Reply to Referee #3

Dear Referee,

Thank you very much for your effort and time to review our manuscript amtd-7-12173-2014 “A linear method for the retrieval of sun-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY data”. We highly appreciate your detailed comments which helped to improve the manuscript substantially. In the item-by-item response hereinafter, 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

The authors present an improved retrieval methodology for SIF retrieval. They are continuing previous work by Joiner et al. (2013). The method has now also been applied to a couple of months of Scia data. The topic is interesting but the paper needs to be substantially revised before I can recommend publication. I encourage the authors to do so because I see enough potential for this paper to become publishable. Below I have listed major and minor comments. I would appreciate if these are addressed point by point. Many minor comments however serve as an example. I emphasize here that the authors should use these comments to also critically look at other parts of the manuscript.

Major comments:
-Overall, I feel that the style of the manuscript needs to be improved. I quite often struggled to understand what the authors are trying to say. Sentences can sometimes be wordy and beside the point. Also, the way of presenting their material is sometimes not very well structured (e.g., introduction and discussion of eq.3 should be moved up to p.12179) or confusing (e.g., discussion of training and test spectra should be more clearly separated). I encourage the authors to reread the manuscript and improve its style. Perhaps a native English speaker can help. I have listed detailed comments until section 3.1 to make my point. Please address these comments. They should also be used, however, to thoroughly improve the remainder of the manuscript, which is necessary. . . I may be nitpicky but these detailed comments are intended to help the authors make their story more clear.

Thank you for pointing out that the manuscript could have better been structured. Hence, we rearranged the methodology section and rewrote parts of it. In particular, equation 3 is introduced earlier, descriptions of training and test sets are clearly separated (with respect to methodology/simulation/real data) and all necessary equations moved to the methodology section. We extended the description of the backward elimination algorithm and discuss implications arising from assumptions made.
We believe that the revised manuscript is much more clear and justifies our approach in an appropriate manner.

-In the description of the retrieval methodology, there is quite some overlap with Joiner et al. (2013). I think the description can be made much more concise and to the point. Also, the focus should be on differences with Joiner et al. (2013). These should definitely be more explicitly stated and argued. The

unique contribution that the authors make in their present manuscript to the retrieval methodology should stand out more clearly.

We agree that differences to the Joiner et al. (2013) algorithm and our unique contribution should stand out more clearly. We revised the manuscript accordingly including a modified derivation of the forward model starting from the most simplistic formulation of the TOA signal. We defined an effective atmospheric ground to sensor transmittance ($T_{\uparrow}^{e}$) to clarify that this an additional model parameter instead of a state vector element in the retrieval from Joiner et al. (2013). Implications of this step are now discussed more in detail to justify our approach. Furthermore, we explicitly mention that the algebraic transformation to obtain our final forward model implies initially an increased number of state vector elements (which is later on reversed by the stepwise model selection). We further emphasized the importance of the automated determination of the optimum number of PCs.

-The authors linearize the forward model by estimating T_up in a pre-processing step, but this comes at a cost (loss of accuracy). As far as I can see, the only reason for linearizing the forward model is computational speed: for the automated determination of PCs according to the BIC, retrieval for a single target pixel has to be run multiple times. Linearization is not needed for the backward elimination per se (or am I missing something?). The authors need to show whether the supposed accuracy improvement when optimizing the choice of PCs outweighs the loss of accuracy due to the linearization.

From our point of view, solving for $T_{\uparrow\downarrow}$ and $T_{\uparrow}$ simultaneously has no significant advantage with respect to decoupling them. Please note that all simulations presented by Guanter et al. (2015) are also based on that decoupling. Those results could serve as evidence to support our statement. We discuss further implications in the revised manuscript in order to justify this approach. In particular, we show by means of simulated TOA radiances that the error due to an in-filling of atmospheric absorption lines is negligible. Therefore, we have tested two scenarios without instrumental noise, a medium and a large difference in SIF emission (2.1 and 4.3mW/m^2/sr/nm, respectively), and calculated $T_{\uparrow}$. The resulting changes are marginal (new Fig. 4 in the revised manuscript).

In the extended sensitivity analysis we are further able to show that our algorithm selects about 7 PCs and 15 coefficients respectively. In this context, it is to be remarked that Joiner et al. proposed to use 25 PCs for an equivalent retrieval window. Despite the significantly lower number of PCs used, the main advantage of our approach is that the number of selected coefficients remains stable regardless of how many PCs are provided.

In view of other studies (Guanter et al., 2013 and Guanter et al., 2015), our extended sensitivity analysis and presented results, it can be claimed that the linearization is not affecting the accuracy. On the other hand, a model selection can also be applied to the non-linear forward model but, as the reviewer is stating, this would be computationally less efficient.

The appropriate comparison for the retrieval simulations of sect.4.3 would be to compare the backward elimination mode (including the linearization) with the full-PC non(!)-linear forward model fit. Also, is computational speed indeed an issue here? Can the authors give an estimate of the increase in computation time when using backward elimination with the non-linear forward model instead of backward elimination with the linear forward model. Other comments and questions that I have about the analysis in sect.4.3 can be found below.

We agree that it would be interesting to apply a model selection also to the non-linear forward model of Joiner et al. (2013), but we consider this to be beyond the scope of this study. Please note that a comparable forward model to our approach was proposed by Guanter et al. (2013) and recently by Guanter et al. (2015). The primary reasons to use a linear model are an easier implementation and an increased computationally efficiency. The non-linear forward model has not been implemented in this study. Consequently, we are not able to provide a definite answer to the question of the effect on the computation time. Nevertheless, we presume that there would be an increase by a factor of 4-6 since that is the number of iterations until the algorithm of Joiner et al.(2013) typically converges.

I encourage the authors to rethink their analysis and way of presenting it. There are interesting results, but it can be presented more convincingly.

We have to acknowledge that the manuscript was not clear enough to justify our approach in an appropriate manner. We therefore substantially revised the relevant parts according to the reviewer`s comments. Furthermore, we have processed the SCIAMACHY data for the 08/2002-03/2012 time span. For this reason, we are able to compare long-term averages of different SIF data sets (from GOME-2, SCIAMACHY, GOSAT). We feel that the revised manuscript provides more interesting and valuable results than before.

-I have stopped doing a detailed analysis of sect.5. As an overall remark, this section is primarily a description of the SIF data sets. Maps, trend plots and biome-specific analyses are very similar to the previous papers and I am struggling to see a unique contribution here. If the point of this paper is an improvement in retrieval methodology, also the focus of this section should be much more on presenting and discussing differences between data sets to support the conclusions from the retrieval simulations, rather than just trying to convince the reader that the retrieval algorithm is picking up a fluorescence signal, which has been done before. If I think about possible ways to improve this section and make it more worthwhile, the PNAS paper comes to mind. You can ask yourself what is needed to produce a SIF database that can readily be compared with GPP flux tower measurements? To my recollection, in that paper corrections were made, for example, for the area of a grid box actually covered by vegetation. Also, can you make corrections for the presence of clouds? This is a just a suggestion to help the authors. I am open to other ways of improving this section.

The criticism is justified and we rethought the way of presenting our results. As already mentioned in the last reply, we now compare long-term averages of SIF data sets in the revised manuscript. A key finding is that the correlation between GOME-2/SCIAMACHY and GOSAT SIF is enhanced with the presented approach but that there are still uncertainties in absolute SIF values. We also extend the analysis of the North-south bias in the training set in order investigate more in detail if this bias is really related to an instrumental issue. This is most likely the case, although the magnitude can be decreased by an appropriate sampling of the training set.

Minor comments:
p.12174,l.20: photosynthetically-active -> photosynthetically active

Done.

p.12175,l.1: Start new paragraph after ‘field’?

Done.

p.12175,l.3: I believe the paper explaining the methodology is Frankenberg et al. (2011a) rather than Frankenberg et al. (2011b), and this reference would be more appropriate here.

Adopted as poposed.

p.12175,l.6: enabled the possibility to evaluate -> ‘enabled evaluation’ or ‘gave the possibility to evaluate’

Formulation has been modified to 'enabled evaluation'.

p.12175,l.9-10: used both the 757 and the 770 nm spectral regions -> I would speak of fit windows instead of spectral regions here? How broad are these fit windows? In view of the discussion of GOME2 retrievals below, you could mention that GOSAT only covers the O2 A band. Then, the reader understands why such small (micro) windows are used in the GOSAT retrievals.

We agree with the reviewer and revised the concerned parts of the manuscript.

p.12175,l.10: and is devoid of significant atmospheric absorption -> Do you mean that even for the isolated Fraunhofer line at 770.1 nm there is residual absorption by oxygen?

Indeed, the fitting window between 769-9—770-25 nm includes two very weak O2 absorption lines on either side of the K I line and there is also a O2 absorption line located at 769.9 nm.
Thus we would like to keep our initial formulation.

p.12175,l.11-12: Using only the single line in 770 nm simplifies the forward model -> Is this true? The method by Joiner et al. (2011) could equally well be applied to the micro windows of Guanter et al. (2012), I would say. So using an isolated line doesn’t really simplify. An important practical problem though is that a measured irradiance should be available.

We agree that the formulation might be misleading and removed the disputatious part of the sentence.

p.12175,l.11-12: single line in 770 nm -> single line at 770 nm (throughout paper!)

Thanks, we went through the manuscript and corrected the mistakes.

p.12175,l.14: The method proposed by Frankenberg et al. (2011b). . . -> You haven’t yet discussed the method by Joiner et al. (2011), which to my recollection relies on having available a measured

irradiance spectrum.

We agree and added the missing information to the text.

p.12175,l.21: narrow -> Why narrow? (See your paper.)

Indeed, narrow can be and has been left out in the revised manuscript.

p.12175,l.28: (2deg x 2deg) which -> What is temporal resolution? Add comma: '2deg), which'.

From our point of view, the information about the temporal resolution is unnecessary at this point because we aim to discuss the intrinsic limitation due to the incomplete spatial coverage. Nevertheless, global composites are only meaningful when averaging retrieval results on larger time scales such as monthly means (because of the presence of clouds and the high retrieval noise). This applies similarly to GOME-2/SCIAMACHY SIF composites, whereas the temporal resolution can be increased a bit (we recommend to use at least eight days for GOME-2 and 16 days for SCIAMACHY). In this context, it might be worth to recap the revisit times:
- GOME-2: 1.5 days,
- GOSAT: 3 days,
- SCIAMACHY: 6 days

p.12175,l.28-29: high retrieval noise -> Do you have a reference for that? Or, in view of your G2 and SCIA maps, can you argue that GOSAT has relatively larger noise errors than G2 and SCIA?

Please note that the main point is the incomplete global coverage of GOSAT and the following sentence relates solely to the better spatial coverage of GOME-2/SCIAMACHY. For this reason, we would like to avoid a modification of the manuscript on this point. However, the relatively high retrieval noise of GOSAT SIF retrievals is documented in Frankenberg et al. (2011b), Guanter et al. (2012) and Köhler et al. (2015) with values about 20-30% for monthly composites. The standard error of the mean (also monthly) reported by Joiner et al. (2103) and our results indicate that the instrumental noise from GOME-2/SCIAMACHY is about half as large. We added the reference Frankenberg et al. (2011b) at this place to address the raised concern.

p.12176,l.1: with lower spectral resolutions -> I would insert here 'but better spatial coverage'.

Done.

p.12176,l.4: instruments which -> instrument, which

Done.

p.12176,l.8: might be -> is

Done.

p.12176,l.9-11: For this reason, . . .continuous spatial sampling. -> Overly cautious and wordy. Sentence can actually be left out, or changed to 'Thus, the representativeness of spatially mapped SiF is enhanced by using data from instruments which provide a continuous spatial sampling.'

Thanks for pointing that out, we adopted the sentence as proposed.

p.12176,l.14-15: spectral and radiometric . . .enables a significant increase -> Why? Is radiometric performance of G2 superior to SCIA and GOSAT?

Thanks for this comment, we now realize that this formulation is misleading. The disputatious part of the sentence has been removed and subsequently, we merged the rest together with the sentence before.

p.12176,l.18: applied on -> applied to

Done.

p.12176,l.20-22: Reading this, I ask myself whether there any other ways to get info about SiF?

We must admit that this sentence is unnecessary and it was therefore removed.

p.12176,l.25-26: consists in -> wordy, replace for example by 'is'

Done.

p.12176,l.26-27: used singular . . . respectively. -> Again wordy: leave out 'used' (throughout paper!). Why make such a formal distinction between PCs (Joiner) and SVs (Guanter) here? It distracts. They are (basically) the same, to my knowledge. For example, simply replace phrase by 'the principal components needed for retrieval' (throughout paper!).

We went through the manuscript and revised numerous formulations according to the raised concern.

p.12176,l.28-29: We test . . . retrieval method -> A bit awkward, perhaps replace by: 'We apply the proposed retrieval method to data from G2 and SCIA to demonstrate its feasibility'

Sentence has been modified to: "The proposed retrieval method is applied to simulated data (Sect. 4) as well as to data from GOME-2 and SCIAMACHY (Sect. 5)."

p.12177,l.1: evaluated at 740 nm -> Why do you mention this here? I don't think it is needed this place, but doing so makes me think about your remark. For example, are you doing something different than

previous authors? It doesn't seem so. Is the wavelength at which SiF is retrieved a distinct property of your retrieval method (e.g., depending on the choice of fit window)? No, because 740 nm just is the peak wavelength of the assumed spectral SiF model, which has been determined in laboratory measurements.

Our intention was to emphasize that we present the first SIF retrieval for the second fluorescence peak applied to SCIAMACHY data. Basically, you are right in the criticism but please note that it would also be possible to average the amount of SIF over the retrieval window. We therefore state 'at 740 nm' in order to avoid any misunderstanding. It has become obsolete to mention it here since we rephrased the concerned sentence.

p.12177,l.8: Experiment -> Experiment-2

Done. Here and throughout the manuscript.

p.12177,l.13: overpass -> equator overpass (also in next section)

Done.

p.12177,l.15-18: The fourth channel . . . Fig. 1. -> Confusing sentence. What point are you trying to make? Perhaps change to 'The fourth channel (590 - 790 nm), which has a spectral resolution of 0.5 nm and a SNR up to 2000, covers the SiF wavelength region (Fig. 1).'

Thanks for the proposed streamlining of the sentence, which has been adopted.

p.12177,l.20: to evaluate . . . peak in 740 -> Peak 'at' 740 mm. Also, you are not retrieving SiF at 740 nm per se, but across the entire range between 720 nm and 758 nm. SiF at 720 nm contributes to the reported 740-nm peak value through the assumed spectral model.

As already stated above, we do not average the amount of fluorescence over the retrieval window and state therefore 'evaluated at 740 nm'. This does not exclude that we retrieve SIF in the entire retrieval window with respect to the assumed reference spectrum. We feel that it would be unnecessary to add a comment on that in the section about the satellite instruments because it belongs to the methodology section.

p.12177,l.22-24: The GOME-2 level 1B . . . device. -> Wordy: 'in photons per second' can be left out. It is perhaps important to mention though that the solar irradiance measurement is done daily principle. For example, 'The GOME-2 level 1b product consists of radiance spectra and a daily solar irradiance measurement.'

Thanks, streamlined sentence has been adopted.

p.12177,l.25: available -> 'used' (GOME-2 level 1b data from the entire mission were available for

your study, I guess. See also next section.)

Done.

p.12178,l.13: mode which -> mode, which

Done.

p.12178,l.19-21: That means . . . limitations. -> Sentence can be completely left out when ‘captured’ in l.18 changed to ‘downlinked’.

Thanks, adopted as proposed.

p.12179,l.2: ‘basically’ can be left out

Done.

p.12179,l.6: main challenge to retrieve SiF -> main challenge when retrieving SiF

Done.

p.12179,l9: to data-driven -> to the data-driven

Corrected.

p.12179,l.10: of the TOA signal → to the TOA signal

Comment does not apply anymore due to the revision of this section.

p.12179,l.10-12: In simplified . . . by SiF. -> I don’t quite understand what you are trying to say here? What exactly is simplified? It seems that you are stating an obvious thing and I would think that this sentence can be left out.

Sentence has been removed within the revision.

p.12179,l.15: I_sc -> Where does the ‘c’ stand for? I understand that you want to distinguish s(urface) from s(olar), but perhaps you can come up with more understandable subscripts.

I_sc has been changed to I_sol to address the raised concern.

p.12179,l.14,eq.1: Why do you use different symbols (L and F) for the radiance?

L has been changed to F.

p.12179,l.208-225: It seems that you are basically repeating and summarising previous papers here, particularly the one by Joiner et al. (2013). Why not refer to that paper and just state final expressions? Alternatively, you can indeed decide to summarise the derivation, but then you need to give more explanation. I think that the current version raises questions with the reader. For example: if you mention the approximation in l.21, state explicitly the reason for doing so (atmospheric scattering small in the near-infrared), relate T_down_up to T_up, and then explicitly argue why you can write. 12184,l.8,eq.4 (no diffuse contribution). Also, p.12183,eq.3 is not consistent with eq.2 (rho_o).
p.12179,l.18: amount of -> can be left out
p.12180,l.2-3: apparent reflectance -> please give definition

We have to acknowledge that the previous argumentation was inconsistent with respect to the equations. Please note that we have revised the section in order to remain consistent when summarising the derivation. 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. Thereby, the spectrally smooth part, which is represented by a third order polynomial in wavelength, can be attributed to surface reflectance and atmospheric scattering. For this reason, we refer to apparent reflectance. 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. If one strongly argues in a physically based way, it is necessary to neglect atmospheric scattering, path radiance and continuum to end up with a separate formulation for surface reflectance and atmospheric transmission. Our data-driven/statistically interpretation refers more to the reconstruction of the signal than to the single components of the equation. The main point thereby is to reproduce the measurement with a sufficient accuracy to be able to evaluate the changes in the fractional depth of solar lines by fluorescence. We therefore revised the manuscript accordingly.

p.12180,l.3-6: A subset . . . SiF retrievals. -> Please leave out descriptions of detector channels. They are already given before and it is not needed to repeat them here. For example, simply say ‘In out SiF retrievals we use a fit window extending from 720 nm to 758 nm’. Does the fit window differ from the one used by Joiner et al. (2013)?

This sentence has been removed within the revision. The latest version (V25) of the Joiner et al. (2014) algorithm uses a fitting window extending from 734-758 nm. This information can be found in Sect. 5.4 “Comparison with existing results” of the revised manuscript.

p.12180,l.8-9: at which we evaluate the amount of SiF -> Suppose you choose a fit window that does not cover 740 nm, for example a combination of windows from 720 nm to 739 nm and from 741 nm to 758 nm. Does your retrieval then fail? Do you think you cannot retrieve SiF at 740 nm with the current retrieval algorithm then? (See previous remark on this point.)

We believe that this is a misunderstanding: We retrieve SIF in the entire retrieval window but evaluate

the amount at the peak of the reference spectrum. Indeed, at this place this explanation is obsolete and has been removed. Please note also that the selection of the retrieval window is now discussed in point 4.4 together with results from the sensitivity analysis in different fitting windows to justify the choice of the 720—758 nm window.

p.12180,l.14: requires a more complex model -> Why?

This might have been a misleading formulation without the information that we mean a more complex model in terms of state vector elements to be derived. However, this bullet has been removed according to the second following comment.

p.12180,l.14-16: Simultaneously . . . accuracy. -> In view of the previous bullet, why not stop at 721.5 nm then?

It would of course be possible to stop at 721.5 nm, but we doubt that there would be a change as long as the atmospheric window is close to the lower/upper limit. Furthermore, we have not tested that retrieval window and removed that bullet according to the following comment.

p.12180,l.8-19: I think point 3 and 4 repeat point 2 and can be left out.

We agree and revised the manuscript accordingly.

p.12180,l.22-23: training set and test set -> Please give a definition of training and test set here or at the start of the next section. Also, I find test set a confusing name (what are you testing?). Why not say target set? For example, 'The training set is a set of reflectance spectra over non-vegetated areas used to determine PCs and the target set is the set of reflectance spectra over vegetated areas for which a SiF retrieval is attempted.' I stopped giving comprehensive and detailed comments here.

We rearranged the manuscript in order to clearly separate between the training set in theory, simulation and application to real satellite data. Therefore we added the most simplistic definition at this point (training set: measurements over non-fluorescent targets; test set: basically all measurements over land). The naming has been chosen according to the usual conventions in statistical textbooks, where the term training set is often used in conjunction with a test set.

p.12181,l.20: sun-glint -> Why mention this? You include only land pixels in your training set.

We suppose that the reviewer means the test set and agree that this is an unnecessary information, which has been removed.

p.1218-12181,sect.3.2: You select different spectra for the target set than Joiner et al. (2013). For example, you don't select sea ice or cloudy ocean. Why? Does it make a difference?

We suppose that the reviewer refers to the training set here. Then we have to admit that our description

was obviously misleading. Our training set is also sampled over the sea ice, ocean and non-vegetated areas (regardless of the degree of cloudiness). Especially the non-vegetated areas have been selected according to the IGBP land classification. We clarified the description in the present manuscript.

p.12182,l.1-2: Actually . . . test set. -> I don't understand this sentence?

We rephrased the sentence in order to clarify this issue: "The selection of the training set is relevant to obtain meaningful results because this data is used to model the planetary reflectance of the desired measurements over land (test set).".

p.12182,l.16-22: Why are discussing the spectral dependence of the surface reflectivity for vegetation here? Vegetated pixels are not included in your training set, are they?

Apologies for the confusing structure at this point. Sections have been substantially rearranged to avoid such a misleading argumentation. In particular, this discussion can be found in the revised section 4.4 "Selection of the retrieval window".

p.12186,l.9-12: Please add reference. How about modelling noise error for Scia?

Unfortunately, we are not able to provide a reference here. So far, we have used the same noise model for GOME-2 and SCIAMACHY as the spectral and radiometric performance is quite similar. We realize that this actually not appropriate and for this reason we do not show the standard error of the weighted average for SCIAMACHY. We will of course account for this issue in further investigations. In this context, it should be mentioned that the noise model (equation 7) should estimate a realistic noise for the simulated radiance. It may have been unclear that the signal level is taken into account. The reference SNR is determined for each measurement (simulated and real) from the averaged radiance between 757.7–758 nm using calculations of the SNR vs. level 1B calibrated radiance performed by EUMETSAT. We clarified this issue in the revised manuscript.

p.12186,sect.3.6: Why not simply use the noise errors reported in the level-1b products?

We agree that it would be best to do so 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).

p.12186,eq.9: Please state explicitly that SD is the standard error of the mean. (Then, formula can perhaps also be left out.)

Thanks for this comment, the equation has been moved to the methodology section and renamed according to its denotation.

p.12189,sect.4.2: Backward elimination is not used here and the forward model is still non-linear, right?

We believe that our manuscript was not concise enough in this section and has been revised accordingly. Actually, the entire retrieval approach including stepwise model selection and linear forward model was applied here. Please note that this section has also been extended in order to enhance the confidence and to proof that systematic effects can be excluded. Therefore, a figure of retrieved minus input SIF in dependence of different illumination angles, water vapour contents and aerosol optical thicknesses has been added (new Fig. 6).

p.12190,l.25: You mention the danger of overfitting and fitting of noise a couple of times, which is the reason why you introduce backward elimination. Can you substantiate and illustrate the claim that you are overfitting for your SIF retrieval methodology? The fact that PCs are thrown out according to the BIC supports this claim, I believe, but it would be nice if you could illustrate the danger of overfitting with some insightful plots for an example case before you actually introduce backward elimination. Alternatively, a short discussion of eigenvalues and noise levels could perhaps also be insightful.

It might be worth to note that the backward elimination has been primarily introduced to provide a solution for the arbitrary selection of the number of state vector elements in the data-driven approach. Furthermore, it has to be mentioned that there is always a risk of overfitting when no model selection criteria is applied. We agree that our statements needed to be substantiated and the Sect. 4.3 "Influence of number of PCs used and backward elimination" should have done so. We realize that this was not the case. Thus, we modified the figure 7 in the revised manuscript. In particular, we now present selected PCs/coefficients, bias, standard deviation and BIC because these are the most important plots, which are necessary to show the dependency on the number of PCs used and to explain the advantages of the backward elimination. The other plots have been removed in order to make the manuscript more clear and to the point as suggested from the reviewer. In addition, we present an error correlation matrix (new Fig. 8) to prove that the PCs with the largest explained variance are selected and to show the validity of the assumption that the reference fluorescence emission spectrum (hf) has no significant correlation to any of the provided PCs.

p.12189,sect.4.3/fig.5: About the bias: I don't understand why the retrieved SIF is always overestimating the input SIF?

Please note that the value (if backward elimination is enabled) is close to zero. Nevertheless, in the revised manuscript we discuss several implications from our ground to sensor transmittance estimate, which could result in an overestimation of SIF. Nevertheless, we consider that this effect is very small.

About correlations: If the backward elimination mode performs better, the correlation should be higher, right?

We removed this plot, but would like to mention that both correlations were above 0.99. It must also be stated that the description was incomplete: we used TOC averages to calculate the correlation coefficient (which is actually not particularly meaningful here).

About SD: what is SD here? You say 'the SD of retrieved SIF values' in the caption, but I guess it is the standard deviation of the difference between retrieved and input SIF? (Standard deviation of retrieved SIF values is not meaningful here.)

We have to admit that the description was not complete at this point. Indeed, we present the standard deviation, but it must also be stated that this is the average of standard deviations from SIF retrievals of the 60 TOC SIF spectra under different atmospheric conditions. Hence, this value represents the precision of the retrieval. It would also be possible to show the standard deviation of the difference between retrieved and input SIF, but this would rather be an assessment of the accuracy than of the precision. This is already done by showing the bias. We revised the manuscript in this regard.

Same for RMSE? So, in this figure bias**2 + SD**2 = RMSE**2 should hold? From the figure, this does not seem to be the case, but I might be wrong.

In this case, the mentioned relation should not be valid, because of the averaging (see last comment) of the standard deviation. Apologies for the inaccuracy in our explanation, which has led to your reasonable criticism.

Furthermore, I doubt whether the comparison between the backward elimination mode and the linear model mode is presented properly. I find it confusing anyway: The number of PCs (x-variable) in the former case is the initial number of PCs that are input to the backward elimination pre-processing step. In the actual retrieval, less PCs are used (how many? are these indeed the PCs with the largest 'explained variance'?).

Great thanks for this comment. It may indeed have been somewhat confusing. Hence, the first plot of the revised figure now shows the number of selected PCs and coefficients. We also added the x-axis label "provided PCs" to avoid any confusion. The caption has been revised accordingly. You can also find a new figure (Fig. 8) in the revised manuscript which depicts an error correlation matrix of a sample fit. We believe that this figure can prove that in general the PCs with the largest explained variance are selected. In this context, it should be noted that only a few PCs (about 5, exact number depends on the fitting window) can explain a very large amount of variance (>99.999%). It might also be worth to note that our approach suggests to use significant less PCs than Joiner et al. (2013). We compare the results to the retrieval using all coefficients to empasize that the backward elimination is able to enhance the accuracy and precision regardless of how many PCs are used.

The number of PCs (x-variable) in the latter case however is the number of PCs actually used in the retrieval. In backward elimination mode, the number of PCs selected for the actual retrieval step should saturate, as you suggest in the text, and from that point on it should not matter whether you further increase the initial number of PCs. Thus, from the saturation point onwards, I would expect data points in fig.4 to be basically repetitions (the actual retrieval is perfectly the same). The figure suggests otherwise (independent data points). This should be explained more clearly (why is there still variability in data points for the backward elimination mode for large numbers of initial PCs?).

Your assumption is true, the number of selected PCs saturates at about seven. In principle, you are also right to assume no variability in data points for large numbers of initial PCs. Basically, this figure also reflects this expectation – however, admittedly there is still a low variability. The reason can be found in the backward elimination algorithm itself: it would be necessary to test all possible combinations of model parameters to find the 'best' fit, which is computational too expensive. For this reason, we perform a stepwise model selection. We added a brief paragraph to explain this in the present

manuscript as suggested.

Continuing this point, an alternative and perhaps more appropriate comparison would be a comparison of retrieval outcomes for the two modes when the same number of PCs is used in the actual retrieval. For example, SIF biases when using eight optimally chosen PCs in backward elimination mode and when using the first (in terms of explained variance) eight PCs in (non?)-linear model fit mode. Then, we have the same variable on the x-axis for both modes. To me, the backward elimination step to determine the optimal set of PCs really is a pre-processing step and not part of the actual SIF retrieval.

We agree that this might also be an interesting comparison, but here we focus on differences arising from the stepwise model selection. Furthermore, both retrieval modes have the same starting point so that the x-axis is appropriate from our point of view. Even if the backward elimination step can be regarded as a pre-processing step, we have not used any formulations which suggest otherwise.

p.12189,sect.4.3: Can a figure be included showing how many PCs on average are eventually selected by the backward elimination step? Otherwise discuss this.

Great thanks for this comment, we included such a figure (revised Fig. 7) and a discussion in the revised manuscript as already mentioned in the penultimate reply.

p.12192,l.6-7: should amount at least eight days -> 'should amount to at least eight days';

Please note that we removed this sentence because this statement might have been subjective.

please give reference p.12193,l.13-14,fig.7: unbiased relationship -> Please provide statistics to support the claim that the NDVI-slope is not significantly different from one, whereas the SIF slope is. The NDVI-slope is larger than one, which agrees with the hypothesis of cloud contamination.

The NDVI comparison has been removed due to several other comments.

p.12192,l.11-p.12193,l.22: Scia SIF values are lower than G2 SIF values. You investigate the effect of clouds using the NDVI, which is a rather indirect approach. Can't you use cloud data directly? For example, cloud fractions from FRESCO or OCRA (don't know actually which ones you used). I think you should also discuss that a difference is not expected because the overpass times are so close together. (I am not an expert on this, but I would say that an hour difference is too small to see differences in vegetation activity. True?)

We fully agree with the reviewer here. Please consider that the data basis to compare GOME-2 with SCIAMACHY SIF was somewhat weak. This issue has greatly improved since we processed the SCIAMACHY data for the 08/2002-03/2012 time span. For this reason, we are able to compare long-term averages of different SIF data sets in the revised manuscript. Instead of the NDVI we present a direct comparison of FRESCO cloud fractions in order to evaluate possible differences in cloud contamination (new Fig. 10). Furthermore, we use TOA reflectance in order to prove if there is a

relative radiometric offset, which is most likely not the case. We explicitly mention in the revised version of the manuscript that a difference is basically not expected since overpass time differs only half an hour.

p.12194,eq.11: missing ˆ2 in formula.

Thanks for pointing that out. We revised the equation accordingly.

p.12199,l.8: SIF signal is absorbed -> shielding or extinction, cloud droplets do not absorb in this wvl range (throughout paper).

We fully agree with the reviewer and went through the manuscript to correct this issue.

p.12208,fig.4: Is SD the same as SD_Fs in eq.9?? I guess not. Why then use an acronym for a plain and ordinary standard deviation here?

We agree that it has been confusing to use an acronym here and revised it accordingly.

sect.5.: I have stopped doing a detailed analysis of sect.5. As an overall remark, I encourage the authors to rethink also this section and ask themselves what point they are trying to convey here.

Thank you very much for the detailed comments until section 5, which helped significantly to improve our manuscript. However, we also substantially revised this section, which has already been mentioned in the replies.