Dear Referee#1,

We would like to thank your suggestions in order to improve our manuscript, which are fully addressed below. Your comments appear in bold.

Best regards,
Omaira García et al.

Specific comments

1) Title

Since a lot of work has been done also on ECC sonde measurements (consistency of the time-series / comparisons with FTIR / trend estimation), the authors could maybe include this in the title (the importance of using 2 different techniques in the paper is expressed by the authors in the first paragraph of Sect. 6.1). This could give more visibility for the paper to the ECC sonde community (this is a suggestion, the authors can decide).

The ECC sonde measurements are widely used throughout the paper, as the referee comments, for supporting the FTIR's results. Nonetheless, the core technique of this work is the FTIR spectrometry and a complete description of this measurement technique is given in the manuscript (especially section 3). Hence, the title of the paper only refers to the FTIR system. References to ECC sonde data are included in the abstract, showing the importance of these data in the study. We think that it is sufficient.

2) Abstract

- l.9-12: "our theoretical calculations indicate that a very precise knowledge of the instrumental line shape is mandatory for a precise g-b FTIR remote sensing of stratospheric ozone". The authors write at other places that a very precise knowledge of the ILS is mandatory for precise determination of the upper stratosphere ozone layer Sects. 3.4; 3.5; 6.1; conclusions). It is a very interesting point, and I wonder if the authors could demonstrate this in a more precise way. They show that the ILS is a dominant source of error, and they give the error contribution in the case of a precise knowledge of the ILS (cell measurements). Could they give a rough estimation of the error on partial columns due to the ILS in case, no cell measurements being available, the ILS would be included as fitted parameters in the retrieval process, and/or in case the ILS is assumed to be ideal?

The influence of the uncertainties in the Instrumental Line Shape (ILS) can be derived by analysing jointly the time series of the modulation efficiency (Figure 5 of manuscript) and the error budget for the significant contributors to ozone partial column errors (Table 3 of manuscript and revised Table 4, please see answer of section 3, point f).
For example, if the ILS is assumed to be ideal before June 2008 the error associated to the modulation efficiency can reach 5% (see Figure 5 of manuscript). According to our error estimation (Table 3 and revised Table 4), this uncertainty implies an error for setup C of about 1.5% in the 2.37-13 km layer and of about 6.5% for the 28-42 km, while for the total column it reaches 2%. If the ILS is not carefully monitored it is a leading error source.
Regarding to include the ILS as fitted parameters in the retrieval process, there is a large interference between the ILS fit and the fit of stratospheric absorbers. Hence, it is not feasible. Instead the ILS information should be obtained by cell measurements.

- l. 17-20: when one reads the sentence, one expects that the trends derived from ECC dataset are also -0.3%yr-1 and +0.3 %yr-1, which is not the case. The trends from FTIR and ECC agree within their error bars, and the FTIR trends are -0.3%yr-1 and +0.3 %yr-1. I think also that it is worth to

mention in the abstract that these trends (from FTIR) are statistically significant (and/or to give the confidence intervals).

- last sentence: I think the authors should be more careful when they link their observed trends to the increased circulation in response to climate change. Indeed what they observe, especially in the lower stratosphere is linked to change in the Brewer-Dobson circulation, but within a such short term period, this can also be due to the inter-annual variability of this circulation (WMO 2011) rather than long-term change due to climate variability (or it could be a combination of both). Also, in the upper stratosphere, if the effect of climate change is indeed found to increase ozone in that layer, it is certainly combined with the effect of the decrease of EESCs (Fig. 3-21, WMO 2011). In Hegglin and Sheperd (2009), the authors avoid the effect of EESCs by showing the differences between two periods where the EESCs are supposed to be equal. But the present paper deals with data that are in the decreasing part of the EESCs time-series, so it is expected to see an impact in the upper stratosphere.

Following the referee's suggestion, the abstract has been slightly modified, including a clarification about the trends from the FTIR and ECC datasets (see text below). In addition, the explication of the observed trends has been reduced here, including more details in the introduction (section 1) and the discussion of results (section 6). Please note that the altitudes that define the layers have been re-defined and thus the estimated ozone trend have slightly changed (please see answers of section 3).

"The linear trends estimated from the FTIR and the ECC datasets agree within their error bars. For the FTIR, we observe a significant negative trend in the upper troposphere/lower stratosphere of about -0.2 $\%yr^{-1}$ and a significant positive trend in the middle and upper stratosphere of about +0.3 $\%yr^{-1}$ and +0.4$\%yr^{-1}$ respectively. Admittedly, a 12-year time series is too short for reliable trend studies, however, it is worthwhile mentioning that such subtropical ozone profile trends are predicted by climate models at the northern subtropical latitudes."

3) Introduction

- First paragraph: as for the abstract, the discussion and references on ozone expected trends should mention the effect of declining EESCs in the upper stratosphere.

- The introduction is incomplete for the good understanding of the context of the present paper: previous long-term evaluations of ozone partial columns time-series have been made at several FTIR European stations (Vigouroux et al., 2008; updated in WMO 2011). The Izaña station is one of these stations. The present paper deals with one additional year of data compared to WMO 2011. This should be mentioned, together with the clear statement of one of the scope of the paper: examine if the "NDACC" retrieval setup can be improved, especially for the measurement of the long-term evolution of ozone.

The introduction has been slightly modified to include the referee's suggestions, as follows (the text modified or included appears in italic):

"In the coming decades some kind of ozone recovery is expected, however, it is difficult to predict how, when, and to what extent it will occur (Weatherhead and Andersen, 2006). Currently it is discussed how climate change will interact with ozone recovery. *The multiple interactions between the components of the chemistry-climate system complicate a clean attribution of changes in ozone to changes in ODSs (ozone-depleting substances) and other factors such as the Brewer-Dobson circulation, anthropogenic emissions of greenhouse gases, stratospheric temperatures, etc (WMO, 2011).* For example, climate models predict an accelerated stratospheric circulation, leading to changes in the spatial distribution of stratospheric ozone and an increased stratosphere-to-troposphere ozone flux (Hegglin and Shepherd, 2009, and references therein). *Nonetheless, the combined effect of the decrease of anthropogenic halogen abundances in the upper stratosphere from the mid-1990s has also to be considered (WMO, 2011).* In order to verify or decline the different climate model simulations consistent long-term observations of the

vertical distribution of ozone are required. Since the expected signals are rather small (e.g. expected trends from −3% to +1% per decade between 1960 and 2100, Hegglin and Shepherd, 2009; Li et al., 2009), only high precision observational datasets are useful.

Within the NDACC (Network for the Detection of Atmospheric Composition Change, e.g. Kurylo and Zander, 2000) high resolution solar absorption infrared spectra have been measured by ground-based FTIR (Fourier Transform InfraRed) spectrometers for up to two decades at globally distributed sites. It has been shown that these measurements can provide very high quality ozone total amounts (Schneider and Hase, 2008; Schneider et al., 2008a; Viatte et al., 2011) and profiles (Schneider et al., 2008b). *Due to its long-term characteristic and its high precision the FTIR data are very interesting for trend studies. Vigouroux et al. (2008, updated in WMO, 2011) estimated ozone trends at several European NDACC FTIR sites. In this work we examine in detail the FTIR error sources and discuss how they can affect the estimated ozone trends. We present three different FTIR ozone profile retrieval setups, including the setup applied by Vigouroux et al. (2008), and discuss its reliability for providing correct ozone trend estimates.*

The study is performed for the ozone super-site Izaña Observatory, where since 1999 FTIR measurements have been performed in coincidence to several other high quality atmospheric ozone measurement techniques (e.g. Brewer spectrometer, Electro Chemical Cell, ECC, sondes, photometric in-situ surface). The Izaña Observatory and its Ozone Program is described in Sect. 2. In *Sect. 3 we present the three different FTIR retrieval setups and perform detailed theoretical error estimations. In Sects. 4 we briefly discuss the quality of the ECC sonde data and in Sect. 5 we present a day-to-day comparison between the three different FTIR datasets and the ECC dataset*. In Sect. 6 we present the ozone seasonality and the trends obtained at different altitudes from the different FTIR datasets and discuss their consistency to the values obtained for the ECC dataset. Finally, the main results are summarized in Sect. 7."

4) Section 2
No comment.

5) Section 3
5.1) General comments on Sects. 3.1; 3.2; 3.3 to clarify the conclusions about the different retrievals setups:

In order to address a correct comparison between the three retrieval strategies we decided re-defined the different layers, such that all layers have a degree of freedom for signal (dof) larger than one. Hence, we guarantee that each layer is sufficiently well detected by the FTIR system. All results of the paper have been re-calculated considering the new layers. For example, Table 1 lists the statistics of the dof time series of the retrieved ozone obtained from the IFS 120/5HR for all setups, considering the new layers.

Layer [km]	Retrieval Setup		
	A	B	C
	M, σ	M, σ	M, σ
2.37-13	1.01, 0.06	1.09, 0.05	1.31, 0.08
12-23	1.10, 0.06	1.23, 0.08	1.49, 0.11
22-29	1.01, 0.06	1.06, 0.05	1.02, 0.05
28-42	1.25, 0.06	1.28, 0.06	1.18, 0.06
2.37-120	3.84, 0.23	4.10, 0.14	4.20, 0.17

Table 1. Mean (M) and standard deviation (σ) of the dof time series of the retrieved ozone obtained from the spectrometer 120/5 HR for all setups. These values are shown for each layer (2.37-13 km, 12-23 km, 22-29 km, 28-42 km) and for the total column (2.37-120 km). The standard error of the mean (SEM) is lower than 0.01 for all setups and layers (not shown).

a) One information is missing before making conclusions based on Tables 2 and 4: how did the authors choose their Tikhonov constraint? By tuning the regularization strength, one could reach with setup A and B (DOFS=3.84 and 4.10 in Table 2), the same DOFS than setup C (4.20). So what was the criterion to choose the regularization strength (minimizing the total error?)? Since the smoothing error is the dominant error, it is important to explain how the authors obtained the DOFS for the Tikhonov setups A and B, especially when comparing with setup C.

In principle, the constraint for setups A and B is a slope constraint TP1: we constraint the slope of vertical profile and the absolute value for the uppermost atmospheric model altitude. The strength of the constraint is determined by starting with a weak constraint and then increasing it until observing a significant increase in the residual of the spectral fit (L-curve criterion). This strategy is different from setup C, where the strength of the constraint is determined by the a priori covariance obtained from ECC sonde dataset.

b) Is the Tikhonov constraint the same for setup A and B? In case the answer is yes, it is interesting to note that the DOFS is increasing at all layers when the temperature is retrieved, and I would be curious to know if the authors observe the same increase when they use OEM: could they give this information with one sentence (DOFS setup C compared to DOFS setup C without temperature retrieval)? The test could be done on a small set of representative spectra – no need to run the entire time-series. Also the errors are decreasing at all layers by using the temperature retrievals and I would like to know if this is the case also for the OEM retrievals (probably the answer will be yes). Then the authors can indeed recommend strongly to the FTIR community the use of the temperature retrievals to obtain more precise ozone partial columns.

Including a temperature retrieval improves the retrieval quality (i.e., reduces the residues of the fitted spectra) and, thus, an increase of dofs. Hence, when comparing the setup C with and without simultaneously fitting temperature we generally observe an increase of dofs (see Table 2).

Layer [km]	Retrieval Setup	
	C without temperature	C with temperature
	M, σ	M, σ
2.37-13	1.20, 0.08	1.29, 0.09
12-23	1.42, 0.11	1.49, 0.11
22-29	1.01, 0.05	1.02, 0.06
28-42	1.23, 0.07	1.17, 0.08
2.37-120	4.10, 0.18	4.18, 0.20

Table 2. As Table 1, but for the dataset used in the Brewer –FTIR comparison (2005-2007, N=475).

Regarding theoretical error estimation, the inclusion of a simultaneous temperature fit also reduces significantly the random error associated with the temperature for all layers, as summarises the following table (Table 3). Note that for the higher layers (22-29 km and 28-42 km) and for setups without fitting temperature, this error source accounts for most of error on the ozone partial columns.

Layer [km]	Retrieval Setup			
	A	B	C (without temperature)	C (with temperature)
	Temp, TPE	Tem, TPE	Tem, TPE	Tem, TPE
2.37-13	0.6, 1.0	0.5, 0.9	1.3, 1.6	0.5, 1.0
12-23	1.3, 1.4	0.3, 0.7	1.1,1.2	0.3, 0.7
22-29	2.7, 2.7	0.3, 0.9	2.7, 2.8	0.3, 0.9
28-42	4.2, 4.3	0.7, 2.0	4.3, 4.4	0.6, 1.6
2.37-120	2.3, 2.3	0.4, 0.7	2.2, 2.2	0.3, 0.6

Table 3. Estimated random errors relative to actual ozone partial columns [%] for typical measurement conditions for the Izaña spectrometer 120/5HR for all setups and for the different layers. TPE [%]: Total Parameter Error due to all input parameters and measurement noise and Tem [%]: Error due to temperature.

A comment about the influence of a simultaneous temperature fit on the dof and on the theoretical error estimation for setup C has been included in the revised manuscript in section 3.2 and 3.3 respectively.

c) Two changes have been made between setup B and C: the use of a realistic Sa matrix instead of Tikhonov regularization and the use of an inter-species constraint between the different ozone isotopologues. Then, I would like to know which one of these two changes made the largest impact on the differences observed in DOFS and TRE. Did the authors make an intermediate setup between B and C?

We would like to remark that in the real atmosphere the different ozone isotopologues are strongly correlated. Thus, an Sa matrix is only realistic if this a priori knowledge is considered. Our retrieval setups show the two possible situations:
1. Setup A and Setup B: a priori information is not available and the constraint is determined ad-hoc by the L-curve criterion.
2. Setup C: it includes all the a priori information available and thus it can be considered as the most realistic constraint.

d) It is clear from Table 4 that the setup C gives a better precision than setup B in the two lower layers, especially in the troposphere. However, the precision is worse in the upper layer, which is also a layer of scientific interest for the study of ozone recovery. So, I would advise the authors to be more nuanced when they write that setup C is the optimal one. Also, what about this Sa matrix at the altitudes above the ECC sonde measurements? Could the authors find another climatology for these altitudes (from satellite measurement)? This loss of precision in the upper layer is due only to the use of OEM or could it be improved by using also an appropriate climatology above the altitudes of the sondes?

Above 30 km setup C has a larger smoothing error since it is stronger constrained than setup A and B at these altitude layers. This constraint is obtained from the variability as observed in ozone sondes (above the ozone sonde we assume a similar variability as at 30 km). It is a realistic constraint. The constraint obtained ad-hoc for setup A and B is much looser and might mean an over-interpretation of the spectra for these altitudes. This over-interpretation leads to larger parameter errors. This becomes better visible in the new layering (please refer to the Table 4 in the comment (f) of section 5.1).

In agreement with the better precision in the lower layers, "setup C" is better when comparing with ECC sondes (Fig. 9). Also, the Brewer comparisons are improved with setup C (even if the temperature retrievals have a larger impact on the precision of total columns). So, to strengthen their conclusion on setup C, I would add in Table 4, the errors for the total columns.

Table 4 have been modified in the revised manuscript by including the significant contributions to the ozone partial column random errors (instrumental line shape (ILS), temperature and measurement noise) and the errors for the ozone total columns. Please refer to the comment (f) of section 5.1.

The discussions about the better (or worse) precision of one setup compared to another are valid in the case of a Tikhonov constraint chosen in order to minimize the total random error. Otherwise one could ask himself if it would not have been possible to obtain similar errors values for setup A and B than for setup C by tuning the constraint. (so same question than comment a): how was chosen the Tikhonov constraint ?)

Please refer to the answer for the comment (a) of section 5.1.

e) How SE (Table 4) is calculated? Using the same Sa matrix for each setup? It should be explained in the text, especially since these values of SE are used to determinate the best retrieval setup.

The smoothing error (SE) is calculated following the formulism given by Rodgers [2000] as $(\mathbf{I}-\mathbf{A})\ \mathbf{S}_a(\mathbf{I}-\mathbf{A})^T$. Here, **I** is a unity matrix, **A** is the averaging kernel, and $\mathbf{S}_a$ the assumed a priori covariance of atmospheric ozone. We use a $\mathbf{S}_a$ matrix that is obtained from the WACCM climatology (Whole Atmosphere Community Climate Model), which is used to calculate the SE for the three setups.

This explication has been included in the revised manuscript (section 3.3).

f) Fig. 4 – Table 4: Maybe the authors should mention that the errors profiles plotted in Fig.4 are the diagonal elements of the error matrices of the different contributions, and that these error matrices have off-diagonal elements. It would be very helpful for the discussions to include in Table 4 the significant contributions to the partial columns errors (temperature, noise, ILS).
Maybe giving a Table (or additional column in Table 4) for the contributions of the systematic errors could help also for the discussions on the possible effect of ILS on the trends (if the authors try to go deeper in the discussion about the ILS, as suggested in comment 2).

The Table 4 of the revised manuscript has been modified by including the significant contributions to the ozone partial column random errors: instrumental line shape (ILS), temperature, measurement noise. Likewise, the errors for the ozone total columns have been also included.

Layer [km]	Retrieval Setup		
	A	B	C
	TPE (ILS, Tem, Noi), SE, **TE**	TPE (ILS, Tem, Noi), SE, **TE**	TPE (ILS, Tem, Noi), SE, **TE**
2.37-13	1.0 (0.4, 0.6, 0.8), 9.5, **9.5**	0.9 (0.4, 0.5, 0.6), 9.3, **9.4**	1.0 (0.3, 0.5, 0.8), 8.7, **8.8**
12-23	1.4 (0.4, 1.3, 0.5), 2.6, **3.0**	0.7 (0.5, 0.3, 0.4), 2.4, **2.5**	0.7 (0.5, 0.3, 0.4), 2.2, **2.3**
22-29	2.7 (<0.1, 2.7, 0.4), 3.2, **4.2**	0.9 (0.6, 0.3, 0.5), 3.3, **3.4**	0.9 (0.5, 0.3, 0.5), 2.9, **3.1**
28-42	4.3 (0.7, 4.2, 0.5), 3.5, **5.6**	2.0 (1.6, 0.7, 0.7), 2.9, **3.5**	1.6 (1.3, 0.6, 0.6), 3.5, **3.8**
2.37-120	2.3 (0.1, 2.3, 0.1), 0.3, **2.3**	0.7 (0.4, 0.4, 0.2), 0.1, **0.7**	0.6 (0.4, 0.3, 0.2), 0.8, **1.0**

Table 4. Estimated random errors relative to actual ozone partial columns [%] for typical measurement conditions for the Izaña spectrometer 120/5HR for all setups and for the different layers. TPE [%]: Total Parameter Error due to all input parameters and measurement noise; ILS [%]: Error due to instrumental line shape; Tem [%]: Error due to temperature; Noi [%]: Error due to measurement noise; SE [%]: Smoothing Error; TE [%, in bold]: Total Random Error.

In addition, a clarification of the vertical error profiles has been included in section 3.3 of the revised manuscript (please see text below):

"The propagation of uncertainty sources for a typical measurement of the spectrometer 120/5HR, and applying the different retrieval setups, is displayed in Fig. 4. This figure shows the error profiles as the root-square of the diagonal elements of the error covariance matrix for the different error sources considered (see Table 3)."

5.2) Section 3.1: context of the work

After a complete description of the context in the introduction part, the authors should explain in Sect. 3.1, what was the strategy used at Izaña in Vigouroux et al. 2008 and WMO 2011. And maybe (if not already done in the introduction), they could say a few words on which station uses which strategy in this previous work on ozone trends.

A brief description about the strategy used at Izaña in Vigouroux et al. (2008) and WMO (2011) has been included in the revised manuscript (the text modified or included appears in italic):

"Setup A can be considered as the "NDACC" approach except for the logarithmic instead of the linear scale retrieval of ozone. *This strategy has been used for the ozone trend estimations at Izaña and Kiruna as presented in Vigoroux et al., 2008 (updated in WMO, 2011).* "

5.3) Section 3.2:

- p.3437, l.15: The avks are usually described as the rows of the matrix A (Rodgers, 2000) not the columns.

When depicting averaging kernels some authors depict the rows, other the columns, and others both of them. We decided to show the column kernels. They document how an atmospheric perturbation is smoothed out by the remote sensing system. The row kernels inform about the atmospheric altitudes that affect the retrieved FTIR data.

- p.3438, l.8: "When interpreting the FTIR time-series it is important to consider the time evolution of avks": I did not find if (where) the authors took this into account in their trend study.

It is important to keep in mind that trends in the response function of the remote sensing system (avks or dofs) can influence the trend: decreasing dofs might underestimate gradually increasing differences between the a priori O3 used as constraint and the real O3, i.e. it underestimates trends. Furthermore, there might be a bias between the climatologic O3 data (the a priori) and the FTIR O3 data due to systematic error sources like spectroscopic line parameters. In this case the magnitude of the bias will decrease with decreasing dofs, thereby leading to a trend even though real atmospheric O3 remains stable. The variability and the drifts in the dof can be observed in Fig. 3. For instance, there is clear drift between 1999 and 2004 meaning that trends estimated for this time period have to be treated with care.
Furthermore, we followed the referee's suggestions and analyzed the differences in the trends obtained from smoothed and unsmoothed ECC sonde time series and briefly mention this in the discussion of Sect. 5.

5.4) Section 3.3:

- p.3438, l.22: "The uncertainties are split into statistical and systematic contributions, 80% and 20% respectively: : :": How these numbers are obtained ? (same question for Fig.4: how is made the distinction between random and systematic part of the errors?) The authors should explain more or give a reference.

These are our assumptions. They are based on our experiences and we think that they are of a realistic order of magnitude. It is important to document the uncertainty assumptions for which the errors are calculated. The reader can easily calculate the errors that would result from higher uncertainties as the ones that we assume. Since the error estimation assumes linearity the reader has just to scale the errors accordingly.

5.5) Sections 3.4 and 3.5:

a) Maybe (only suggestion) change the titles into:
- 3.4 Long-term consistency of FTIR measurements
3.4.1 ILS
3.4.2 Comparison between: : :

OR
- 3.4 Long-term consistency of the ILS
- 3.5 Comparison between: : :

The titles of section 3.4 and 3.5 have been changed into:
3.4.Long-term consistency of the ILS.
3.5. Comparison between IFS 120M and IFS 120/5HR.

b) When a significant bias is observed between 120M and 125HR ozone measurements (for the 31-42 km), is it taken into account for the trends calculation? Are the columns corrected? This should maybe be said / justified.

The biases between the IFS 120M and 120/5HR ozone partial columns were obtained with side-by-side measurements during only two months at 2005 and, therefore, we can not guarantee that these biases are constant over time. So, in order to avoid possible artefacts in estimating the FTIR's trends the ozone partial column time series from IFS 120M (1999-2004) was not corrected. It is important to mention here that the whole FTIR ozone partial column time series is consistent to ECC sonde partial column time series: no significant biases were observed between two datasets (please see section 5 of the manuscript).

6) Section 4
- p. 3442, l.10 and l.15: Schneider 2008b instead of Schneider, 2008a

The reference has been corrected in the revised manuscript.

7) Section 5
- p.3444, l14: as shown "in" Fig.9

The phrase has been corrected by including the word "in" in the revised manuscript.

- The authors have chosen not to smooth the ECC sonde profiles with the FTIR averaging kernels. However, I think it would help their discussion to do so, for example:

a) p.3444, l.25: "The smoothing error might explain a large part of the discrepancy between FTIR and ECC: : :" It would be interesting to see if this discrepancy remains when the ECC sonde profiles are smoothed with the FTIR avks before they are integrated into partial columns. Also in that case, one would expect (from Table 4) that the comparisons would improve between setup A and B, especially in 22-29 km layer (since TPE is decreasing), but not anymore between setup B and C, where mainly the smoothing error is improved.

b) p.3448, l.17-21: the authors explain the difference in the FTIR and ECC sonde annual cycle by the smoothing error. This could be proven by applying the avks on the ECC sonde profiles. (idem p.3445, l.18)

c) If the authors could compare the trends of ECC sonde with and without the smoothing, this could give an indication on the effect of the smoothing error on the FTIR trends (their issue about the trend in the DOFS which could lead to an artificial trend – Sect. Conclusions)

Not smoothing the ECC sonde data with the FTIR avks guarantees that the ECC time series and, thus, the ECC annual cycles and trends observed are completely independent of the FTIR time series. Hence, the ECC sonde trends are not influenced by possible trends of the FTIR avks. Nonetheless, as referee

suggests, performing the intercomparison between the smoothed ECC and FTIR data would allow for documenting the influence of the smoothing error. Naturally, a comparison to smoothed ECC sondes will not be affected by the smoothing error (then ECC and FTIR have the same smoothing error). We made this study and in Section 5 we briefly mention to what extent this increases the agreement between FTIR and ECC profiles. This effect is summarized in the following Table.

Layer [km]		Retrieval Setup		
		A	B	C
		M±SEM, σ	M±SEM, σ	M±SEM, σ
2.37-13	ECC not smoothed	-4.8±0.6, 9.7	-5.9±0.6, 9.2	-6.7±0.5, 8.5
	ECC smoothed	-2.8±0.4, 7.2	-3.6±0.4, 7.2	-4.1±0.5, 7.3
12-23	ECC not smoothed	0.4±0.4, 6.3	0.8±0.4, 6.0	1.0±0.4, 5.8
	ECC smoothed	3.3±0.3, 4.8	3.3±0.3, 5.0	3.7±0.3, 5.0
22-29	ECC not smoothed	7.1±0.2, 3.6	6.9±0.2, 3.4	7.2±0.2, 3.5
	ECC smoothed	7.1±0.1, 2.3	7.0±0.1, 2.3	8.2±0.1, 2.2

Table 5. Statistics of relative differences [%] between the coincident measurements from the ECC sonde and FTIR data for all setups and the different layers. ECC sonde data with and without smoothing by FTIR avks are shown. M±SEM: Mean and standard error of the mean, σ: Standard deviation.

Regarding the effect of the smoothing error on the FTIR trends, we have performed the test proposed by the referee. In order to evaluate possible artificial trend in the FTIR's dofs time series, we have analyzed the time series of the differences between smoothed and unsmoothed ECC sonde data. For all setups we have observed that there are no significant trends in the 2.37-13 km and 12-23 km layers, but that there is a significant trend (at 95% of confidence) for the 22-29 km layer. For example, for setup C, the trend is of 0.12±0.10 DU/yr (i.e., 0.11±0.09 %/yr) in this layer (see figure 1).

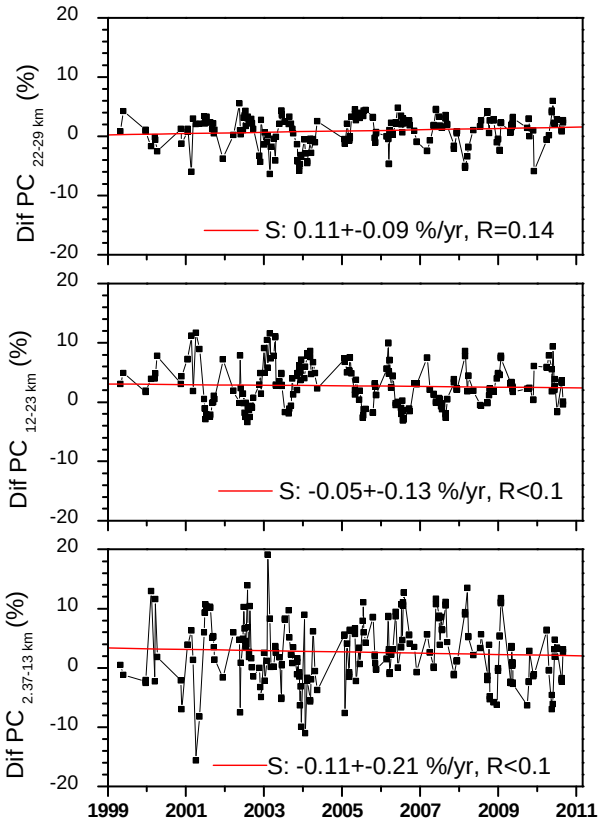

Figure 1. Time series of the relative differences between the smoothed and unsmoothed ECC sonde data with avks from setups C. In the legend the slope (S) and the correlation coefficient (R) of the best fit is shown.

Therefore, in particular in the middle/upper stratosphere the FTIR ozone trends are affected by our avk time series. However, it is also important to mention here the reduced number of coincident measurements between ECC sonde and FTIR dataset, only 263 cases during 12 years, which are mainly recorded from April to October. The strategy of not smoothing the ECC sonde time series and, thus, of evaluating the trends from FTIR and ECC with the whole time series takes advantage of a larger number of measurements (1887 for FTIR and 515 for ECC sonde data). Thus, the results obtained are more representative

These facts consolidate our applied strategy not to smooth the ECC sondes. Otherwise both instruments might show similar trends, due to the fact that there is a trend in the avks. It also highlights the importance of super-sites, like the Izaña Observatory, that concentrates numerous independent measurement techniques, allowing for a comprehensive intercomparison of techniques.

These results will be mentioned in Section 5.

8) Section 6

a) Context of the work: the authors should add one sentence to compare the obtained trends between the current paper and WMO 2011 (and to explain the differences).

Section 6 we already include a comparison between the trend obtained in the current paper and several theoretical and experimental works, such as Li et al. (2009), Steinbrecht et al. (2009), Vigoroux et al., (2008), WMO (2011). We think that a more in detailed discussion about the obtained trend and other studies is out of the scope of our technical paper.

b) One could expect that improving the precision on the FTIR ozone partial columns (from setup A to setup C) would improve the precision on the obtained trends. However it seems from Fig. 12 that this is not the case (errors bars are similar –even slightly larger for setup C and the 11-21 km layer). This could be due to the fact that the "noise" due to atmospheric processes (see Sect. 6.1) is more important than the noise due to the precision of the ozone retrievals. It is worth to mention this result of the retrievals setups comparison study: the better precision achieve with at least setup B (for setup C – it depends on the answers of the authors to the Sect. 3 comments) has no (or few – not clear with only a figure, and not given numbers) impact on the confidence interval on the trends, in the currently used model.

The day-to-day ozone variability is much larger than the random uncertainty of the ozone data. Hence, the confidence interval for our trend estimation as obtained from the bootstrap method is of course mainly determined by this day-to-day variability.
Nonetheless, high quality measurements, like those obtained from setup C, are very important to minimize possible artificial trends caused by drifts in the error sources. For example, a simultaneous temperature fit minimizes the artificial trend that might be caused by a drift in the temperature uncertainty (e.g., it might be -1°C in 2000 and gradually improve to ~0°C in 2010). Another example is the ILS uncertainty. There might be a drift in the ILS (see Fig. 5 of manuscript). If this possible drift is not adequately considered (e.g., by using the ILS results as obtained from the cell measurements for the retrieval of ozone) an artificial trend will be the consequence.
It is also important to correctly constrain the solution. An appropriate constraint to correctly interpret the variability as seen in the measured spectra. A wrong interpretation might lead to wrong trends. (e.g., if the lower stratosphere is over-constrained and the upper stratosphere is under-constrained, a lower real stratospheric trend will be – to some extent – interpreted by the FTIR system as an upper stratospheric trend).

Precise data are important for trend studies. We will expand the error discussion in Sect. 3.3 accordingly.

c) In the troposphere, the values of the trends with the different setups agree well within the error bars. However the conclusion is different: significantly positive for setup A; non significantly for setup B and C. What is surprising is that the larger impact on the trends comparisons occurs in a layer where the theoretical calculations of the random errors (Table 4) show the less impact: the temperature retrievals (from setup A to setup B) only improve the TRE by about 3%. Could the authors explain more what is happening at this layer when the temperature retrievals are performed? I guess the retrieved temperatures are more different than the a priori ones (from diurnal radiosondes) in that layer? Are the retrieved temperature realistic (i.e compatible with the radiosondes error bars) in that layer? The ECC sondes give a value closer to the setup A, but the conclusion (non significant trend) is the same as setup B and C. Would it be possible to obtain the trend from the surface data (since at the altitude of Izaña they are representative of the free troposphere)?

Many factors might affect the ozone trends in the troposphere (temperature error, the simultaneous temperature retrieval,…). In order to investigate the quality of the retrieved temperature profiles, we have analyzed the differences between the surface temperature (in-situ temperature measurements) and the retrieved temperature and a priori temperature (from radiosondes and NCEP) at the Izaña's altitude. We observed a very good agreement between these data: a mean difference of about 1.9K and a scatter of ±2.9K (±1σ) between the surface temperature and the retrieved values and of about 2.9K (±2.6K) between the surface temperature and the a priori values (for setup C). Note that these differences are of order of magnitude of the assumed uncertainty for the temperature profile below 50 km (see Table 3 of the manuscript), but they do not explain the differences in the trends observed (Table 4). Nonetheless, we think that these differences should not be over-interpreted. Please be aware that the uncertainty bars for the trends as estimated for this first layer are very large. It is a good idea to look on the night-time surface in-situ data, which for Izana are representative for the free troposphere and they can be compared to the FTIR data (Sepúlveda et al., 2012). For this data we observe a significant small negative trend (-0.12±0.07 %/yr), which better agrees with the setups that simultaneous fit the temperature profile.

d) We see from Fig. 12 that the error bars on the trends obtained by the ECC sonde measurements are larger than the FTIR ones (especially for the 11-21 km layer). This is also an interesting result. Is it due to lower precision (5-10% for profiles as given in the paper) or to a different (lower frequent) sampling of the time-series (or combination of both)?

The ozone amounts around the tropopause are highly variable, while in the troposphere and middle stratosphere, the variabilities are smaller. This high variability is especially well observed by the ECC sonde since this technique provides vertically highly resolved data. Furthermore, there are less ECC observations than FTIR observations (there is only one ECC sonde measurement per week). The high variability and the sparser sampling make the trend estimations more uncertain.

- p.3445, l.24: techniques (not technics)

This typographic mistake has been corrected in the revised manuscript.

- p.3445, l.25: Maybe (suggestion), the authors could be less assertive because some papers have been published on multi-regression models applied to short time-series (ex: Bodeker et al., JGR, 1998).

This affirmation has been modified in the revised manuscript (see text below):

"The ozone trends are estimated by using a bootstrap re-sampling method (Gardiner et al., 2008), which models the total variation in ozone by a function F(t) and allows for separating the annual cycles from possible long-term trends".

- p.3446, l.15: ": : :the bootstrap method, which assumes that the residuals are Gaussian: : :". I think this is not correct (Gardiner et al., 2008, p.6722, "This method allows the uncertainty associated with any of the model parameters to be evaluated without making any assumptions about the statistical distribution of the residuals").

We agree with the referee since the bootstrap method proposed by Gardiner et al. (2008) does not assume any statistical distribution of the residuals. With this sentence we wanted to say that for our study we assume that the error has a Gaussian distribution, and not that the bootstrap method is only valid for Gaussian distributions This issue has been clarified in the revised manuscript (see text below):

"The significance of linear trends is estimated by assuming that the residuals are Gaussian and uniform over the whole analyzed time period."

- p.3447, l.13-16: For the quality – in general - of FTIR ozone retrievals in the upper stratosphere, the authors could maybe refer to Vigouroux et al. 2008: FTIR measurements at Jungfraujoch show very good agreement with Lidar measurements at Hohenpeissenberg.

This reference has been included in the revised manuscript in Sect. 5, where the empirical validation is performed. Therefore, we modified this paragraph (the text included or modified appears in italic):

"For the upper stratosphere the quality of the FTIR ozone time series has not been empirically validated in this work by *day-to-day* inter-comparisons due to the lack of respective ECC data. *Nonetheless, previous works show very good agreement between the ozone measurements obtained from FTIR and other measurement techniques in this layer, such as ground-based LIDAR and millimeter-wave radiometer (Kopp et al., 2002; Vigouroux et al., 2008).* "

- p. 3448, l.8: troposphere (not tropopause)

This typographic mistake has been corrected in the revised manuscript.

- p.3448,l.16-21: see comment 7b)

9) Section Conclusions

- p.3449, l. 15: 1999 (not 19990)

This typographic mistake has been corrected in the revised manuscript.

- p.3449, l.25 – p.3450, l.2: see comment 7c). The effect of a trend in DOFS could also be tested by artificially decrease the DOFS obtained by the 125HR to the values obtained with the 120M, by tuning the regularization constraint. It would be interesting to know the influence on the ozone partial columns of such a "jump" in DOFS. Could this be tested?

Trends in the FTIR avks are assessed by estimated trends in the DOFS for the different altitude layers. Furthermore, we compared the trends of ECC sonde with and without the smoothing, as the referee suggested in the comments of section 5 (please refer to the answer of these comments).

- Since a large part of the paper is about the comparisons between the different setups, the authors should give their conclusions about this part (precision on the data themselves and implication for the trends).

The conclusion has been modified as follows (the text modified or included appears in italic). Now the implication of the results for trend studies is discussed:

"*In this paper we document the quality of the ozone profiles obtained from ground-based FTIR systems and discuss its application for long-term studies. We investigate three different retrieval setups: (A) an ad-hoc constraint for ozone and no temperature profile retrieval, (B) an ad-hoc constraint for ozone and a simultaneous temperature profile retrieval, and (C) an ozone constraint based on an ozone climatology (optimal estimation retrieval) and a simultaneous temperature profile retrieval.*

Our theoretical error assessment reveals that the measurement noise and the uncertainties in the ILS and the applied temperature profile (for setup A) are the leading error sources. In particular the retrieved middle/upper stratospheric ozone amounts are strongly affected by ILS and temperature uncertainties. We reveal that the temperature error can be significantly reduced by performing simultaneous temperature profile retrieval. The ad-hoc constraint retrievals offer more DOFS in middle/upper stratosphere than the optimal estimation retrieval. At lower altitudes it is vice versa. Consequently the ad-hoc constraint retrievals might over-interpret ozone variability at higher altitudes.

For an empirical quality assessment we use a coincident ECC sonde ozone profile dataset as reference, whose quality, in turn, has been checked, independently from the FTIR data, by a comparison to Brewer total column measurements. During the 12 year period of 1999–2010, the agreement between the vertical ozone distribution obtained by the FTIR and the ECC sondes is very satisfactory. We show empirically that the FTIR system is well able to capture the day-to-day ozone variability in the troposphere, tropopause region, and middle stratosphere. Furthermore, both techniques reveal very similar annual seasonality. *For the ozone retrieval setup that applies a constraint based on an ozone climatology and includes a simultaneous temperature profile retrieval we observe a slightly better agreement than for the other setups. These observations confirm* our *theoretical quality assessment.*

We estimate the trends for the 1999-2010 time period for the ECC and FTIR datasets. In the middle stratosphere we observe a significant positive trend (95% confidence interval) in both datasets and all FTIR retrieval setups (the FTIR also reveals a significant positive trend above 30km, where there are no ECC data available). In the upper troposphere/lower stratosphere region the FTIR observes a significant negative trend (95% confidence interval), which cannot be confirmed by the ECC dataset. At these altitudes ozone amounts are very variable and the trend estimates are rather uncertain. This is especially true for the ECC trend estimates, since there is only one ECC observation per week (compared to several FTIR observations per week). In the troposphere we observe no significant trend neither in the ECC nor in the FTIR datasets.

A main reason for this satisfactory agreement is the fact that we take a lot of care in documenting the ILS (see Fig. 5), thereby avoiding artificial trends due to drifts in the ILS. A regular ILS monitoring, applying low pressure gas cell measurements, is very important for FTIR trend studies. Furthermore, we think that a simultaneous temperature retrieval is important, since it can significantly reduce the risk of artificial trends caused by possible drifts in the temperature uncertainty, thereby theoretically increases the reliability of the FTIR trends. In our study we observe that the temperature retrieval modifies the estimated trends but that the respective modifications remain with the trends' uncertainties. Using a realistic constraint instead of an ad-hoc constraint does not significantly affect the observed trends. The realistic constraint is important for reproducing the large day-to-day variability (see comparisons in Sect. 5), but it does not significantly affect the estimated trends. Finally, one should consider the temporal evolution of the dofs when using remote sensing data for trend studies. For example, if there is a bias in the remote sensing data, this bias will very likely decrease with decreasing dofs, thereby giving rise of an artificial trend.

In summary we think that correctly estimating the small expected ozone trends is a difficult task for any measurement technique. In this context super-sites like the Izaña Observatory, that concentrate numerous measurement techniques, are important. They allow for intercomparing the techniques, thereby documenting the long-term consistency of the profile datasets ad their suitability for estimating ozone trends."

10) Section References
- Barret et al: De Mazière, M (not Maziére, D. M)
- Lazante et al.: analysis (not anayliss)
- Redondas et al.: sensitivity (not sensitiviy)

Theses references have been corrected in the revised manuscript.

11) Tables and Figures

- legend of Fig. 6: add that these plots are for setup C.

The caption of Fig. 6 has been modified following the referee's suggestion.

References

Gardiner, T., Forbes, A., Mazi´ere, M. D., Vigouroux, C., Mahieu, E., Demoulin, P., Velazco, V., Notholt, J., Blumestock, T., Hase, F., Kramer, I., Sussmann, R., Stremme,W., Mellqvist, J., Strandber, A., Ellingsen, K., and Gauss, M.: Trend analysis of greenhouse gases over Europe measured by a network of ground-based remote FTIR instruments, Atmos. Chem. Phys., 8, 6719–6727, 2008.

Hegglin, M. I. and Shepherd, T. G.: Large climate-induced changes in ultraviolet index and stratosphere-to troposphere ozone flux, Nat. Geosci., doi:10.1038/NGEO604, 2009.

Kopp, G., et al., Evolution of ozone and ozone-related species over Kiruna during the SOLVE/THESEO 2000 campaign retrieved from ground-based millimeter-wave and infrared observations, J. Geophys. Res., 107, 8308, doi:10.1029/2001JD001064, 2002.

Kurylo, M. J. and Zander, R.: The NDSC-Its status after 10 years of operation, in: Proceedings of XIX Quadrennian Ozone Symposium, pp. 167–168, Hokkaaido University, Sapporo, Japan, 2000.

Li, F., Stokarski, R. S., and Newman, P. A.: Stratopsheric ozone in the post-CFC era, Atmos. Chem. Phys., 9, 2207–2213, 2009.

Nair, P. J., Godin-Beekmann, S., Pazmi´no, A., Hauchecorne, A., Ancellet, G., Petropavlovskikh, I., E., L., and Froidevaux, L.: Coherence of long-term stratospheric ozone vertical distribution time series used for the study of ozone recovery at a northern mid-latitude station, Atmos. Chem. Phys., 11, 4957–4975, doi: 10.5194/acp-11-4957-2011, 2011.

Rodgers, C.: Inverse Methods for Atmospheric Sounding: Theory and Praxis, World Scientific Publishing Co., Singapore, 2000.

Schneider, M. and Hase, F.: Technical Note: Recipe for monitoring of total ozone with a precision of 1 DU applying mid-infrared solar absorption spectra, Atmos. Chem. Phys., 8, 63–71, 2008.

Schneider, M., Redondas, A., Hase, F., Guirado, C., Blumenstock, T., and Cuevas, E.: Comparison of ground-based Brewer and FTIR total column O3 monitoring techniques, Atmos. Chem. Phys., 8, 5535–5550, 2008a.

Sepúlveda, E., Schneider, M., Hase, F., García, O.E., Gómez-Peláez, A., Dohe, S., Blumenstock, T., and Guerra, J.C.: Long-term validation of tropospheric column-averaged CH4 mole fractions obtained by mid-infrared ground-based FTIR spectrometry, Atmos. Meas. Tech., 5, 1425–1441, 2012.

Steinbrecht, W., Claude, H., Sch¨onenborn, F., McDermid, I. S., Leblanc, T., Godin-Beekmann, S., Keckhut, P., A. Hauchecorne, J. A. E. V. G., Swart, D. P. J., Bodeker, G. E., Parrish, A., Boyd, I. S., K¨ampfer, N., Hocke, K., Stolarski, R. S., Frith, S. M., Thomason, L. W., Remsberg, E. E., Savigny, C. V., Rozanov, A., and Burrows, J. P.: Ozone and temperature trends in the upper stratosphere at five stations of the Network for the Detection of Atmospheric Composition Change, Int. J. Rem. Sens., 30, 3875– 3886, 2009.

Viatte, C., Schneider, M., Redondas, A., Hase, F., Eremenko, M., Chelin, P., Flaud, J.-M., Blumenstock, T., and Orphal, J.: Comparison of ground-based FTIR and Brewer O3 total column with data from two different IASI algorithms and from OMI and GOME-2 satellite instruments, Atmos. Meas. Tech., 4, 535–546, 2011.

Vigouroux, C., Mazi`ere, M. D., Demoulin, T., Servais, C., Hase, F., Blumenstock, T., Kramer, I., Schneider, M., Mellqvist, J., Strandberg, A., Velazco, V., Notholt, J., Sussmann, R., Stremme, W., Rockmann, A., Gardiner, T., Coleman, M., and Woods, P.: Evaluation of tropospheric and stratospheric

ozone trends over Western Europe from ground-based FTIR network observations, Atmos. Chem. Phys., 8, 6865–6886, 2008.

Weatherhead, E. and Andersen, S.: The search for signs of recovery of the ozone layer, Nature, 441, 2006.

Weber, M., Dikty, S., Burrows, J., Garny, H., Dameris, M., Kubin, A., Abalichin, J., and Langematz, U.: The Brewer-Dobson circulation and total ozone from seasonal to decadal time scales, Atmos. Chem. Phys., 11, 11 221–11 235, 2011.

WMO: Scientific Assessment of Ozone Depletion: 2010, Global Ozone Research and Monitoring Project-Report No.52, World Meteorological Organization, Geneva, Switzerland, 2011.