals (measured radiance minus fitted radiance) are not larger for vegetated surfaces than for surfaces included in the training.

We claim that it is essential to capture as many atmospheric states as possible within the training data set, which includes soundings over ice. Vegetated surfaces should inherently be excluded from the training data in order to accurately retrieve SIF. The fluorescence emission may otherwise be incorporated into the PCs.

Residuals for vegetated surfaces are indeed not larger than for other surfaces. Please note that we mention the average residual sum of squares (RSS) in Sect. 5.2 (previously 3.7, “Quality control”) and use the RSS to filter 'bad' retrievals. In this context, it should be considered that measurements over vegetated areas would be removed if higher RSS values would occur in such regions. We also mention that high RSS values can be attributed to the South Atlantic Anomaly. From our point of view, an

additional map would not add fundamental information.

Training and test data set: Section 3.2 describes the selection of training and test data sets without making clear what the purpose of theses dataset are. Please add a discussion to the fundamentals.

We have re-arranged the 'Retrieval methodology' section in order to make our manuscript more concise. A brief paragraph about the training and test set has been added to the fundamentals according to your suggestion.

Eq.1: i) Where is this equation coming from? Please add a reference or derive (Joiner et al. (2013) does not give more information). ii) It should be mentioned that this is already an approximation of the RT and the assumptions made (e.g., plane parallel atmosphere) and their potential influence on the retrieval should be discussed.

We agree that the manuscript suffered from a lack of information at this point. For this reason, we have rewritten this section according to the raised concerns. In the revised manuscript, we derive the forward model starting from the most simplistic formulation of the TOA signal.

Sec.3.3: The terms “apparent reflectance” and “atmospheric path radiance” have not been discussed in respect to Eq.1 and 2. Please use consistent terms and describe how these terms relate to the terms used in Eq.1 and 2.

Indeed, we have to acknowledge that the previous argumentation was inconsistent with respect to the equations. Please note that we have revised the section including new Eq. (1), (2) and (5) in order to remain consistent. In particular, we argue that the planetary reflectance (without SIF, including path radiance, surface reflectance and atmospheric transmittance) can be modeled by a combination of spectrally low and high frequency components. Apparent reflectance refers to the surface reflectance at the TOA and therefore to the spectrally smooth part, which is represented by a third order polynomial in wavelength. Thus, both surface reflectance and atmospheric path reflectance are represented in our forward model by the same polynomial in wavelength. Although the continuum absorption of the atmosphere is part of the apparent reflectance (low frequency component), we formally refer to an effective atmospheric transmittance for the oscillating component of the spectrum. That part is represented by the atmospheric principal components.

Eq.3: Eq.3 implies that the total reflectance is a product of transmittance (sum over the PCs) and a polynomial. This is conceptual different from Eq.1 and 2 with the additive term $\rho 0$ ($\rho 0 + \rho s \cdot T$). Maybe this clarifies when introducing a proper definition of apparent reflectance to Eq.2. Otherwise the connection to Eq.1 and 2 is unclear.

As stated in the reply to your last comment, the raised concern is justified. Please note that the former Eq. 1 and Eq. 2 are not necessary to derive our forward model. Nevertheless, it has to be emphasized that the main information for the retrieval comes from the solar Fraunhofer lines. It is less important how contributions from atmosphere, surface reflectance and fluorescence are modeled as long as the measurement is reproduced with a sufficient accuracy to be able to evaluate the changes in the fractional depth of solar lines by fluorescence. Here, we use low and high frequency components to

model the signal. The previous manuscript was not concise enough and has been revised accordingly as mentioned in the last replies.

Description of the retrieval: Sec.3.4,P12184L17: "... for the actual retrieval (Eq.3)..." Eq.3 is the forward model not the retrieval. It includes the to be derived state vector elements. However, a detailed description of how to obtain these parameters is missing. I suppose you perform a least squares fit without any prior constraints (?). Please add a description.

Thanks for this comment. Model parameters and state vector elements are explicitly named in the revised manuscript. Your assumption is right, an ordinary least squares fit is applied, which is also documented in the revised section 3.4.

Eq.5: Please explain the intention of introducing γ and the advantage (?) of enhancing the dimensionality of the problem. With $\gamma = \alpha \cdot \beta T$, the elements of γ cannot be considered independent but within the process of backward elimination they are handled as if they were independent. According to Eq.3 the apparent reflection (polynomial) is multiplied with the transmittance (PCs), i.e., each PC is multiplied with the same polynomial. In contrast to this, Eq.5 allows to have different polynomials for different PCs. Physically and in respect to Eq.3, this seems to be not meaningful.

We believe that the manuscript was not clear enough to justify our approach. Therefore, the relevant sections have been revised accordingly. In summary, the planetary reflectance is modeled by combination of a third order polynomial in wavelength (identified as apparent reflectance) and a subset of atmospheric PCs (identified as effective atmospheric transmittance). The consequence is a significantly increased number of state vector elements compared to Joiner et al. (2013), whereas the backward elimination algorithm selects only a subset of those candidate state vector elements. Our sensitivity analysis shows that the number of actually selected coefficients is (on average) nevertheless smaller. Please note that the main point is to reconstruct the signal with a high accuracy and our approach is more statistically/data-driven than physically based.

Minor comments
Citations: Check citation style within the entire manuscript.

The citation style has been chosen according to the AMT reference guidelines.

P12174L14: It is not the first time, see Joiner et al. (2012).

We added the information that SIF at 740 nm is retrieved for the first time from SCIAMACHY in order to clarify this issue (Joiner et al., 2012 used the Ca II Fraunhofer line at 866 nm).

P12176L23-P12177L5: Please references to the corresponding sections of the paper.

Adopted as proposed.

Sec.3.2: As you state “the selection of the training set is highly relevant”. A map would help to show which soundings have been selected and improve reproducibility.

From our point of view, a map of one selected training set would not improve the reproducibility since spatio-temporal composites are needed to obtain meaningful results. Our intention was to emphasize that it is necessary to ensure that non fluorescent targets serve as a basis for the training set. Therefore, the real satellite data must first be partitioned as opposed to our sensitivity analysis. It might have been misleading to mix the pure retrieval methodology with the application to the real satellite data. We have separated the methodology from the application in the revised manuscript to clarify this issue. Please note that the relevant section has been extended and moved to Sect. 5.1 within the rearrangement of the manuscript.

Sec.3.2: “...the training set is sampled on a daily basis.” What does this mean?

It means that a new training set is sampled every day when the retrieval is applied to real satellite data from GOME-2. Due to the reduced number of soundings, we sample the training set on a three-day basis when SCIAMACHY data is used. We extended the description of the sampling in Sect. 5.1 to make this point more clear. In addition, a comparison of different sampling methods has been added to the “North-to South bias in training set” section (Sect. 5.6). It might be worth to note that it is expected to capture instrumental artifacts and degradation by sampling the training data in such short time periods.

Sec.3.2: Please describe the applied cloud/aerosol filtering for the training data set (if any).

Please consider that we state that there is explicitly no restriction for cloud fractions.

Fig.2: Due to the instrumental resolution, the used micro-window at about 722nm may be influenced by H2O (?) absorption (e.g., dip in the middle of the window). This can result in errors of the derived apparent reflection and hence transmission. How large is the potential influence on the derived SIF?

Thanks for this comment. It is true that the micro-window might be influenced by water vapour absorption. Too high retrieval results may occur, but the contribution of SIF in the lower wavelengths of the fitting window is decreased for two reasons: an increasing distance to the maximum of the prescribed spectral function at 740 nm and atmospheric absorption (except for the atmospheric micro-window). In order to address another comment, we extended the sensitivity study by showing (amongst others) the retrieved minus simulated SIF in dependence on the water vapour column in new Fig. 6. Results suggest that this effect is indeed negligible. We added a paragraph to the revised manuscript where the potential influence is mentioned (Sect. 3.3).

Sec.3.5: “... previously presented steps...” Which previously presented steps do you mean that reduced the number of coefficients?

We revised the section and removed the misleading phrase.

Sec.3.5: “Eq.3 contains i · j ...” i and j are only index variables. You could define, e.g., N and M . Additionally, Eq.3 does not contain N · M +but N + M + 1 coefficients.

Thanks for the proposed correction. We defined n_pc and rewrote the forward model as it is implemented. Please note that this is an algebraic transformation and hence Eq. 3 actually contains 4*n_pc+1 coefficients. We agree that it was not obvious and thus we clarified this issue in the revised manuscript.

P12185L3: Define “a few”.

We only stated “a few” because the exact number depends on the fitting window. We agree that this can lead to confusion and added the necessary explanation to the text as follows: “...(about 5, exact number depends on the fitting window).”

P12185L5: Please explain what you mean by overfitting. Joiner et al. (2013) used more PCs than you do (?). Does this imply they were overfitting?

Overfitting means that the number of model parameters is too high - the model is too complex. In consequence, it looses its predictive performance by fitting noise. We added this information to the Backward elimination section of the revised manuscript. It should be considered that there are many methods to compare and select models. Applying different criteria will lead to a different choice for the final model. It is true that our approach suggests to use less PCs than Joiner et al. (2013), but the retrieval algorithm is different. In particular, we have potentially 4*n_pc+1 state vector elements while Joiner et al. (2013) are solving their forward model only for 4+n_pc+1 state vector elements. Therefore, it is not possible to answer the question, but there is always a risk of overfitting when no model selection criteria is applied.

Eq.6: Why do you not use the noise estimate provided in the L1 files.

The reason for using that noise model is the estimation of a realistic noise for the simulated radiance. Indeed, it would be best to use the noise estimate attached to the L1 files for the real satellite data, but we just adopted the noise estimation of the sensitivity study for the sake of simplicity. Please realize that a similar procedure was used by Joiner et al. (2013).

P12187L11: Please compare the obtained RSS values with values expected from instrumental noise. This comparison can tell you if you are already fitting noise (too many PCs) or if important PCs are missing.

We appreciate this point and agree that it could be a good alternative to avoid an overfitting in our retrieval. However, we have preferred to stick to our backward elimination algorithm approach, as it also allows us to minimize the choose the optimum number of model parameters.

P12187L16: Define “high average”.

Done.

P12188L7: Check unit of water vapour columns.

Thanks, we corrected the unit to g/cm^2.

Sec.4.2,P12189L3: Please explain why you decided to use the first eight PCs but not the just introduced backward elimination algorithm.

Thanks for this comment. The formulation might have been misleading. The backward elimination selects the required PCs automatically, for this reason we stated "...providing the first eight PCs...". We, rephrased the sentence in order to clarify that we use the approach as described.

Fig.4: Fig.4 shows a good correlation between simulated and retrieved SIF. However, from this plot alone, systematic effects cannot be excluded. As an example, high H2 O values could always be hidden in the lower end of the error bar. You could enhance the confidence in the retrieval by showing that retrieved minus simulated SIF does not depend on H2O, solar zenith angle, AOD, etc.

We agree and would like to thank you for that proposal. You can find such a figure (new Fig. 6) in the revised manuscript.

P12191L13: "10 PCs" When applying the algorithm to real data, some PCs may also have to account for instrumental effects which was not the case for the simulations. Are 10 PCs still enough in this case? How often does the back elimination algorithm decide that 10 PCs (i.e., all) are needed.

Great thanks for this comment. We evaluated the number of selected PCs when our approach is applied to GOME-2 data. It has emerged that the algorithm selects on average six PCs and decides in about 3% of cases that ten PCs are actually needed. This finding corresponds to our sensitivity analysis, where on average seven PCs are selected. In the previous version of our manuscript, we implemented the retrieval on a trial basis for SCIAMACHY data. Meanwhile we processed data for the 08/2002—03/2012 time span with and found that things are changing fundamentally when the approach is applied to SCIAMACHY data. In this case, the algorithm decides in about 40% of cases that ten PCs are needed. Therefore, we provided 20 PCs to the algorithm, whereas on average 14 PCs are selected. The probability that all 20 PCs are selected is below 1%. In consequence of these findings, we provided ten PCs to the retrieval for GOME-2 and 20 PCs when SCIAMACHY data is used. You can find this information also in the revised manuscript in Sect. 5.1 Application to GOME-2 and SCIAMACHY data.

Sec.5.1,P12193: You use NDVI as proxy for cloud contamination. This seems to be very indirect. What if cloud contamination effects NDVI much less than the amount of vegetation. In this case the NDVI to NDVI graph would still have a slope close to 1. Therefore, your conclusions "it must be concluded that...most likely..." seems to be too strong. Why not using FRESCO?

We agree that a comparison with cloud fractions from FRESCO data is more meaningful. Please consider that the data basis (three month of SCIAMACHY SIF) was actually too weak for a comparison. In the revised manuscript we were able to enhance the comparison with the recently processed data from SCIAMACHY. We now present a comparison of long-term averages of TOA reflectance, FRESCO cloud fractions and SIF retrieval results (new Fig. 10) instead of the NDVI comparison.

Fig.7,Fig.9: Please discuss potential reasons for slopes being different from 1 (especially Fig.9). V25 uses more PCs; if one of the additional PCs is not orthogonal to the SIF spectrum $hf \cdot T\uparrow$, I would expect too low SIF values. Could this explain the observed discrepancy? If so, I would expect that your retrieval also critically depends on the number of PCs (even though simulated data - Fig.5 - does not suggest this). Additionally, please add a colour bar.

The reason for the observed discrepancy is most likely not associated with the number of PCs used. We assume that the difference can be attributed primarily to the usage of a reduced fitting window (734–758 nm) in the V25 results described in Joiner et al. (2014). We discuss this potential reason in the revised manuscript (Sect. 5.4). In view of this, it must be stated that there are still large uncertainties in absolute SIF values and it is very difficult to judge which values are closer to reality. Furthermore, we do not expect a critical dependency on the number of PCs in our retrieval approach, since the retrieval results from GOME-2 (ten provided PCs) and SCIAMACHY (20 provided PCs) are similar. Please note that Fig. 9 has been replaced by a hexbinplot (form of bivariate histogram) that now includes also a legend.

Eq.10: This is not the standard error of the mean. Please check the equation. Additionally, why introducing Eq.10 when Eq.8 was introduced for the same purpose?

Apologies for confusion. It is actually the standard error of the weighted mean. We revised this issue. Our intention was to show two error estimates: one for instrumental noise only and one for instrumental noise plus natural variability.

Fig.10: Please show the selected areas in a map (e.g., together with the selected training data set, see earlier comment).

Please note that the concerned comparison has been removed. A similar comparison between monthly averages of GOME-2 and SCIAMACHY SIF is now included in Fig. 11, whereas only a sample crop region is evaluated in order to investigate the temporal consistency. However, the analysed area has been added to Fig. 15 of the revised manuscript.

P12195L18: “At this point...” Could it help to compare with GOSAT? The lack of validation data sets is a general problem for satellite based SIF retrievals. Are there plans to improve the situation.

Thanks for this comment. Indeed, the only possibility to assess the validity of results from the data-driven approaches, besides sensitivity analyses, is at present a comparison to physically-based SIF retrieval results from GOSAT data. The much higher spectral resolution of the GOSAT FTS enables a simpler retrieval approach restricted to narrow fitting windows and hence more accurate results. It

should be mentioned that GOSAT and GOME-2/SCIAMACHY SIF values are expected to be different owing to different overpass times, evaluated wavelengths (GOSAT SIF is evaluated between 755–759nm), illumination and especially observation geometries. Nevertheless, we may expect a good correlation between those data sets. Therefore, we compare long-term averages (overlapping period) of GOME-2 SIF (our approach and V25 from Joiner et al., 2014) and the recently processed SCIAMACHY data with GOSAT results in the revised manuscript. It appears that there is an enhanced correlation to the GOSAT SIF data set for the presented retrieval algorithm. Nevertheless, it remains unclear which absolute values are more accurate. As far as the authors can tell, there are currently no plans to set up a validation site(s) that would be appropriate.

Sec.5.3: Please mark the approximate region of the SAA in one of the maps.

The SAA center has been added to Fig. 15 of the revised manuscript.

Sec.5.5: Please show the analysed area in a map.

The analysed area has been added to Fig. 15 of the revised manuscript.

Sec.5.5,P12200L2: "...is robust against..." This conclusion seems to be too strong. Fig.13 shows that changing the cloud detection threshold from 0.25 to 1. can result in SIF averages being about 30% off. As the sample without cloud filter (threshold 1.) includes also all cloud free scenes, it can be assumed that the influence on individual soundings is much larger than 30%.

We agree with the reviewer and modified the formulation to: "...robust against moderate cloud contamination.". Consequently, we modified the relevant part in the conlusions to: "...moderately affected by cloud contamination.".

P12201L3: ENVISAT was launched in March 2002.

It is true that ENVISAT was launched in March 2002, but the data availability goes only back to August 2002. We therefore gave an estimated time frame. Please note that we removed this statement because we present the entire time series in the revised manuscript.

P12201L10: If the proposed method could also be adapted to CarbonSat (which is probably the case), CarbonSat (Bovensmann et al., 2010; Buchwitz et al., 2013) should be cited in this context because it enables carbon cycle analysis based on simultaneous measurements of XCO2 and SIF.

We agree and added the citation as proposed.