Final Response, Ralf Sussmann, Karlsruhe Institute of Technology, Garmisch, 12 August 2011.

We thank both Christian Frankenberg and the anonymous referee for very interesting referee-suggestions which significantly improved the paper, and added a corresponding statement to the acknowledgment section. We thereafter present our point-to-point response followed by replies to the two Short Comments.

I) Response to Referee Comment by C. Frankenberg

“What is the actual motivation to use narrow microwindows? Wouldn’t larger windows (as TCCON does) result in less interference errors if proper lines of the interference species are retrieved alongside?”

This is an interesting question. Since we have demonstrated in this paper a practically interference-free retrieval with residual interference errors in the order of 0.1 % it will probably be a major effort to demonstrate further improvements. Anyway as you say our study has been performed within the state-of-the-art practice of using narrow mid-IR micro windows. Using narrow micro windows was historically motivated by the relatively high computation power requirement for the forward calculation of the very high resolution solar spectra. A successful exploitation of wide micro windows of solar spectra has recently been performed in the near-IR in the frame of TCCON (with about an order of magnitude lower spectral resolution, however). Increasing the width of micro windows leads generally to a tradeoff between reduced retrieval noise errors but increased forward model (parameter) errors resulting from accumulated systematic residuals. This is one more reason to cut out narrow micro windows – to avoid in-between regions showing prominent systematic residuals. On the other hand, our study has shown that there is one class of systematic errors which can lead to error contributions with opposite sign resulting from neighboring micro windows – namely systematic interference errors due to forward model (parameter) errors. Therefore, it might be possible that using a strongly increased number of micro windows, or even one very wide micro window, would not necessarily increase overall errors. Answering this question, i.e., whether the net effect - including retrieval noise, smoothing errors, interference errors from propagation of smoothing errors of the interfering species to the target species retrieval, as well as all forward model (parameter) errors including systematic interference errors - is positive or negative may be answered via studies utilizing the concept for quantifying absolute interference errors presented in our paper. This would be a major (computation) effort, however.

“impact of polynomial baseline fit?”

Using narrow micro windows there is practically always a significant impact of fitting a baseline with a higher-order polynomial. The standard approach is therefore to use just a linear background slope. We added this information to Table 6 (original numbering).

“Page 2967, line 3, … more correct to just state “agrees well with the WFM-DOAS v2.0 …”

Done.

“page 2967, line 21: Bousquet may not be the best reference for the loss term.”

We added the reference to Lelieveld et al., ACP, 2004.

“Page 2980, TCCON: Please cite the TCCON overview paper by Wunch et al, 2011”

Done.

“Page 2985, pressure: Do you use on-ground pressure sensors? Interpolating NCEP to a high altitude station might cause some errors (how large do you estimate them to be?).”

We added the following text to Page 2985:

Not all NDACC sites perform quality controlled surface pressure measurements as TCCON sites do. Therefore we investigated the quality of NCEP pressure information and its interpolation to an elevated site. For this purpose we performed a multi-year comparison of NCEP-derived pressure for the Garmisch station (743 m a.s.l.) versus the TCCON pressure sensor (1-min values from a high-quality pressure transducer which is regularly quality checked against a mercury barometer). We found a bias of -0.21 hPa with a standard deviation of 1.6 hPa.

“Is alpha fixed once and for all or does it depend on station and time of year?”

In our suggested setup alpha is fixed in time to the mean optimum of a 1-year test-ensemble. This is done per station individually. We added this information to the figure caption of Fig. 3 in the revised manuscript. Obviously, also the fixing in time could be released and something more sophisticated be considered, e.g., we tested even optimizing alpha per spectrum of the full time series. We did not suggest this here as a general strategy because i) the impact of such kind of alpha fine tuning (corresponding to dofs changes in the order of 0.1) on the total column retrieval is uncritical on the per mille level, and ii) this kind of optimization requires some batching environment which is to our experience hard to be transferred to and implemented at all NDACC sites.

“The author claim (page 2978, line 9) that the new scheme better integrates the measured absorption-line profile. In the current version, this claim is unsubstantiated by evidence. To prove this, a plot of alpha vs. reduced 2 of the fit would tell us whether or not the true fit quality really improves by fitting a profile.”

We agree and added such a L-curve to Fig. 3. It shows evidence for this claim as anticipated.

“This curve (in principle, an L-curve), would also enable a more objective choice of the regularization parameter“

We agree and added the L-curve along with the second derivative of the double-log L-curve to Fig. 3, and amended the text in the last paragraph of Section 2.5:

Figure 3a shows the L-curve, and Fig. 3b its second derivative which shows an optimum for an α corresponding to dofs $\approx$ 2. Figure 3c shows that at the same time one gets a dofs $\approx$ 2, a minimum for the diurnal variation is obtained (0.23 %, 1 σ). This is nearly a factor of 2 lower than the diurnal variation of 0.39 % which is obtained in case a simple vmr-profile scaling approach is used (dofs $\equiv$ 1, see point on the very left hand side of Fig. 3c). Together with the L-curve this provides evidence that the optimized Tikhonov profile retrieval accounts for true profile variations in a way that helps to better integrate the measured absorption-line profile, ...

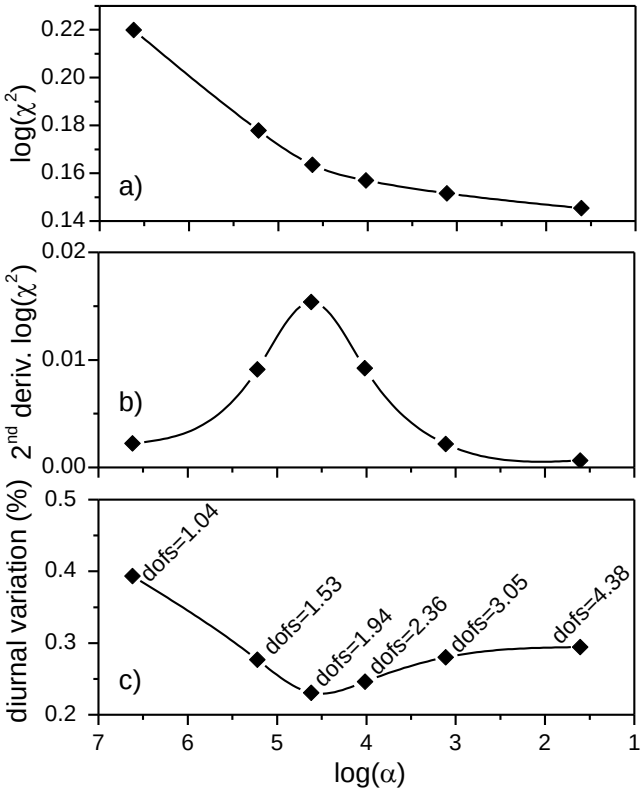

Fig. 3. Optimizing regularization strength α of Tikhonov L_1 retrievals of CH_4 using a test ensemble of all Garmisch year 2007 measurements. a) Mean L-curve, i.e., goodness of fit (χ^2) of as a function of α. The residual term within χ^2 is the overall rms-residual from the spectral fit and the noise term within χ^2 is calculated from the wave number interval 2615.25 - 2615.40 cm^{-1}. b) Second derivative (curvature) of the L-curve. c) Mean diurnal variation (1 σ). Corresponding numbers for the information content (dofs) are indicated (using a diagonal measurement covariance with a signal-to-rms-noise ratio of 500).

“RMS is typically a bad indicator of the goodness of the fit. Is SZA dominating your SNR variations (causing the RMS variability) or is it that systematic residuals are more apparant at higher airmass (also causing the RMS variability but for a very different reason).”

We agree but our initial intention was to find a quality selection criterion for the quality of the spectra (spectral noise) rather than for goodness of fit (see also answer to following comment). We used rms as a valid proxy for spectral noise because the rm-residuals are clearly dominated by spectral noise (R = 0.98, slope 0.88) and not by systematic residuals (correlation to chi2 of R =0.28) in our application, and rms is readily available from the fit output of the standard code. Anyway we agree that it is more correct to calculate true rms-noise from “out-of-band rms”. We did so (using the 2615.25 -2615.40 cm^{-1} interval) and redid Figure 4 using now “rms-noise” instead of “rms-residual”. Indeed it looks practically the same as expected.

“Why don’t the authors use a reduced 2 measure?”

We thank for this hint and added such quality selection criterion. However, this cannot fully replace a quality selection with respect to spectral noise, as a scatter plot of χ^2 versus rrms-noise / dofs shows (added as a new Fig. 4). There are still outlyers for low χ^2 which are due to very high spectral noise, and we combined both criteria. To explain this we added the following text to the second paragraph in Section 2.6:

First of all we use a threshold for χ^2 as a measure for the goodness of fit as shown in Fig. 4. However, we found that there are still some outliers for low χ^2 which are due to spectra with bad quality. To remove these we added another quality selection threshold for spectral rms-noise divided by the information content (dofs) as outlined in Fig. 4. The reason for using the spectral rms-noise to dofs ratio is as follows. Figure 5a (upper trace) shows that the time series of spectral noise contains a seasonal cycle ...

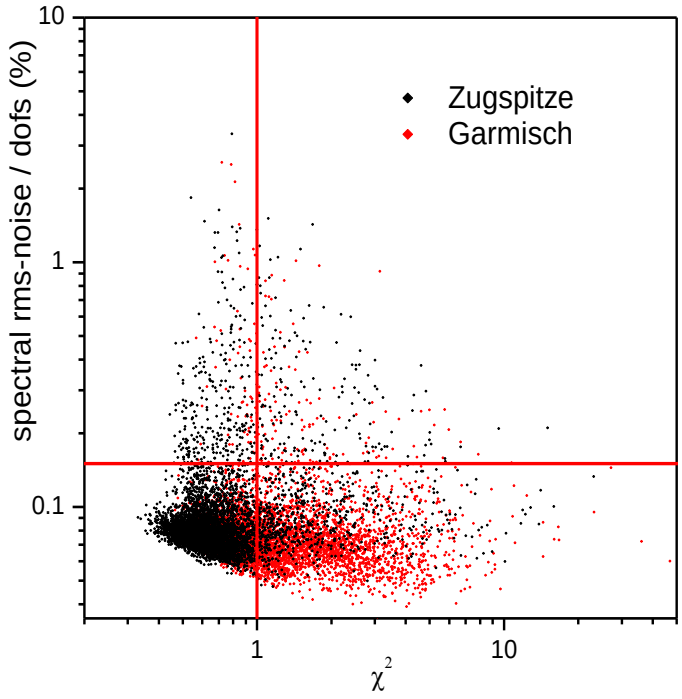

Fig. 4. Quality selection criteria and thresholds for goodness of fit (χ^2) and spectral quality (rms-noise) relative to information content (dofs). Data points are for 5 years of measurements.

II) Response to Referee Comment by Anonymous Referee #2

"As someone not connected with these ground-based networks, it is not clear to me why minimum diurnal variation equals optimum precision. Is this related to the solar zenith angle and the atmospheric path viewed? How, exactly? The paper would benefit from further explanation on this topic."

We added the following explanation to the beginning of Section 2.7.2:

The precision of the retrieved CH_4 columns (mostly limited by the impact of clouds) is estimated from the 1-σ diurnal variation of retrievals from single spectra (derived from average of several scans, ≈4-7-min integration), averaged over all individual days of the multi-annual time series. Based on the assumption that CH_4 columns are constant during each individual day, this is a means to obtain a proxy for the precision of remote sounding column measurements of CH_4. In reality, part of the diurnal variation will be caused by real variations in CH_4 over the day. Therefore this method gives an upper limit for the precision (see, e.g., Warneke et al., 2006).

"It is not clear to me what the physical meaning of the ratio of the spectral residuals to the degrees of freedom for signal is. I think some further explanation is required of what this quantity actually means in order to understand why it is of benefit in this context."

We added this explanatory text to the second paragraph of Section 2.6 (which is also partly in response to a comment by C. Frankenberg as to χ^2):

First of all we use a threshold for χ^2 as a measure for the goodness of fit as shown in Fig. 4. However, we found that there are still some outliers for low χ^2 which are due to spectra with bad quality. To remove these we added another quality selection threshold for spectral rms-noise divided by the information content (dofs) as outlined in Fig. 4. The reason for using the spectral rms-noise to dofs ratio is as follows. Figure 5a (upper trace) shows that the time series of spectral noise contains a seasonal cycle with a maximum in winter (minimum in summer) which is due to the changing solar zenith angle. This means that a classical quality criterion using a simple threshold for the rms-noise would eliminate more measurements in winter than in summer. However, Fig. 5a (lower trace) indicates that the dofs shows a seasonal cycle with same phase. This is a result of the absorption line depth changing with varying solar zenith angle in such a way that during winter there is higher dofs, as the lines are deeper. Deeper lines from (winter) spectra with higher noise level (less sun light than during summer) can be analyzed at a comparable quality (retrieval noise) level as the weaker lines from summer spectra, which show a lower average noise level. Thus, an optimized quality criterion can be utilized using a threshold for the ratio of the spectral rms-noise and dofs, see Fig. 5b

“I am not convinced that the work presented in this paper establishes the absolute accuracy of the retrievals. Therefore, unless the authors present a convincing argument otherwise, I would suggest that the title be changed.”

We claim that our paper improved both “precision” and what often is called “relative accuracy”, the latter being related to time-dependent bias and latitudinal bias or station-to-station bias.

We thereafter separate “relative accuracy” from “precision” which we both claim to improve with our paper.

Confusion between “accuracy” and “precision” is inherent in available error terminology, and Rodgers (2000, sect. 3.2.5, p 50) has made a very helpful statement on this:

“Errors are traditionally classified as systematic or random according to whether they are constant between consecutive measurements, or vary randomly. Related terms are widely used are “precision” and “accuracy”, where precision measures the variability between repeated measurements of the same state, and accuracy measures the total difference between the measurement and the truth, and includes both random and systematic errors. In practice the distinction is somewhat vague, because error sources may have time variability on a range of scales, and a source which is randomly on one scale may be systematic on another. “

Using the Rodgers statement as a basis we think the claim to have achieved a “high-accuracy-and-precision retrieval” is adequate for our paper title due to the following 3 independent arguments:

i) Our paper has reduced interference errors. H2O-CH4 interference errors are small and random type on the minutes-time scale, but large and systematic on the semi-annual scale: random-type variability of water vapor columns on the minutes-scale is, e.g., only in the order of <0.5 % (1 sigma) for Zugspitze (see Fig. 4 in Sussmann et al., ACP, 2009, or Fig. 3 in Vogelmann et al., AMT, 2011). However, there is a systematic and reproducible seasonal cycle of water vapor columns which is characterized by, e.g., a factor of up to 100 between the summer maximum and the winter minimum for Zugspitze (see Fig. 5 in Sussmann et al., ACP, 2009). Variability of water vapor columns on both scales will cause interference errors,

random type interference errors on the minutes-scale which would be attributed to ”precision”, and systematic interference errors on the seasonal scale which would contribute to “relative inaccuracy”. Our paper shows that HDO/H2O-CH4 interference effects on the semi-annual scale cause a CH4 bias up to 5 % depending linearly on the HDO or H2O columns. This is a systematic and repeatable (non-random) effect which can even be corrected by a bias correction. Therefore, HDO/H2O interference errors are related to (rel.) accuracy, not to precision on the dominant, semi-annual scale. While our optimization of interference errors helps to improve both accuracy (on the seasonal scale) and precision (on the minutes scale) by the same amount in relative terms, the above numbers show that the effect upon accuracy is a factor of 100/0.005 = 20 000 larger in absolute terms. One might use the term “relative accuracy between seasons” for such an optimized retrieval minimizing seasonal bias due to (systematic) interference errors on the semi-annual scale.

ii) (Following text will be added to the summary and outlook section).

Another outcome of our paper is that it has laid the cornerstone for obtaining a minimized station-to-station bias, i.e., improved relative accuracy for methane retrievals of the NDACC network. To improve this situation considerably became an obvious need after a study by Dils et al. (2006) had shown unacceptable numbers for the quality of NDACC methane retrievals, which induced Bergamaschi et al. (2007) to comment that “the precision of the mid-infrared FTIR measurements of 3 % and the relative accuracy of 7 % is significantly below the precision and (relative) accuracy targets of <1-2 % of SCIAMACHY measurements“. The problem of the Dils et al. (2006) study had been strongly differing retrieval strategies used by the participating groups from different stations (e.g., inconsistent HITRAN versions and priors). Therefore, we have demonstrated in our paper how to implement a harmonized retrieval strategy comprising one common spectroscopic line list, one consistent source of prior information, one regularization approach, one common source of pressure-temperature information, one set of to-be-retrieved interfering species for all stations, and one common quality selection approach. We have described and applied such harmonized retrieval strategy to the 3 test sites Wollongong, Garmisch, and Zugspitze in this paper. The benefit of these measures with respect to improved station-to-station accuracy is currently quantified using the TCCON network as an intercalibration standard (Forster et al., to be published). First results of this study show that a station-to-station accuracy of the order of 0.5 % is the result of the harmonized retrieval strategy described in this paper.

iii) Finally one formal statement. There are at least two differing definitions of accuracy in the literature, one (e.g., Rodgers, 2000) including random errors (“precision”) to be part of overall accuracy in addition to “bias”, the other (e.g., Bevington, 1969) including only “bias” to accuracy. The first is the more common one, and under this definition any improvement of “precision” is at the same time an improvement of “accuracy”.

“I think that it would be relevant to note in this paper that methods for “optimal” micro window selection have been considered by others”

We have added a related remark at the end of Section 2.3 and included the references to Dudhia et al. (2002), and von Clarmann and Echle (1998) to the revised manuscript.

See also our reply to the first question of referee C. Frankenberg.

“Abstract: The meaning of “seasonal bias” is not clear in this context. Does this refer to the systematic errors due to H2O/HDO interference? “

Yes indeed. We changed the sentence to: Dominant errors of the non-optimum retrieval strategies are systematic HDO/H_2O-CH_4 interference errors leading to a seasonal bias up to ≈5 %.

“Page 2968, lines7-8: “the sources and sinks on the regional scales” should be “sources and sinks on regional scales””

Done.

“Page 2969, line 23: “15 yr” should be “15 years”.”

Done.

“Page 2970, line 1: “parameters compilation is subject” should be “parameter compilation is the subject””

We respectfully point to the fact that the commonly used term in this context is “line parameters compilation”.

“Page 2970 Line 6: “errors in case a non-optimum” should be “errors in the case where a non-optimum””

Done.

“Page 2972: Lines 1-2: Please expand all acronyms.”

Done.

“Section 3: Could you state here what spectral region the SCIAMACHY retrievals are using? Which version of the spectroscopic parameters are used by the WFM-DOAS v2.0 retrieval? It might be a good idea to emphasize in this section why the SCIAMACHY dataset is a good dataset for comparisons here. Would it be possible to comment on how the SCIAMACHY column averaging kernels compare to the ground-based column averaging kernels?”

We added this text: These retrievals are utilizing CH_4 absorption features in channel 6 (1000–1750 nm) along with HITRAN 2008. SCIAMACHY data are a useful set for comparison because they are sufficiently sampled in time to show a significant seasonality. In addition the WFM-DOAS total column averaging kernels are close to 1 in a range between well above the tropopause and the surface. This means that the retrievals integrate the column with a high sensitivity similar to the characteristics of the ground-based soundings (see averaging kernels in Fig. A1).

“It would be more of a fair comparison to show SCIAMACHY 04/05 against Zugspitze 04/05, …”

We added the seasonality derived from Zugspitze 04/05 data to the Figure of the revised manuscript. The agreement is comparably well, but the uncertainty by interannual variability is of course larger for this reduced data set.

“Section 4: Page 2989, lines 9-10. This sentence should be corrected for grammar.”

Done.

“Page 2989, Line 12: “located at” should be “located in”.”

Done.

“Page 2989, Line 14: Unmatched bracket. Also, suggest changing “This is representative for the: : :” to “This spans the range of: : :.””

Done.

“Page 2990, Lines1-2: This sentence is awkward and the meaning unclear to me. I would suggest rephrasing it for clarity.”

Done.

“Page 2990, Line 6: Move quotation marks to encompass “internal tension” for consistency with other parts of the document.”

Done.

“Appendix A: Are the authors saying that the non-uniformity of the column averaging kernel with respect to altitude is caused mainly by H2O and HDO interference errors? Presumably there are other reasons for non-uniformity of the column averaging kernels (like the shape of the temperature and CH4 profiles themselves?) I would suggest adding some clarification here.”

No, we did not intend to say this, and of course we agree with this reasoning. We cancelled this sentence in the revised manuscript to avoid misunderstandings.

III) Response to Short Comment by M. De Maziere

“… see the last 2 columns in Tables 4 and 5, where the reasoning made in the paper should lead to an abs IF error for MW(135) that is identical with opposite sign to the rel. interference error for MW(24).

We respectfully disagree. It would be an ad absurdum exploitation of our concept to use results from MW(24) containing three “1-MW-perturbations” at the same time (minus MW 1 “and” minus MW 3 “and” minus MW 5), to “validate” another strategy (MW(135) in this case). Our concept of absolute interference error quantification uses only retrievals with on single-MW perturbation at a time. Another major problem of the “validation” you suggest is the strongly increased scatter of the retrieval results of such strategy (like MW(24)) which has only 2 MW´s left. As can be seen from Tables 4 and 5 this leads to strongly increased uncertainty (given in brackets) for the rel. interference errors which you would be trying to exploit, e.g., +0.14(8) for Garmisch MW(24).

“is the selection of the best retrieval strategy based on the rel. interference error as determined in this paper still valid?”

We added the following text and validation to the end of Section 2.7.3 of the revised manuscript:

These numbers show that the HIT00 MW(135) strategy is favorable over the HIT08 MW(1234) strategy for the medium-humidity site Garmisch. We expect that the disadvantages of the HIT08 MW(1234) strategy become even more pronounced for the wettest site Wollongong. To show this we calculated analogous numbers for Wollongong, see Table 6. Indeed, the absolute interference error of the HIT08 MW(1234) strategy approaches the unacceptable 1 % level for Wollongong (-0.82 %). Also the “internal tension” is even higher compared to Garmisch with a strong negative interference error contribution from MW 1 (-0.87(3) %) and a strong positive contribution from MW 4 (+1.18(2) %)), see Table 6.

To conclude this section, we validate our concept of calculating absolute interference errors. We give 4 validation examples. First, we derived for Garmisch from the HIT00 MW(135)

strategy an absolute interference error of -0.10 % (Table 6), and for the HIT08 MW(12345) strategy an absolute interference error of -3.71 % (Table 6). If we combine this information we would expect a relative interference error for HIT08 MW(12345) / HIT00 (MW135) of -3.71 %-(-0.10 %) = -3.61 %. This agrees with the relative interference error, which we can independently derive directly from Fig. 7, i.e., MW(12345) / HIT00 (MW135) = -3.71(7) %, with a discrepancy of +0.10 %. The second example is the analogous exercise for Wollongong: here we derived for the HIT00 MW(135) strategy an absolute interference error of +0.14 % (Table 6), and for the HIT08 MW(12345) strategy -5.22 % (Table 6). If we combine this information we would expect a relative interference error for HIT08 MW(12345) / HIT00 (MW135) of -5.22 %-0.14 % = -5.36 %. This agrees with the relative interference error which we derive from Fig. 7, i.e., MW(12345) / HIT00 (MW135) = -5.58(6) %, with a discrepancy of +0.22 %. The third validation example is again for Garmisch; i.e., combining the absolute interference errors for the HIT00 MW(135) strategy with the HIT08 MW(1234) strategy (Table 6), where we would expect a relative interference error for HIT08 MW(1234) / HIT00 (MW135) of -0.40 %-(-0.10 %) = -0.3 %. This agrees with the relative interference error which we derive directly from Fig. 9, i.e., HIT08 MW(1234) / HIT00 (MW135) = -0.51(4) % with a discrepancy of +0.21 %. The fourth validation example is the analogous case for Wollongong. Again from the absolute interference errors (Table 6) the expectation for the relative HIT08 MW(1234) / HIT00 (MW135) interference error can be derived (-0.82 % - 0.14 % = -0.96 %), and this agrees with the direct determination from Fig. 9 (-1.16(4) %) with a discrepancy of +0.20 %.

All validation results are summarized in Table 7. The overall validation result is that our method of absolute interference error estimation yields results with an accuracy at the ≈0.2 % level or better. This confirms the validity of the concept to calculate the absolute interference error from the negative of the sum of the relative interference errors. This also means that our quality ranking of the different retrieval strategies with respect to absolute interference errors (Tables 4-6) is significant. In particular, we conclude that within this 0.2 % uncertainty the HIT00 MW(135) retrieval strategy (absolute interference error +0.14 % for Wollongong, -0.10 % for Garmisch) is to be favored over the HIT08 MW(1234) retrieval strategy (absolute interference error -0.86 % for Wollongong, -0.4 % for Garmisch).

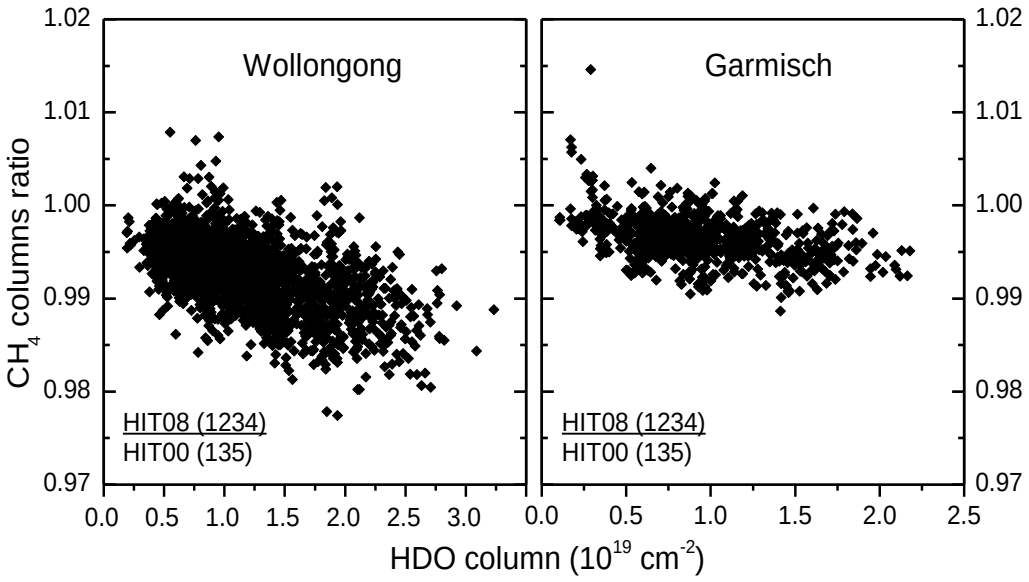

Fig. 9. Ratio plots showing significant relative HDO-CH_4 interference errors which are dominated by the unfavorable HIT08 MW(1234) retrieval strategy while the recommended HIT00 MW(135) retrieval strategy is practically interference free (see Section 2.7.3).

Table 6. Comparison of the HIT08 MW(12345) and HIT08 MW(1234) retrieval strategies versus the recommended strategy HIT00 MW(135). Numbers are for Garmisch (standard font) and Wollongong (bold). Use of HIT08 MW(12345) is out of discussion (Wollongong absolute interference error -5.22 %) but also the use of the HIT08 MW(1234) strategy is strongly discouraged because of i) high absolute interference errors (e.g., -0.82 % for Wollongong) and ii) strong "internal tension" (strong rel. interference error contributions from differing micro windows with opposite sign, e.g., -0.87 % versus +1.18 % for Wollongong).

micro windows used		MW (12345)	MW (2345)	MW (1345)	MW (1245)	MW (1235)	MW (1234)	MW (135)
HITRAN 2008	dofs						1.75	
							1.75	
	diurn. variat. (%)						±0.28	
							±0.31	
	rel. IF error (%)		-0.64(4)	+0.22(1)	+0.08(0)	+0.74(2)	+3.31(9)	
			-0.87(3)	**+0.38(1)**	**+0.13(0)**	**+1.18(2)**	**+4.40(6)**	
Garmiisch, **Wollongong**	abs. IF error (%)	-3.71					-0.40	
		-5.22					**-0.82**	
HITRAN 2000	dofs							1.80
								1.80
	diurn. variat. (%)							±0.26
								±0.27
	rel. IF error (%)		+0.01(2		+0.12(0)		-0.03(4)	
			-0.07(1)		**+0.10(0)**		**-0.17(3)**	
Garmisch, **Wollongong**	abs. IF error (%)							-0.10
								+0.14

Table 7. Four validation cases using two independent ways of estimating relative interference errors. The discrepancy is a measure for the uncertainty of the method of estimating absolute interference errors. For details see text.

retrieval strategies	HIT08 MW(12345) / HIT00 MW(135)		HIT08 MW(1234) / HIT00 MW(135)	
site	Garmisch	Wollongong	Garmisch	Wollongong
rel. interf. error estimated from abs. interf. errors (Table 6)	-3.61 %	-5.36 %	-0.30 %	-0.96 %
rel. interf. error derived from Figs. 7 and 9	-3.71(7) %	-5.58(6) %	-0.51(4) %	-1.16(4) %
Discrepancy	+0.10 %	+0.22 %	+0.21 %	+0.20 %

"If the various retrieval strategies have different sensitivities to water vapour interferences and if these are really 'perturbing' the CH4 retrieval, why are the diurnal variations on each row in Tables 4 and 5 almost the same in all cases?"

The diurnal variation of columnar water vapor, e.g., above Zugspitze, is in the order of a few 1/10 mm while the seasonality covers a range of 10 mm (see e.g., Sussmann et al., ACP,

2007). Therefore, the contribution of interference errors to CH4 diurnal variation is in the order of a few 10 times lower than the maximum interference errors we derived from full annual time series and which are in the range of a few 1 %. I.e., the interference contribution to the CH4 diurnal variation is in the 0.01 % range. We observe 10 times higher diurnal variations, namely in the few 0.1 % range. These are understood to be dominated by cloud apodization effects. The latter are partly compensated for by our (Tikhonov) profile retrieval, and the efficiency of this mechanism depends slightly on the details of the retrieval strategy which may explain the existence of small differences in diurnal variation.

“And can the authors clarify why a minimal diurnal variation is not another criterium to be verified by the selected "best" retrieval strategy ?”

We think the diurnal variation should be verified. Maybe this had been to weakly expressed. Therefore we added some more examples for the verification of the diurnal variation to Section 2.7.2 of the revised manuscript:

One note on the HIT08 MW(1234) strategy. This strategy might be favored by the “esthetic” reason that it comprises latest version HITRAN. However, it will be shown in Section 2.7.3 that this strategy causes significantly higher interference errors than our recommended HIT00 MW(135) strategy comprising HITRAN 2000. Table 6 shows two further disadvantages of using HIT08 MW(1234): for Garmisch (Wollongong) there is an increased diurnal variation of ±0.28 (±0.31) compared to using the HIT00 MW(135) strategy which leads to ±0.26 (±0.27). Also the information content is lower using HIT08 MW(1234), namely 1.75, compared to 1.80 attainable by using HIT00 MW(135).

“The paper discusses essentially the precision of the retrievals. I haven’t found the arguments for a high accuracy of the selected retrieval strategy”

See our answer to anonymous referee #2”.

IV) Response to Short Comment by F. Hase

“Recently, it became evident that the current NDACC IRWG CH4 guideline falls short under certain conditions. …We acknowledge that the work presented by Sussmann et al. is a careful study of the current NDACC infrared working group (IRWG) CH4 retrieval guideline. The authors prove that other schemes can be found which are superior to the current official NDACC guideline, especially for wet sites.”

We thank for this endorsement.

“However, we have objections with regards to the proposed retrieval scheme to replace the current NDACC retrieval guideline, as the new scheme still might require further optimization.”

The intention of our paper was to initiate a major science progress in a special situation, namely the NDACC community performs CH4 retrievals which are dominated by artifacts. These artifacts were successfully quantified in our paper for the first time (up to ≈5 % with the current NDACC guideline), and minimized down to the ≈0.1 % level (comparable to TCCON) via a new retrieval strategy, which allows for the first time to reliably retrieve the true (1-%-amplitude) CH4 seasonality in the northern hemisphere.

We agree with you that further improvements beyond our paper may be achieved as research goes on naturally during coming years. However, we don´t see in your comment a reasoning or geophysical target for “further optimization”, which you wanted to achieve

beyond the (≈0.1 %) error level achieved as an outcome of our paper, nor did you present a related time line.

In that context we like to mention that “a decision on a NDACC retrieval guide line” will be performed on a different level, as AMTD is not a forum targeted to that. In addition to taking science input into account such decision will likely include other aspects like contributions to existing international projects. In particular, our group is responsible for CH4 deliverables in such project (to which you contribute), and these are required on much shorter time scales than the time needed to implement all “further optimization” you might desire at some end.

“Our main concern is (1) that two of the proposed microwindows (MW 3 and MW 5 in Sussmann et al.) suffer from strong H2O (HDO) interference”

We are surprised about this statement, because our paper proves the opposite. We performed a quantitative interference error treatment and demonstrated negligible interference errors from MW 3 and MW 5 on the ≈0.1 % level, see our Table 4 (for using the recommended retrieval strategy including MW(135) along with HITRAN 2000). We remain unclear whether you missed our error numbers in Table 4, or your desire is to significantly improve beyond the ≈0.1 % level, which might be ambitious.

Statements like “suffer from strong H2O (HDO) interference” are failure prone if based on qualitative considerations only. You might consider to carefully distinguish between i) the qualitative visual finding of (possibly) interfering lines in a spectral window, and ii), the quantity of a resulting interference error. The interconnection between i) and ii) is not at all trivial: not every strong interfering line leads to a strong interference error, the magnitude of interference errors depends often more on the existence of hidden spectroscopy errors than on obvious spectral overlap of interfering lines with target species absorptions. A way out is to perform a quantification of absolute interference errors. Obviously, this was missing in your consideration.

“Our main concern is (2) that the highly variable interfering species H2O (HDO) is not treated with the required care to minimise the interference error”

We thought a careful treatment of the interfering species H2O/HDO had been the main subject of all our paper and the outcome would be a success (showing H2O/HDO interference errors below ≈0.1 %).

Anyway, obviously you “believe” that additional (joint or offline) retrieval of water vapor profiles might lead to a further significant reduction of interference errors below the ≈0.1 % level achieved by profile scaling of water vapor. However, you have not shown this quantitatively yet, since all figures in your comment do contain only qualitative information (i.e., relative interference errors).

We think your suggestion of a “pre-determination” of water vapor profiles (either offline or online using out-of-band lines) should be reflected with some care: i) pre-determination may help to reduce part of the interference effect, namely the one from propagation of smoothing errors due to the variability of the interfering species to the target species - as described in detail by Sussmann and Borsdorff (ACP, 2007). Such improvement is the dominant effect in the limiting case that spectroscopy for the interfering species is perfect, the kernels are ideal, and/or the a priori does not differ from the retrieval. ii) If such ideal conditions are not fulfilled, you have counteracting effects. E.g., if there is a spectroscopic inconsistency between the water vapor lines(s) used for predetermination and the interfering water line(s), a pre-determination will increase the systematic part of the interference error, because the local residuum of the interfering species around the target line will be increased and miss-interpreted by the target species retrieval. You always have a tradeoff between such effects i) and ii) and principal considerations don´t help to decide whether the net effect of a pre-

determination of water vapor profiles would be positive or negative – again, you need a full quantitative treatment of absolute interference errors. This is missing in your comment.

Here we meet another problem related to your suggestion of an offline pre-determination of water vapor profiles: we know you are performing such pre-determinations since years for all kind of retrievals, but we have not seen a rigorous quantification of the interference errors related to that approach. In order to perform a quantitative error estimate for this approach, you consequently would have to propagate the error analysis of the water vapor profile retrieval through the error analysis of the subsequent methane retrieval including all positive and negative effects i) and ii) described above. You have not done this up to now and indeed this seems nearly impossible to us. So in a sense you have to believe or hope that the above positive effect i) dominates over the possible negative effects ii).

There is also a major strategic drawback of using (out-of-band) water vapor profile retrievals as input: as you know only few groups have experimented with water vapor profile retrievals (including yours and ours), and these retrievals are far from the level of being a validated NDACC standard approach that would be readily available for all sites (with their strongly differing humidity levels).

Note in this context, that Sussmann and Borsdorff (2007) have shown a method for joint (in-band) profile retrieval of interfering species that minimizes interference errors and includes a rigorous interference error quantification. We have not applied this approach to joint (in-band) water vapor profile retrievals in case of methane retrievals because i) even with simple water vapor profile scaling we achieve interference errors <0.1 % and therefore see no geophysical need for further improvements, ii) the computation effort would be is relatively high, and iii), this approach would be difficult to be transferred to all NDACC groups.

Last but not least we don´t see the innovative aspect of your suggestions to investigate again “new” micro windows in a merely qualitative approach as we all did since 15 years. We think what was needed was an approach for absolute quantification of interference errors. This is available now from our study and will hopefully be applied in future studies testing “new” micro windows.

“Our main concern is (3) that a remigration to outdated HITRAN line lists is required.”

We have shown that HITRAN 2000 contains smaller errors in the spectral range of our application than HITRAN 2004 or HITRAN 2008. We think it is scientifically correct to use the best available data base – even if it is not the last version – and forward at the same time the message to the HITRAN community that there is a problem with the new versions. What you suggest is to hide this finding and try to find complicated workarounds with lower quality HITRAN 2008.

Please note also that we have an agreement with the NDACC infrared working group to only use official version HITRAN line lists. This agreement is fulfilled with our suggestion to use HITRAN 2000.

On the other hand you suggest “to apply an ad-hod correction of HITRAN 2008 line parameters to improve fit quality“ - and thereby ignore this very NDACC agreement.

End of response