We thank the referee for taking the time to carefully review the manuscript and for many helpful suggestions.

I acknowledge the presented concept, but at the same time I see the following critics which should be solved before the final publication (please find the motivation background within the detailed comments later):
(1) Algorithm description: the algorithm is rather poorly described. Neither there is structure of the algorithm sufficiently presented with clear references for details nor are the implementation of the new improvements (in particular the implementation of the accounting for the diurnal gradient) clearly and in detail described.

The algorithm description has been improved. Basically, the improvements consist of a clearer presentation as well as additional details and references. The responses to the detailed comments below will be more specific about the improvements.

(2) Besides that (but closely related) there is practically missing scientific discussion on the importance of the new improvements to the scientific community. It is not clear if the newly introduced concepts are new also outside the author team, are there alternatives, are they applicable or have perhaps already been applied for other algorithms, models, instruments?

In terms of the new diurnal chemical gradient capability in Sasktran, there is one comparable alternative, the VECTOR model (McLinden et al., 2006), although Sasktran is a fully spherical RT model. A fully spherical RT model with diurnal chemical gradients has not been used for NO_2 retrieval.

We now modify the wording in the abstract as follows:

The diurnal variation of NO_2 along the line of sight is included in a fully spherical multiple scattering RT model for the first time. Using this forward model with built-in photochemistry, the scatter of the differences relative to the correlative balloon NO_2 profile data is reduced.

Related changes to Sect 2.2 are discussed below.

(3) validation study: there are either conclusions made based on a single profile or on a very general statistical behaviour without discriminating between type of balloon instrument, its measurement/retrieval technique, location, season, time (SZA) etc.

These points are addressed below in response to detailed comments. But a response will be given here a well: the original manuscript was already clear that no conclusion was drawn based on a single profile. There are two single profiles that are shown as examples (Fig.2 left) and Fig. 3. However, Fig. 2 (right panel) demonstrates the benefit of improved precision in a statistical sense (i.e. large sample size).

We now elaborate on the use of an extended fitting window in the opening sentence of Sect. 3 as follows:

One difference in the method described above (Sect. 2) compared to previous OSIRIS NO_2 algorithms (Haley and Brohede, 2007; Bourassa et al., 2011) is the use of spectral information at longer wavelengths which allows the NO_2 absorption optical depth to be more precisely quantified.

Figure 3 is also a sample case. At p9L17, we now write "The overestimate for this sample case…"
In this same paragraph (P9L24), the reader is already made aware that the conclusion about improved NO_2 retrieval precision using the 2D model (i.e. with diurnal chemical gradients along the line-of-sight) instead of the 1D model (i.e. no horizontal chemical gradients) is supported by various statistics obtained using a large sample of coincidences.

Also it happens often that too much positive conclusions are suggested although plots indicate that it is not always the case.

After >15 years of forward model and retrieval algorithm development, the improvements are not expected to be large. In one case, the improvement is stated as being slight (p9L5). This general comment is addressed in the responses to the detailed comments below as well, but we summarize here that the reviewer:

1) in one case, looked at the magnitude of the mean biases, rather than their statistical significance. Statistical significance accounts for the existing variability (or scatter), not just the mean.
2) in a second case, incorrectly estimated the size of the error bars at the top of Fig. 2 (right panel).

Detailed comments:
P1, L19 "LOS column densities": is the definition the same as for slant column densities (SCDs) used later? If yes, be consistent, otherwise provide definition.

To be consistent, we now use slant column densities in the abstract.

Introduction, 1st paragraph: The statement needs references.

There are multiple statements in this paragraph, so we have provided references for all of them.
We now write:

Nitrogen oxides, such as NO and NO_2 are the reactive nitrogen-containing species in the middle atmosphere and are produced mainly from the breakdown of nitrous oxide in the stratosphere (Crutzen, 1971). Oxides of nitrogen dominate ozone loss in the middle stratosphere, whereas in the lower stratosphere, they react with oxides of chlorine and bromine, such as ClO and BrO, to reduce the halogen-catalyzed destruction of ozone (Salawitch et al., 2005; Wennberg et al., 1994).

P2, L17: be more specific what gradients do you aim to consider (nadir instruments also sees strong gradients). Maybe you want to say that a limb instrument has in addition to deal with horizontal gradients.

We now write:

The photochemistry of NO_2, which is particularly rapid near the day-night terminator, leads to horizontal gradients within the field of view, particularly for limb sounders such as OSIRIS …

P3, L18-19 "The accuracy is due to two main factors: very high signal-to-noise: : :":

Higher signal to noise ratio mostly improves precision (not the accuracy), doesn't it?
We use the following definition for accuracy used in the fields of chemistry and physics, taken from the Random House dictionary: "the extent to which a given measurement agrees with the standard value for that measurement."

Thus, accuracy is not simply a lack of bias, it also involves precision. In other words, individual measurements are not accurate even if the mean of these measurements is unbiased. No change is made to the revised manuscript.

Last sentence P3 "occulation/limb occultation": skip occultation on P4, L1

We have changed "occultation" on P4L1 to "geometry".

Sect. 2. The basic algorithm description is limited to 1(!) sentence, not to say that the chapter 'methods' starts right away with subsection "Algorithm settings". To what then these settings are applied? Please provide background, an algorithm diagram would be helpful.

The "background" is provided in the introduction (P3L9-14). There should be no need to rehash the detailed algorithm description that is available in the papers cited in the introduction (e.g. Sioris et al., 2003; Sioris et al., 2007). The changes included in the 'new' algorithm are what should be described in this manuscript in addition to a brief overall description. P3L9-14 in Sect. 1 should be clearer as to the improvements relative to the existing literature. To this end, we now elaborate as follows:

The current algorithm was developed to demonstrate that improved accuracy is possible through the combination of a better forward model and better forward model inputs than used by Sioris et al. (2007), and additional, longer wavelengths than those selected by Sioris et al., 2003. The main focus is on the lower stratosphere (and upper troposphere).

The basic algorithm description is not limited to one sentence. The algorithm consists of three main parts:

1) spectral fitting (which is described in P4L9: "multiple linear regression")
2) forward model (specified in Sect. 1, namely "SASKTRAN", with appropriate references, P2L26)
3) inversion approach (specified in Sect. 1, namely "MART", with appropriate reference, P3L7)

It seems the main problem is that the title of Sect. 2.1 should not be "Algorithm settings". We have renamed it "Algorithm description and settings". With regard to an algorithm diagram, we provide a reference to Haley et al. (2004) in the second sentence of Sect. 2.1:

"A generally appropriate diagram of the algorithm appears in the work of Haley et al. (2004)."

P4, L6-8: Also this statement requires references. The previous work of "many groups" must be cited. From one side the methods and previous work cannot be known for unfamiliar readers, from another

side the authors should give proper credit to previous work done in the field. When filtering in particular with regard to NO_2 and limb satellite and balloon measurements the reference list is not long anymore.

We now refer to Ogawa et al. (1981) in the opening sentence of Sect. 2.1. Only some examples are given of balloon-based NO_2 vertical profiling via spectroscopic techniques. There are dozens of papers when applying the filters suggested by the reviewers, so no attempt is made to be exhaustive. Limb-viewing satellite instruments making profile measurements of NO_2 include MIPAS, ACE-FTS, ACE-MAESTRO, SCIAMACHY, OSIRIS, SAGE II, SAGE III, SME, LIMS, etc. This list is also too long for this paper. We now write the following sentence only about NO_2 limb-scattering satellite instruments in the introduction:

Besides OSIRIS, other space-borne limb scattering instruments that have measured NO_2 vertical profiles include Scanning Imaging Absorption spectrometer for Atmospheric Chartography (SCIAMACHY) (Bovensmann et al., 1999), Solar Mesosphere Explorer (Mount et al., 1984), and Stratospheric Aerosol and Gas Experiment (SAGE) III (Rault et al., 2004; Rault, 2004).

In order to cite previous work by the many balloon groups, we have selected some very relevant examples and now modify the sentence at p4, L7-8 as follows:

This 2-step approach is used by many groups (e.g. Pommereau, 1982; Ferlemann et al., 1998; Renard et al., 2000) including some of the balloon remote sensing teams providing data used in this study.

P4, L9 What spectral fitting technique is used, is the fit done in optical thickness domain (e.g. DOAS) or radiance domain, or...? Some paragraphs above there are provided a number of references of author's previous work. Please cite here the paper where the particular part of the algorithm is described in detail. And again, it is clear that brief introduction to the algorithm would be helpful, otherwise the reader must search through at least the 4 author's papers cited above to puzzle the basic algorithm buildup.

We definitely agree with the reviewer. We now write in the opening sentence of Sect. 2.1:

The algorithm is a classic 2-step approach of spectral fitting of optical depth spectra with absorption cross-sections to determine slant column densities (SCDs), followed by vertical fitting to invert the SCD profile (e.g. Ogawa et al., 1981).

The following sentence is added to the opening paragraph of Sect. 2.1 to add to the brief overall description of the algorithm:

The profile is retrieved by iteratively updating it based on MART such that the simulated NO_2 SCDs agree with the observed ones.

We also include the following sentence in the next paragraph to begin the description of the spectral fitting:

The first step of the data analysis, namely the spectral fitting, involves a reference spectrum as in Eq. 1 of Sioris et al. (2003).

P4, L15 “Water vapour absorption is neglected despite maximal absorptions of >0.1%”:
There is an ongoing discussion about water vapour absorption on short wavelengths (e.g. Lampel et al., AMT, 8, 4329, 2015). On what source, absorption database, cross sections etc. the considerations of authors are based? Have you investigated effect of possibly different placement of absorption lines and their strengths?

We use HITRAN 2008. We have not investigated the effect of different absorption line positions and strengths. We now write:

Water vapour absorption is neglected despite maximal simulated absorptions of >0.1% in the fitting window (434.8-476.7 nm) in the upper troposphere as it is not spectrally correlated with NO_2 absorption over this window and would greatly increase the forward modelling computational burden. Note that SaskTran treats water vapour as a line-by-line absorber assuming a Voigt line shape and spectroscopic parameters from the HITRAN 2008 database (Rothman et al., 2009).

P4, L21 at this stage one can puzzle out that an iterative algorithm is used. Again please make the basic algorithm setup clear. E.g. what is done by simulated and measured radiances, fit results, perhaps a diagram might help.

The algorithm is now described as being iterative in the opening paragraph of Sect 2.1 as mentioned above. The number of iterations is specified later (p5L28).

P4, L26 what is SCIAMACHY? The instrument must be introduced already earlier as an example of another limb sounding satellite instrument since the OSIRIS is not the one and only limb instrument. In fact there are not so many limb sounding satellite instruments that it could be worth to introduce some of them, at least those, the authors are citing.

SCIAMACHY is now mentioned as a space-based limb scattering instrument that measured NO_2 in the introduction (see above).

P5, L10; P6, L1 “Odin” -> “OSIRIS”

No change is made since the satellite (Odin) scans the limb, not OSIRIS. However, we noticed that we failed to mention that OSIRIS is onboard Odin in the introduction, so we now write on p1:

The photochemistry of NO_2, which is particularly rapid near the day-night terminator, leads to horizontal gradients within the field of view, particularly for limb sounders such as OSIRIS (Optical Spectrograph and Infrared Imager System) (Llewellyn et al., 2004) on the Odin satellite, …

P5, L23 Perhaps add “since at least” before the reference in the brackets

“At least” serves the same purpose as “over”, so no change is made. Perhaps the reviewer’s comment was misunderstood.

Same sentence as above: “to recover : : : along-track distribution : : :”: While the retrieval of vertical distribution is described in authors’ earlier publications, this aspect for their algorithm is very new. It is also not clear from the manuscript if at all and, if yes, how in detail the along-track distribution is recovered. The text below in this subsection just suggests that one is inquiring (1D) profiles with a 2D or 3D photochemistry correction. If it is like that then citing here a more general publication where 2D fields are inquired from correlated scans along orbit without any comment is misleading.

This is another helpful comment by the reviewer. It was already implied that we are retrieving single vertical profiles since we described the inversion as “vertical fitting” but to be more explicit, we now write in the opening sentence of Sect 2.1 “to invert the SCD profile” rather than “to invert to the SCDs”. Given this change, we feel that the reader should not be misled into thinking that we are retrieving along-track information using several limb scans simultaneously. Additionally, we have placed “and along-track” in brackets in the sentence on P5L23. The Fesen and Hays reference was chosen since they reviewed the early work on algebraic reconstruction techniques. The work of Thomas and Donahue is cited simply since it is the first paper applying this technique to an atmospheric problem to the best of our knowledge.

P5, L29 “Five orders of scattering are used” Perhaps you mean "up to five orders". Anyhow, please explain what you mean here otherwise it reads as a technical slang.

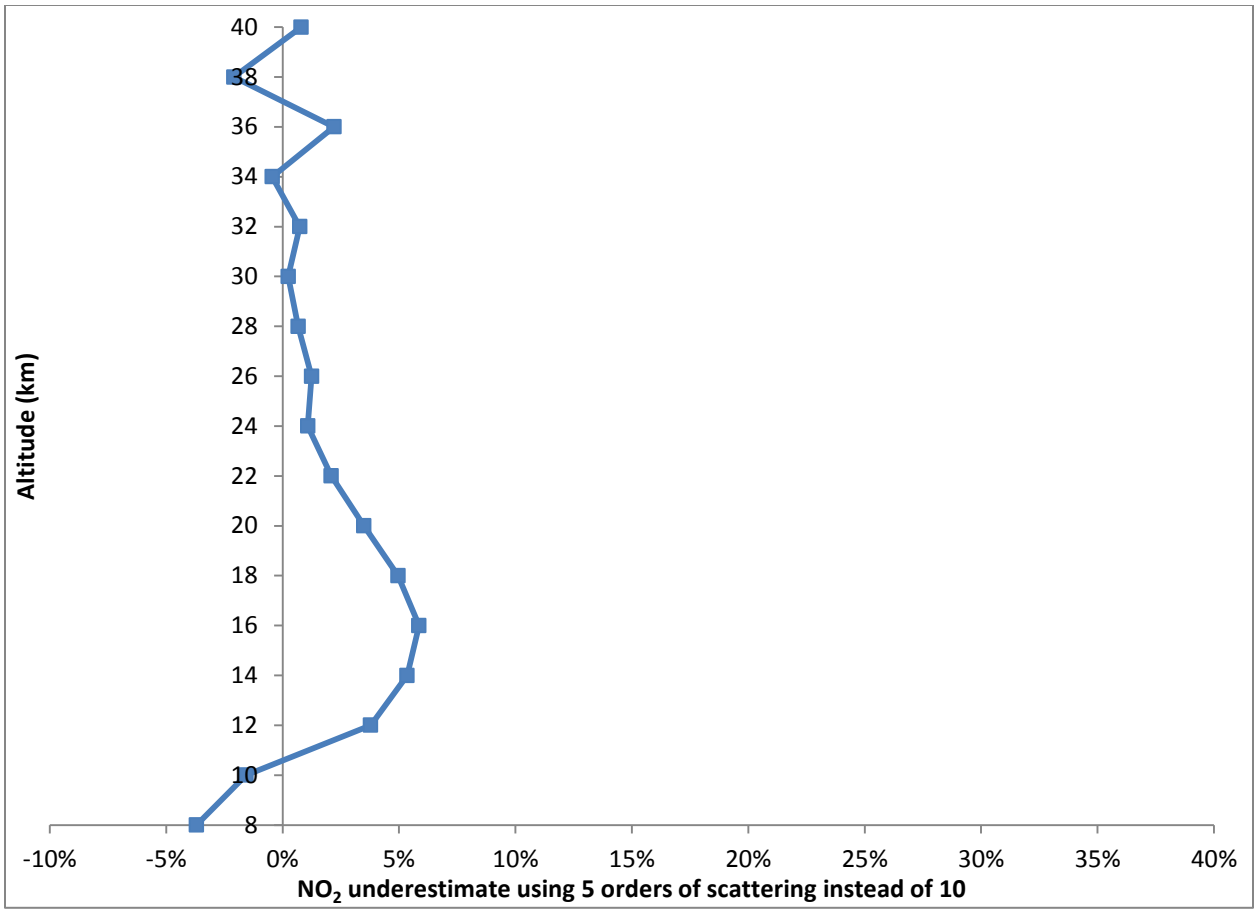
In our opinion, this is not slang and the reader has already been informed in the introduction that SaskTran uses the successive orders of scattering approach. We now modify this sentence:

The radiance is summed over five orders of scattering (Sioris et al., 2004).

Are five orders enough for scenes above (thick) clouds? What is the accuracy then? Or do you filter complete scans for measurements above the cloud scenes out?

Limb scans with clouds below the field of view are not completely discarded. Clouds have a reduced impact for OSIRIS because of the large SZAs and because pre-retrieved scene albedo is used in the NO_2 inversion. However, the reviewer’s question prompted us to study the sufficiency of 5 orders of scattering for cases with thick clouds below the lowest tangent height (TH) used in the retrieval. We find that the retrieval typically has <1.3% retrieval errors in the top half of the retrieval range when using 5 orders of scattering in the forward model of the retrieval yet the pseudo-observations are computed with 10 orders of scattering. However, the figure below shows that the negative bias due to using only 5 orders of scattering becomes larger in the UT/LS region (-6% at 16 km) where higher order scattering can be more important and the inversion is non-linear due to strong extinction and thus retrieval errors are magnified. Note that clouds below the FOV are not taken into account in the NO_2 inversion other than by raising the scene albedo.
We now write:

The radiance is summed over five orders of scattering, which is sufficient for non-cloudy cases (Sioris et al., 2004) and also leads to ≤6% errors over the retrieval range for a case with a solar zenith angle (SZA) of 75° with an optically thick ice cloud occupying two altitude levels within the boundary layer.



Sect 2.2. This short section hides the main improvement of the manuscript. Therefore it is much too short. More details are needed to understand your method and also scientific discussion in the context of the current state of the art is completely missing:
- it is important to say e.g. that photochemical correction capability has been demonstrated already earlier (McLinden et al., JGR, 2006 as an example) even without the 2D or 3D RTM model and discuss what is then benefit (is there benefit at all?) of the 2D or 3D model with respect to the previous corrections? Can now the gradient be accounted better? Why? Are there some studies done in that respect? What are the results?

We did not compare the simulated diurnal chemical gradient in SaskTran to that of VECTOR (McLinden et al., 2006) since the PRATMO photochemical model is essentially the same in both. One of our goals was to add this capability to a fully spherical radiative transfer model, namely SaskTran. Thus, retrievals with 1D and 2D forward model modes of Sasktran are compared rather than comparing with VECTOR. But the main goal of the paper is to demonstrate improvements in retrieved NO_2 relative to the other two fully processed products ('fast' and the operational v3.0 product), neither of which used a forward model with built-in photochemistry. So, improvements relative to an algorithm using VECTOR (e.g. Sioris et al., 2007; Sioris et al., 2004) are outside the main focus of the paper.
Sect 2.2 now begins with:
McLinden et al. (2006) developed the capability to account for diurnal chemical gradients in a pseudo-spherical RTM. The high spatial resolution capability which is required for modelling

(horizontal) diurnal gradients of NO_2 within the fully spherical SaskTran RTM is described by Zawada et al. (2015).

- Also the possibility of simultaneous retrieval of different limb scans along orbit by a tomographic concept must be discussed being a really clear benefit of a 2D RTM model as e.g. already demonstrated for a rather similar retrieval of NO2 by Pukite et al., ACP, 2008 for SCIAMACHY. In such a case even photochemical modelling can be redundant to a large extent. Nothing is said if such a concept can be an alternative for the OSIRIS retrieval or it is even already considered at some retrieval step. If yes, please describe how the benefit by correlating the scans is considered in the described 2D and 3D algorithm and how it matches with the photochemical modelling. Anyhow can you provide more details what you are doing in the new 2D and 3D algorithm, what exactly are implications to the parts of retrieval: forward modelling, spatial reconstruction, iteration process? Is there at least some correlative processing of the profiles obtained for different OSIRIS vertical scans in some of these steps to fully use 2D, 3D model capability or they are processed separately?

As indicated above, the limb scans are processed separately. We now write the following at the end of Sect 2.2:

Note that the typical horizontal sampling of OSIRIS limb scans is too sparse to allow for a tomographic retrieval (Hultgren et al., 2013). A two-dimensional tomographic retrieval of NO_2 from a special set of SCIAMACHY limb scattering measurements with finer horizontal sampling was performed by Puķīte et al. (2008). Their tomographic retrieval provides an alternative approach to account for the diurnal gradient of NO_2 in the orbital plane (i.e. along the LOS).

- The latitudinal resolution of 5 deg looks rather course (following this horizontal distance from TP means that LOS crosses altitude region of _25 km, so almost the whole NO2 retrieval altitude range is within this horizontal region). Can the model assume at least liner variation between the grid points? Otherwise this could be a reason for the rather poor improvement of the 2D corrected retrieval with respect to its 1D alternative? Do sensitivity studies exist regarding this? How the horizontal model grid is matched with measurement and retrieval grid?

As indicated above, the reviewer has misunderstood the original manuscript and we apologize for any confusion. The 5° resolution is sufficient for the photochemical modelling.

P7, L2: “the OSIRIS 2D and 3D profiles are retrieved: : :” if the profiles/OSIRIS scans along orbit are not correlated in any of the retrieval steps (which is not really clear from the few information provided, see one of the previous comments) then they are individual 1D profiles. One then cannot speak about 2D or 3D profiles like it would be if really a simultaneous (e.g. tomographic) retrieval for different scans would be performed. Please avoid misunderstanding and say e.g. “profiles with 2D or 3D photochemical correction”.

We refer the reviewer to p6L15-23. To reduce the possibility of reader confusion, we now write the sentence at p7L2:

Even though the OSIRIS profiles retrieved with the 2D and 3D mode of SaskTran account, to varying degrees, for the diurnal gradients expected for the OSIRIS viewing geometry, photochemical modelling is also required to scale all of the OSIRIS NO_2 profiles to the local time of the balloon measurement.

P8, L28 "1D inversion approach": do you mean retrieval without modelling diurnal gradients?

We apologize for this line, and realize that it could have been very confusing. We now write '1D forward model' instead.

P9, 1st paragraph: The difference in the figure is very small. I doubt that from so few measurements (just 33 scans and so small difference) it is possible to make a clear conclusion that the new fit window is better. Please indicate the sample standard deviation on the figure to clarify this. In order to increase the confidence could the improvement be proved on a much larger dataset? You do not need to stick to profiles matching the balloon measurements here, do you?

Assuming that the vertical resolution is 2 km, there are 12 independent pieces of vertical information in the left panel of Fig. 2. The odds that all layers show smaller standard deviations using the larger fitting window purely by chance is $1/2^{12}$. Thus, we are confident at the 99.9% level that there is an improvement. We consider 99.9% confidence to be sufficient. The reviewer requested the sample standard deviation, but the x-axis of the left panel of Fig. 2 is the standard deviation. We do not need to stick to balloon coincidences here. We simply needed a sufficiently large number of samples (i.e. N=33).

Another aspect: one should be certain that there are no covariance between the measurements and retrieved profiles for such a conclusion to be valid, e.g. if there is some kind of a-priori effect or correlated instrument feature at certain wavelengths (for what reason ever) one also could get lower variability: : :

There is no *a priori* influence in a MART retrieval. As indicated in the paper, the measurements in Fig. 2 come from a variety of latitudes, seasons and years. Furthermore, there are no known systematic artifacts within either fitting window and even if there were, it would also be unlikely that they would correlate with the spectral signature of NO_2 absorption. Thus, we are reasonably certain that there is no covariance. Sample radiance spectra from OSIRIS are shown in Fig. 3 of Haley et al. (2004).

P9, L5 add "left plot" or "left panel" after "Figure 2"

We now write "Figure 2 (left panel)".

P9, L9 the improvement is very small, add "slightly" before "improved"

The improvement is ~20% averaged over the 7.5 to 40.5 km range. We now write "…by 20%...".

P9, L10 "and the upper altitude limit of the retrieval moves slightly higher: : :": It is not said what do the color-shaded regions in Fig.2 mean? Supposing they represent error (one standard deviation), then this statement on L10 contradicts the figure: error for the extended window is larger at the uppermost altitudes. What is the reason for this strange result since one should expect better performance?

We now write in the caption of Fig. 2:

“The retrieval uncertainty for each profile is bound by a shaded area of matching colour.”

There is no contradiction as the reviewer claims. Here are the retrieval uncertainties:

Altitude (km)	435-477 nm window	435-451 nm window
36.5	8.00E+07	9.48E+07
37.5	8.02E+07	9.66E+07
38.5	8.04E+07	9.84E+07
39.5	8.06E+07	9.95E+07
40.5	8.08E+07	1.00E+08

We agree that, at first glance, the reviewer’s observation appears correct. We have also verified that there is no error in making the plot. Since the lower limit of the x-axis is 0, perhaps the reviewer is not realizing that the error bar at 39.5 km for the smaller fitting window is not fully shown since it extends to negative values (error >100%). No change is made to the manuscript.

P9, L11 add “right plot” or “right panel” after “Fig. 2”

We now write “in the right panel of Fig. 2”.

P9, L14 “The significant upper tropospheric NO2 enhancement: : :” The enhancement is not significant if comparing with the error: the relative enhancement for the new fit window might be large at these altitudes but both profiles still lay within the error bars of each other. Therefore it would be good to skip "significant" here. Can you provide more proof that this is generally the case also for other profiles? At least it seems contradictory to other retrieval methods in Fig. 6 and 7 where the retrieval shows negative bias with respect to the validation dataset.

The reviewer has misunderstood this sentence: we are not comparing the profiles retrieved from the two fitting windows; we are referring to the statistically significant enhancement at 9.5 km, relative to the NO_2 number density (or even volume mixing ratio) at 12.5 to 18.5 km, also observed by mini-SAOZ.

Can you include 1D retrieved for the fitting window from v3.0 algorithm in Figs. 6 and 7 as well?

As the reader reaches Figs. 6-7, the focus is on the comparison of the new retrieval results (using the 2D and 1D forward models) with those of previously published, somewhat related algorithms, namely ‘fast’ and v3.0. Adding the 1D retrieval using the fitting window from the v3.0 algorithm blurs this focus. Fig. 2 (left panel) has already clearly demonstrated a significant, albeit slight, improvement in precision over a sufficient sample size of profiles.

Paragraph discussing Fig. 3 on P8, L16-26: At 20 km altitude 2D&3D versions have larger values than 1D, is there an explanation?

At 20.5 km, the difference between retrieved NO_2 when using 2D and 1D forward model modes is +8% or 8×10^7 molec/cm^3. Similarly, for 3D versus 1D modes, the difference at 20.5 km is +10% or 9×10^7 molec/cm^3. Both of these differences are approximately equal to the single profile precision. However, what is probably occurring is that the 1D mode forward model, which neglects the diurnal variation,

overestimates the NO_2 at 22.5 km where the NO_2 concentration is larger and this overestimation results in an underestimation at 20.5 km.

We now write in the revised manuscript:

The slightly lower NO_2 number density at 20.5 km retrieved with the 1D forward model (Fig. 3) is probably a minor oscillation due to an overestimation immediately above at 22.5 km where a secondary NO_2 number density peak is observed by OSIRIS.

All versions disagree with balloon above 21 km or so, can you also explain this? I suppose you selected the best plot for the illustration here. How is about other individual profiles?

There is disagreement at most/all altitudes above 21 km for the different OSIRIS products. However, OSIRIS reproduces the height of the minimum in NO_2 observed by SAOZ at 27 km. The lower maximum occurs at 24 km for SAOZ and at 22.5 km for OSIRIS. The largest differences in retrieved NO_2 number density relative to SAOZ are at the highest two altitude levels observed by the balloon sensor. This raises the possibility that the difference is partly driven by the NO_2 assumed in the SAOZ retrieval above balloon float altitude. It is obvious from other coincidences that the assumed profile above balloon float can be an issue, particularly in the tropics (as is the case here), since the NO_2 peak is at a higher altitude. The profiles retrieved with the 2D and 3D mode forward model versions indicate that the height of the OSIRIS largest NO_2 maximum is at 31.5 km, well above the highest altitude for SAOZ NO_2 (29 km).

The reviewer's supposition is not correct: we chose this coincidence based on it being the one with the largest expected error for OSIRIS if the twilight gradient in NO_2 is not included in the OSIRIS forward model. For most of the coincidences, this "diurnal effect" error is small since OSIRIS is not measuring sufficiently across the day-night terminator.
OSIRIS tends to be high by ~10% relative to the balloon data in general near ~28 km. This general bias is not completely understood but it appears to be slightly smaller for the new datasets (9% relative median bias for '2D mode', 11% for '1D mode') relative to existing products (13% for both 'fast' and v3.0). For this case shown in Fig. 3, there is a ~20% positive bias at this altitude and a positive bias of a similar magnitude at 22.5 km. The latter bias has to do with the altitude of the lower peak being shifted between OSIRIS and SAOZ. We believe the bias is due to inadequate vertical sampling by OSIRIS. The vertical sampling of OSIRIS for this limb scan was >2.0 km in the vicinity of this enhancement at 24 km (with the nearest THs being at 22.1 and 24.2 km). If the spectrum at 24.2 km observed by OSIRIS did not capture much of this enhanced NO_2 layer, it would end up appearing as a local maximum in NO_2 number density at 22.5 km.

We feel a comment on the high bias above 21 km for this single profile is not too valuable here given that:

1) this is a single profile, and the general bias near 30 km is already discussed (p10L28).
2) the focus of the paper on the lower stratosphere (p3L14).
3) the symptom of inadequate vertical sampling is already discussed in Sect. 4.
4) this is not the best validation opportunity for the mid-stratosphere since the balloon float altitude is below the NO_2 peak altitude (similar discussion already exists at p11L10-17). Note that the error in assumed NO_2 above float dampens quickly with decreasing altitude such that the balloon data in the lower stratosphere should suffer minimally from this source of error.

I would recommend providing the comparisons of the individual profiles for the whole validation dataset as a supplement.

A spreadsheet containing the individual profile comparisons has been generated and will be included as as a supplement. We now refer to the available supplementary material on p9L28.

P10,L1, Fig.4 Why sample size of 2D is not shown in the figure. Please mention if the sample agrees with that of 1D.

Sample size of 2D is identical to 1D. We thought this statement was made, but the reviewer is correct that it was missing from the original manuscript. We now write in the Fig. 4 caption:

Sample size versus height for 2D mode is identical to that for 1D mode.

P10,L23-25: Can you comment what factors (RTM, forward modelling, chemical modelling, inversion) are impacting how the processing time for the different dimension versions, i.e. is the increased calculation burden in 2D and 3D just due to the chemical modelling?

The processing time increases greatly from 1D mode to 2D mode partly because the high resolution model is used in the latter mode (Zawada et al., 2015). In other words, the forward model (RTM) is the main factor. The doubling of the processing time for a single profile retrieval between 2D and 3D is again driven by the forward modelling, not the chemical modelling or the inversion. In the first paragraph of Sect. 2.2, we now add a sentence on the high resolution model:

"This capability comes at the price of a large increase in computing time (see also Sect. 3)."

P11, L10 & Conclusions P13, L28: the biases clearly exceed 10% (I see by eye up to 13% bias, i.e. by one third more than 10%).

In both sentences identified by the reviewer, a '~' is used. Basically, the reviewer should interpret "~10%" as having one significant digit and thus 14% is not different from ~10%. As noted above, the median bias is 11% for 1D and 9% for 1D and 2D forward models, respectively. We have now added 'median' to both sentences.

It is also evident that 2D product has larger bias than 1D, in particular at altitudes below 20 km. please discuss.

The NO_2 retrieved with the 2D forward model underestimates slightly since it does not account for the diurnal variation along the incoming beam. This slightly increases the insignificant negative bias (which exists whether the 2D or 1D forward model mode is used). The vertical profile of the bias is very similar between 2D and 1D, the difference between them is only a few percent and is not statistically significant enough to be discussed in the paper. The 3D forward model should be best, but it is far too time-consuming to be useful operationally and thus was not considered other than in Fig. 3.

P11, L13-L15 "In contrast to 'fast' and v3.0 products, there are no altitudes in the lower stratosphere (below 24 km) with statistically significant average and median biases for the 1D and 2D products": this statement is incorrect: in Fig. 7 the bias clearly exceeds error at 15.5 – 20.5 km for 2D: : :

The reviewer has missed the '…average *and* median…' in their excerpt.

P12, L11-16,Fig.9: There is no significant differences in the correlation for most of the altitudes between the different products, as a consequence clear conclusion on a better performance of a certain algorithm cannot be derived. Can you comment on the relatively high correlation of the fast product in the lowermost stratosphere? Is it due to the better forward model?

The higher correlation of the fast product relative to the '1D' product is not due to the better forward model since fast also uses Sasktran in '1D mode'. The sample size is too small at the lowest altitude (12 km) for 'fast' to make any strong conclusion about relative performance. This is the only altitude with a large difference in correlation. Nevertheless, the higher correlation for the 'fast' product at 12 km can be partly explained by limiting the correlation calculation for all products only to the coincidences for which 'fast' data is available. When this was done, the other products also had higher correlations at 12 km, since the 'fast' subset fortuitously consisted of better comparisons with the balloon data.
The reviewer's first statement in this comment ignores the other statistics used to quantify "better performance". No change is made to the manuscript since the original manuscript (p10L2) already points out that the main focus of the validation is the 13-31 km altitude range, where the sample size appears to be sufficient.

P14, L11-L14: Although the presented concept of photochemical correction seems promising, from the results shown (in particular on Figs. 6&7 but also on other plots as commented above) it is hard to speak about a clear improvement for the new algorithm.

The improvement is subtle relative to the operational algorithm (v3.0). The main advantage of the new data product relative to the v3.0 product is the improved performance in the 15-17 km range (as already written starting at p13L26). Specifically, there is a significant positive median and mean bias in the v3.0 product at these altitudes. Fig. 8 also shows that there is much more scatter in the v3.0 product than the other products at these altitudes. The correlation coefficient at 15-16 km is also the worst for the v3.0 product (Fig. 9). Relative to 'fast', the improvements of the new products (with and without diurnal NO_2 gradients in the forward model) are more obvious: the 'fast' product was limited in its vertical range at both ends (top and bottom). The reduced scatter of the differences versus the balloon data (Fig. 8) and improved correlation relative to 'fast' is clearly evident in the 20-30 km range. Furthermore, the overall shape of the profile is improved relative to 'fast' as the high bias near 28 km, and the low bias at 18-20 km are both reduced.

Especially at the lowest altitudes the bias (see Fig. 7) is the largest.

Because the number of balloon coincidences is somewhat limited, and very limited at 12 km, it is important that the reviewer considers both the median and mean bias. And because the number of coincidences is a strong function of height at the lowest altitudes, the reviewer should look at the statistical significance of the bias, not just the magnitude of the bias. This can be done by looking at whether the bias profiles differ significantly from zero given the variability of the individual differences (shadings in Figs. 6 and 7). The original manuscript already discusses biases in terms of statistical significance (p11L13-15).

I recommend to extend the study by investigating the behaviour of the agreement for different conditions (e.g. season, location, SZA, different instruments, techniques : : :)

We felt strongly before the validation and feel even more strongly after having done the validation that balloons provide the best opportunity to validate the OSIRIS products in the lower stratosphere, which is our main region of interest in this study (P3L14).
We agree with the recommendation of the reviewer, but unfortunately there are not enough coincidences in any season (N<20). Furthermore, seven coincidences occur in the tropics where the seasonal cycle of stratospheric NO_2 appears to be small. Regarding latitude (or 'location'), only one region would have a sufficient number of coincidences (northern high latitudes), so the latitudinal variation of retrieval performance can hardly be studied in a statistical way either. There are four techniques used: ascent/descent occultation, limb occultation, limb scattering, and in-situ. Since only one technique has a sufficient number of coincidences, namely limb occultation, it is not possible to make any general, statistically-founded conclusions about how the agreement varies between the correlative measurement techniques. There is also an insufficient number of coincidences for any one correlative instrument. Because Odin is in a terminator orbit, the solar zenith angle range of OSIRIS is quite limited as shown in Fig. 2 of Haley et al. (2004). The coincidences are even more limited in SZA (68.6 to 90.5°). In the abstract we now modify one sentence to read: "…a range of solar zenith angles (68.6 to 90.5°)…".

A study comparing the real gradient (from OSIRIS), gradient simulated by the chem. model and the correction on profile would be interesting. Do the gradients simulated by the chem. model follow real gradients in all cases? Certainly also the effect of horizontal model resolution should be investigated which seems to be much too coarse (see above).

The reviewer's suggestion is interesting, but OSIRIS is changing latitude as well as local time during the orbit. At some times of the year, the latitudinal variation can be very strong, e.g. across the polar vortex edge or a Noxon cliff, and the photochemical model would not be expected to capture the dynamical variability that exists over small latitudinal ranges (or orbital segments). The reviewer's final comment would be correct if we were doing a two-dimensional retrieval. However, we are not doing so and apologize again for any confusion due to the sentence at p8L28. The revised manuscript should be clearer in terms of the methods employed.