Response to reviewer 2

No comprehensive description of their method is provided

The section describing the fitting procedure (section 2.2.3) has been re-written and expanded upon, we hope that it is now clear.

Page 6, lines 27-29: "A model spectrum is then built on the high resolution model grid, which typically has a spacing of 0.01 nm..." How is this model spectrum build? By a convolution with a Gaussian ILS with a FWHM of 0.01 nm?

The model grid is simply a 0.01 nm spaced grid onto which all reference spectra are interpolated by cubic spline. The model spectrum is the result of inputting the fitted parameters into equation 9. The fitting process is now described in more detail, and hopefully more clarity, in section 2.2.3.

How are the effects of the dark current and "bias" determined and corrected? Remark: the latter is typically called "offset" rather than "bias" (see e.g. Platt and Stutz, 2008).

This is explained in section 2.2.3. Thank you for highlighting that fact, we now refer to it only as the offset signal.

Flat spectrum: Esse et al. retrieved the "flat spectrum" by a simple averaging. In contrast, Lübcke et al. (2016) retrieved the instrument effects (flat spectrum but also further instrument effects such as temperature effects) by a Principal Component Analysis (PCA). For me, the PCA approach appears to be more comprehensive. See also my comment 2.3. Please motivate the choice of a simple averaging instead.

The choice of averaging was to directly measure the instrument properties rather than retrieve them from measurement data. This was chosen as it does not rely on selecting measurement spectra for a training set, rather it can be premeasured in the lab before measurements take place. Other instrument effects could remain and it would be possible to extract these separately from the flat spectrum using a similar method to Lübcke *et al* (2016) but this would remove one of the main benefits of the iFit method. This has been explained in the text (p. 6 l. 1).

Background polynomial: Is equation (10) correct? Physically, all light attenuation effects are on the same footing, i.e. both absorption and scattering effects are summands in the argument of the exponential function. In principle, the scattering effects can of course be written in the presented way e.g. as $P(\lambda) = \exp(P^*(\lambda))$ where $P^*(\lambda)$ is the "real" broad-band scattering polynomial. But this is strictly different from the polynomial as it is used in DOAS. In particular, its coefficients will be different to the polynomial $P(\lambda)$ denoted in equation (3). Please clarify.

The discussion of the background polynomial has been expanded to clarify this point (p. 5 l. 11). A single polynomial function is used to account for a number of effects including the Mie and Rayleigh scattering as well as the transmission of the optics.

Ring spectrum: First, I have analogous doubts concerning $R(\lambda)$ in equation (10) and equation (9). Second, please make more explicit how is the Ring spectrum retrieved.

The Ring spectrum was correctly applied in the forward model, but we agree that the description was inconsistent between the two equations. This has now been updated. The ring is retrieved as part of the overall fit (p. 6 l. 18 and table 1).

Sky spectrum: How is it constructed? By adding the absorption effects of background O_3 and NO_2? If yes, how is the correct background amount determined?

This description of the fitting process was perhaps confusing, and so has been changed. Hopefully it is now clearer. All parameters given in table 1 are fitted simultaneously.

Plume spectrum: Is also O_3 included in the fit step from sky spectrum to plume spectrum? If not, how are the diurnal variations of stratospheric O_3 contributions and assessed and corrected? Furthermore, why is particularly BrO (but no other gases) included in the SO_2 fit scenario? The BrO absorption is rather negligible in the typical SO2 fit ranges.

As above this comment stems from the description of the algorithm, which was perhaps confusing. O_3 is included in the fit to account for the changes in path length throughout the day. Fitting the BrO spectrum is indeed redundant at these wavelengths, and so has been removed when reanalysing the data.

Instrument line shape function ILS: "The ILS was measured using a mercury lamp to be 0:50 (±0:01) nm" (page 8, line 27). The uploaded data does provide ILS (only) at 302nm (instrument H15972) with a FWHM of 0.58nm and at 301nm (instrument FLMS02101) with a FWHM of 0.60 nm. Please provide the full mercury spectra for a presented instruments. (Remark: both uploaded ILS have indeed an about Gaussian shape. A super-Gaussian model proposes exponents of (only) 2.3 and 2.1 and the asymmetry is rather small as well.)

The ILS is a unique property of the instrument, although it varies in general with temperature and wavelength. Accordingly, for a given instrument (and similar temperature) all convolution operations have to use one identical ILS. Applying different ILS can cause a significant decrease in accuracy. Furthermore, all compared spectroscopic should apply an ILS retrieved at the same wavelength in order to be consistent. Ideally, this wavelength is chosen in the wavelength range, e.g. at 315 nm. For practical reasons, the mercury line either at 302nm or at 334nm should be chosen. Please clarify which measurement results for the ILS are used. I propose to add a table to the manuscript which lists all instruments used in this study and their spectroscopic properties.

However, later a modelled ILS with "an ILS width of 0.56 nm" (page 10, line 9) has been used instead of the measured ILS. Please clarify why instead of the exact measurement results an apparently wrong ILS is used at this step.

One change to the updated manuscript was to replace the Gaussian ILS with a super-Gaussian. Figure 4 now shows the Hg spectra and fitted 302 nm line and the instrument properties are summarised in table 2. The fitted super-Gaussian ILS is now used for all analyses.

How are the wavelength shift and stretch determined and corrected?

The wavelength shift and stretch are fitted as part of the forward model. This has been clarified in the text (p. 6 l. 18).

Furthermore, physical-logical the wavelength shift and stretch should be actually applied on the measured spectrum in order to correct for temperature-driven variations of the instrument during the measurement. Analogously, the "flat spectrum" should be applied on the simulated spectrum in order to correct for the instrument effects of the real instrument. I can imagine that these inconsistencies are mathematically identical and may thus lead to the same results. Please clarify why/whether these inconsistencies are required.

We disagree with this statement, as to include the shift in the forward model as a fitted parameter it needs to be applied to the model grid, not the measurement grid. The shift could be predetermined and applied to the measurement, however it was found that fitting it was more robust. The flat correction should be applied to the measurement as it is a property of the instrument pixels, not the wavelength. This has been emphasised in the text (p. 6 l. 15).

What actually does "perform fit"? Is it an ordinary DOAS fit? Or is it iteratively minimising $\tau = log\left(\frac{I_{right\ hand\ side}}{I_{left\ hand\ side}}\right)$ ***by means of varying the SO2 column density?***

The measured intensity spectrum is fitted by the forward model by minimising the residual between the two (in intensity). The fitting algorithm has been described in more detail in the text and is now hopefully clearer (p. 6 l. 23).

Is a stray light correction applied?

Yes, using the intensity between 280 and 290 nm. This is explained in the text (p. 6 l. 13).

Missing comparison with the state of the art

The manuscript is motivated by the possible underestimation of the SO_2 ***slant column density in a volcanic gas plume. However, no iFit results for such a scenario have been provided. Please explain this inconsistency.***

Data from the ENIC scanning station of the FLAME network has now been included as an example when this is the case (section 4.2.2). From this a clear underestimation due to the contamination can be seen.

According to the list of literature provided in the manuscript (and to my knowledge), the approach from Lübcke et al. (2016) is the current state of the art to face such background contamination issue. I highly recommend a direct quantitative comparison of these two methods when applied on the same data (ideally contaminated data) in order to reveal the major differences.

Although we see the merit of a direct comparison, we believe that this would be beyond the scope of this manuscript as it would require developing a full analysis routine for the method provided by Lübcke *et al.* (2016). We believe that the advantage of having a "point and shoot" retrieval method is a significant advantage when there is no large training dataset with which to perform the PCA. Conversely, the method presented by Lübcke *et al.* (2016) could be preferable when access to the spectrometers is not possible, and so the flat spectrum cannot be characterised. These points have been further emphasised in the text (p. 10 l. 23).

Esse et al. propose to use iFit for evaluating data recorded by permanent monitoring stations. Monitoring stations are typically not temperature controlled in order to improve their robustness

and to lower their power consumption (see e.g. Galle et al., 2010). iFit has been tested exclusively for temperature stabilised instruments and Esse et al. concluded that "care must be taken if the spectrometers are not temperature stabilised" (page 11, line 9). Accordingly, Esse et al. have not provided evidence that iFit is suitable for monitoring stations. This is in particular in sharp contrast to the approach from Lübcke et al. (2016) which presented their results for contaminated data from monitoring stations.

None of the measurements shown here were temperature controlled. We also now present additional spectra from the FLAME network on Etna showing contaminated data (section 4.2.2).

I agree with Esse et al. that also a comparison with standard DOAS (recorded background spectrum) appears to be mandatory. For the arguments stated in the very first paragraph, I expect that a standard DOAS approach performs in general (i.e. for non-contaminated scenarios) better than iFit. In contrast, the narrative in the current manuscript is rather one-sided, highlighting possible problems in the DOAS approach only. I highly recommend that any subjective valuations are neglected from the technical manuscript parts (e.g. "methods", and "results"). Differences between iFit and DOAS should be discussed later in the "discussion" part where all evaluating statements should be supported by (quantitative) evidence.

I agree with this statement that where an uncontaminated reference spectrum is available and the appropriate corrections (e.g. I_0 effect) are applied then the standard DOAS algorithm will outperform iFit as iFit implements a relatively simplistic radiative transfer model to fit the measured spectra. This is shown and commented in the comparison between iFit and DOAS (p 10 l. 6).

Esse et al. conclude "the lack of a requirement for a reference spectrum means that iFit would be especially well suited to deployment in permanent scanning stations" (page 11, line 14). Permanent scanning stations scan typically from horizon to horizon and thus automatically recorded the reference spectrum. Applying iFit thus does not provide any gain in measurement time in particular at permanent scanning stations. Furthermore, probably only few permanent measurement stations are at all affected by background SO_2 contamination. Accordingly, possible benefits from iFit are limited to those stations. Please limit your conclusions with respect to those scenarios where you can provide evidence that iFit at least does not perform more poorly than the alternative approaches.

We have adjusted this statement to reflect the fact that iFit favours automated analysis as no definition of a reference is required. This is not a comment on the time taken to perform the scan, but the removal of potential sources of error. We agree that not all stations will often suffer from contamination, but would like to emphasise that all stations could (and likely will at some point). Therefore methods which use a synthetic reference are preferable.

Page 4, lines 11-19: Is there any need to discuss the option of a high-pass filter? Is a high pass filter used in iFit or for the DOAS retrieval in this manuscript? If not, this paragraph appears to be redundant. The figures 4, 5, 10 show results of the total absorption cross section.

The high pass filter was only included to simplify the equations – however we agree that it could confuse the reader as both iFit and DOAS analyses presented do not actually use one. It has now been removed and a broadband polynomial included in the equations.

Page 4, line 26: "The ILS can either be a mathematical function (such as a Gaussian)..." The ILS is a property of the instrument and has in general an arbitrary shape. Although it can be indeed often approximate in good agreement by a Gaussian line shape function, the real ILS itself is not a mathematical function!

This was a poor choice of wording, and has been changed in the manuscript to state that it is a property of the instrument but can be approximated with a mathematical function.

Page 4, lines 28-31: Equation 6 is not true! The convolution operation and the scalar multiplication operation are commutative!

Thank you for highlighting this error, it has been removed from the manuscript.

Page 6, lines 29-30: "A wavelength-shift is a common correction in DOAS...". This is correct, however, when the spectra are wavelength-calibrated prior to the DOAS fit this shift is typically in the order of 0:001 nm. Do Esse et al. refer to the additional wavelength shift parameter which is usually allowed between the measurement spectrum and the absorption cross-sections in order to partially compensate for the convolution with a (slightly) wrong ILS? Anyway, this wavelength shift is typically limited to ±0.2nm rather than ±2 nm. A wavelength shift of 2nm appears to be absurdly large. Please clarify.

The 2 nm padding is included for two reasons: firstly to accommodate the wavelength shift (which as the reviewer points out is typically small) and secondly to avoid the edge effects incurred by the convolution of the model spectrum with the ILS. This has been clarified in the text (p. 6 l. 19). Typical retrieved shifts are of the order of 0.1 nm.

Page 7, line 27: "The wavelength region used for these results (304 - 320 nm) is common to most scattered sunlight retrievals of SO2." This is not true. In DOAS - the predominant remote sensing technique for volcanic SO2 - the used wavelength range starts almost exclusively at 310nm (e.g. Lübcke et al., 2016), 312nm (e.g. Theys et al., 2017; Kern and Lyson, 2018), 314nm (e.g. Lübcke et al., 2014; Dinger et al., 2018), or 326nm (e.g. Hörmann et al., 2013). The reason for these lower limits is, that the applied approximation in DOAS are only justified as long as the "absorbance" is not much above 0.1. For the data presented in Figure 10, this means the DOAS retrieval should start not lower than at 314 nm. This limitation does not have to hold for an intensity based fit, however, Esse et al. have to make sure that all presented DOAS data are retrieved for an absorbance below 0.1. Otherwise their DOAS results would be too low and a quantitative comparison between iFit and DOAS therefore flawed.

This comment was meant to reflect the general wavelength region commonly used (e.g. near to the absorbance features of SO_2) not the specific wavelengths used in retrievals, but we can see how it is misleading and so this phrase has been clarified. As already mentioned we also now perform the fits between 310 – 320 nm which is more in line with standard DOAS retrievals.

We also realise that the displayed absorbance features on figures 4, 5 and 10 were actually optical depth (as described in the text) not absorbance, we apologise for this error. The figures have been updated. The peak absorbance value during the traverses was 0.12, but typically it was ~ 0.07 in the plume.

Page 8, line 31: "In particular, use of higher wavelengths leads to an overestimation in the retrieved SO2, possibly due to the reduced strength of the SO_2 absorption spectrum at higher wavelengths". This interpretation of the findings is not supported by further evidence. Furthermore, the "reduced strength of SO_2 absorption" can be expected to result in a larger fit error but there is no obvious reason why this should cause a less accurate result. In fact, I would interpret the findings other way round: the lower the wavelength the larger is the underestimation in SO_2 due to saturation effects and a decreasing solar background radiation (see Platt and Stutz, 2008). At least for DOAS the "absorbance" should be below 0.1 in order to keep the applied approximations justified. With these fundamental limitations in mind, I consider the results for the wavelength range from 310-320nm the most accurate. Furthermore, I expect that starting at 314nm or 326nm would give larger and even more accurate results in particular for the cells above 500 ppmm. In consequence, iFit would overestimate the SO_2 slant column density. Please provide evidence for your interpretation. In particular, I highly recommend to present (additionally or exclusively) DOAS results when the wavelength range starts at least at 310 nm.

The discussion of the cell data have been rewritten. For iFit use of higher wavebands results in a large offset in the retrieved SO_2 SCD due to the program supplying an SO_2 SCD to fit features in the residual. This discussion has been included in the updated manuscript (p. 8 l. 22).

Some minor formal objections

Inconsistent use of brackets (e.g. compare equation 7b and equation 8)

This has been corrected to ensure the equations are consistent.

The (slant) column densities are sometimes denoted by a_i and sometimes by α. They have to be denoted by the same consistent letter throughout the manuscript. Furthermore, they should always hold the index.

The column densities are now all referred to as a_i with the index.

I_0 and I^*_0 appears in several forms throughout the manuscript. They should be consistently denoted by a strictly constant sign.

This has been corrected.