We thank both Referees for their valuable and detailed comments. We feel that the manuscript has improved significantly after addressing all the questions. Below the Referee comments are bolded and our answers are written in normal text.

Anonymous Referee #1

Kajos et al. present comparison of concentrations of selected VOCs from a forested field site using 2 PTR-MS and 2 GC-MS instruments. The main conclusion from the paper is that independent settings, calibration approaches and variable instrument sensitivities can lead to relatively high uncertainties of reported concentrations, which could penetrate to models. Consequently, the collocated measurements with more than one instrument are advised to aid in confidence.
Simultaneous measurements of VOCs by multiple instruments are not usually performed, so the paper has a unique value to the PTRMS and GC community in learning about the accuracy of long-term measurements. In addition, simultaneous measurements make it easier to diagnose problems (e.g. compounds breaking through in the microtrap, leaks, etc.).
It is unfortunate that the paper does not give a deeper scientific insight into observed discrepancies, as the recommendations could be synthesized to be even more useful for the community. Maybe this could still be done. While I recommended the paper is published subject to minor corrections, I do think the clarity of the paper has a high potential for improvement and may require a major effort.

General:
1) The authors did a good job with transparent presentation of concentrations from different instruments. Although for certain compounds comparison looks quite impressive, there are also some "scary" discrepancies highlighted. Significant problems should be identified and either pointed out since the beginning or excluded before the comparison is made. Otherwise the reported discrepancies might exaggerate how bad the accuracy really is. I think it is true that certain problems cannot be easily spotted if there is only one instrument, so acetone and methanol are probably good examples here to show. This leads me to a question: Has a strict quality control been performed on all the data (assessment of quality of zero air, breakthrough volumes, water condensations, leaks, contaminations, assessment of materials used in the sample lines, stabilization of "sticky" compounds in calibrations, etc.)?

The reviewer makes a very good point here. The justification of this study derives from the fact that most longer term measurements have been conducted with a single analyzer, thus making it difficult to assess problems in the measurements. Thus our aim is to quantify the uncertainty of the measurements conducted with these typical on-line analyzers operated with the procedures normal to the field experiments. We have highlighted this in our revised introduction

We have removed data in which an obvious problem can be noticed, i.e. unphysical concentration values, calibration periods etc. As the measurement campaign was relatively short there are only very few of those incidences. However, we have abstained from removing data on the reasons of bad correspondence between different instruments.

The quality control (QA/QC) used during the campaign and data analysis is the same as used for our standard measurements with these instruments. The (QA/QC) procedures have now been added as supplementary material to this paper.

2) The story is generally difficult to follow. While English is generally clear, there are language imperfections in various places. Consequently, the story, while generally interesting, makes impression of incoherent. It is surprising to see relatively little attention taken to comparing the actual results, the discrepancies and to an investigation of actual or potential problems behind those discrepancies.

We have tried to make the text more coherent and also added more in depth investigation of the discrepancies.

3) Nowhere in the manuscript are shown the actual calibrations. Does the fact that they were performed relatively rarely (once per week?) is sufficient to say how stable the sensitivities were? Did the sensitivities vary diurnally as relative humidity changed?

A more detailed description of the calibration procedure was added to the SI. We have not tested the humidity dependence of the sensitivities. However, when analyzing the long term calibration data (95 calibrations between July 2011 and August 2015) we see that the ratio of the sensitivities of two consecutive calibrations is close to unity indicating that our calibration interval is not too long.

4) Typically much better agreements have been reported for methanol and acetone (e.g. Davison et al., 2009), but, here, it might be confusing that some instruments (PTRMS2) were not collocated and did not exactly coincide in timing, so perhaps a perfect agreement should not be expected? Also, it is surprising to see so drastically lower sensitivity for methanol in dynamic vs manual calibration which I guess might not have been an issue if the longer purging times and MFCs with appropriate materials (e.g. Kalrez instead of Viton o-rings) had been used.

It is true that on occasion better agreement have been accomplished for methanol and acetone. We agree with the referee that the co-location is one of the issues making the situation worse in our case. This is particularly important for water soluble compounds such as methanol and acetone as they are readily lost in the canopy explaining some of the discrepancies in our result. The tubing in the different calibration setups differ from each other potentially giving rise to the differences in the methanol calibration as more stainless steel parts are exposed in the dynamic calibration system. This causes additional losses decreasing the sensitivity with regard to methanol in particular. Additionally we would like to point out that there are challenges in the methanol measurements with the GC as well. In our field study, only one GC-MS was measuring methanol. The methanol signal of the GC-MS was corrected based on post-campaign calibration that indicated methanol loss of 55% with water in the system. A small error in this number drastically change the GC-MS derived methanol concentrations. See page 3776, lines 25 onwards. As a note, the inlets in the study by Davison et al. (2009) were also not collocated. As instrument comparison was not main aim of that study, the description of measurement procedure is not as transparent as it could be thus making it harder to evaluate their study against ours.

Specific:

5) P3755 L2, Is this the best sentence to start the abstract from? What do you mean by "real time"? Even the fast GC is not a real time measurement or even close to real-time. Here, you show 1 hour data.

We agree that "real-time" is not the best expression here. It was changed to "automated *in-situ*"

6) P3755 L14 "notably for methane" – methane cannot be measured by PTR-MS. Did you mean methanol?

Yes. Corrected now.

7) P3755 L18 "This mismatch indicates that the systematic uncertainty in the sensitivity of a given instrument can lead to an uncertainty of 50–100 % in the methanol emissions measured by commonly used methods.". Is it if the Viton elements are used in MFCs or was there an issue with the equilibration time of methanol (e.g. if the calibration steps were short?)? I would suggest discussing more the reasons for these and how these could be avoided. This reviewer is not convinced that careful measurements of methanol can lead to so high uncertainties.

The MFCs of the automatic calibration system do not have any Viton elements, instead, Kalrez was used. We believe that the Kalrez elements do not explain the discrepancies between the instruments.

8) P3756 L14-23 I am confused by this paragraph. I guess you probably want to say generally that the inaccuracies in the models may often result from inaccuracies in measurements or do you suggest that earlier measurements might be inaccurate?

We changed the last sentence of the paragraph to: **"**Evaluation of these models require accurate measurements of atmospheric properties including the VOC concentrations."

9) P3756 L24 "Traditionally . . ." and you cite fairly recent papers. Then (L.28) you have: "Recently,.. ." and you cite papers from 2 decades ago?

This is true, we modified the text to: "VOC concentrations have often been measured by collecting samples into canisters or onto adsorbents with subsequent off-line analysis with gas chromatography-mass spectrometry (GC-MS) or gas chromatography connected to a flame ionization detector (e.g. Grosjean et al., 1998; Na et al., 2001; Hakola et al., 2009; Sauvage et al., 2009). In addition to automated measurements based on both GC-techniques and chemical ionization techniques have been developed and utilized (e.g. Lewis et al., 1997, Hakola et al., 2012)."

10) P3757 L17 "Typically, a long-term measurement setup consists of a single analyzer, which is periodically calibrated.". I do not think it is typical to calibrate rarely (e.g. 1-2 times a day may be more appropriate for long term measurements), but I agree that the discrepancies may derive from too seldom calibrations and zero air measurements. This is an important point to make in the paper from the beginning. A recommendation for frequent calibrations and zero air measurements have not been made by the authors but could be relevant for including in conclusions.

We replaced "periodically" with "regularly". PTR-MS1 sampled zero air (BG) every 2nd hour and PTR-MS2 every 3rd hour. Thus in this case the discrepancies were unlikely due to the BG measurements. See also answer 3).

11) P3758 L3 "The main aim of this study was to evaluate how reliable the real-time measurements of aromatic and oxygenated VOCs are when a single stand-alone instrument is used.". It would be interesting to see a deeper insight into the actual reliability of the measurements in the light of the specific factors that influence it.

We have added a more in depth analysis of the discrepancies.

12) P3758 L5 again you refer to "real-time" measurements, but maybe you just mean "online" measurements or it might make sense to say "close-to-real-time".

See answer 5)

13) P3759 L19 "cycle" is unclear and can be confused with the duty or MID cycle.

To avoid confusing usage of cycle, we modified the sentence to: "The measurements were obtained at six heights (4.2, 8.4, 16.7, 33.6, 50.4 and 67.2 m) for one minute at a time at each height before switching to the next height."

14) How much data is rejected between heights switching from PTRMS2? For example, is one minute data comprising all 60 s or is it something like 50 s or less?

The sampling cycle is now explained in more detail: "One sampling cycle (MID cycle) lasted for 55 s, during which 27 different compounds were sampled sequentially each with a 2 s dwell time. Instrument related signals, such as primary ions and oxygen (which is used as a marker for impurities inside the instrument) were sampled with a dwell time of 0.2 s. This means that each compound was effectively sampled for 9×2 s during one hour."

15) P3759 L23. Were all the 6 lines of equal length (100 m)? Where in the set-up did the switching occur? What type of valve was used (what type of materials)? Was the tubing in any of the lines heated?

The switching was performed close to the PTR-MS with a solenoid valve (chamber volume of 11 μl, FKM diaphragm). From each measurement height, the sample air was drawn into the PTR-MS2 via a heated 100 m-long-inlet line with a constant flow of 45 l min^{-1} (Teflon PTFE, 14 mm id). In this comparison study, only the measurements taken at 8.4 m were used. The lines were heated a few degrees above the ambient temperature.

16) P3760 L12 ". . .cannot be used..." seems too strong here. Enhanced PTR-MS selectivity is possible in variable E/N mode (e.g. Maleknia et al., 2007, Misztal et al., 2012).

Instead of cannot be used, we softened the sentence" Therefore it is not used for exact identification of the measured compounds." and mention the possibility to improve the selectivity: "However, the selectivity of the instrument can be enhanced by varying the E/N (Maleknia et al., 2007; Misztal et al., 2012)."

17) P3760 L21 120-140 Td. 110-140 Td has commonly been used (e.g. Ghirardo et al., 2010).

Corrected

18) P3760 Until this point it is unclear to the reader what the dwell times were for each instrument and m/z (perhaps you could consider tabulating this information). Later in the text 2 s is shown as a dwell time for PTRMS2. Was it used consistently for all PTRMS instruments and m/z (including m/z 21)?

Dwell times were added to the text

19) What was the cumulative dwell time for PTRMS2 per each tower level per compound (i.e. dwell time of compound multiplied by the number of MID cycles in a minute (minus the trimming time due to height switching). It needs to be clear what 1 min data from each height effectively was. Possibly, it was just a few seconds for each mass per minute depending on the dwell time and how much you had to trim due to height switching.

See answer 14)

20) P3760 L23 The sentence "Therefore both H3O + and H3O +H2O ions are taken into account in the data processing" is too general for general audience. I assume you mean the data are normalized to hydronium ions and hydrated hydronium ions to account for variation of the signal due to additional protonation of VOCs from colliding water clusters. Do you also consider humidity dependent sensitivities from calibrations at range of humidities (e.g. de Gouw et al., 2003) or was the zero-air source and standard gas of the same humidity?

To explain the normalizing more clearly, we wrote: "Therefore, the measured signals were normalized to 10^6 primary ion counts using both H_3O^+ and $H_3O^+H_2O$ signals. As benzene and toluene react slowly with the $H_3O^+H_2O$, only H_3O^+ ions were considered for the normalization. Heavier water clusters were not taken into account."

Humidity dependence of the sensitivities was not taken into account. However, the zero air generator was using ambient air, thus the relative humidity of the BG signal is the same as the relative humidity of the sample air.

21) P3761 L5-13. This paragraph refers to Taipale et al., 2008. However, some information should be more transparent without a reader having to refer to a separate paper, in particular:
a) Was the normalization done to water cluster at m/z 37, 38 or 39?

We have added the m/z was used for water cluster measurements to the text: **"**The signals of primary ions are measured during each measurement cycle, but in order to maximize the lifetime of the detector, the count rates

of the primary ions were measured at the m/z 21 and m/z 39, corresponding to the isotopes $H_3{}^{18}O^+$ and $H_3{}^{18}O^+(H_2O)$, respectively."

b) Was the ion at m/z 55 included to normalization for water clusters?
Only the first water cluster $H_3{}^{18}O^+(H_2O)$ was included in the normalization. This is mentioned in the text now.

c) Was the normalization performed to drift pressure?
Yes drift tube pressure was normalized to 2.0 mbar. This was added to the text.

d) Was the procedure consistent for all PTR-MS instruments, their calibrations and ambient measurements?

As said in the discussion paper, "Exactly the same procedures were applied for the calibration and VMR calculation for both PTR-MS instruments".

22) When normalizing signal to primary ions, do you pre-average the signal at m/z 21 or do you divide by each point in an MID cycle. If the latter, how are the detection limits affected (i.e. if you are introducing the noise from m/z 21)?

The primary ion signals are pre-averaged using 5-minute running mean. This information was also added to the text.

23) P3762 L10. I like the idea to calculate the detection limits for each ZA air measurement. One thing to clarify is to say what period those detection limits correspond to. Unlike it is the case with the sensitivities, the detection limits will be different for differently averaged data, for example, 1 s and 1 min. Calculation of detection limits corresponding to ambient measurements at 1 min depends on the total counting statistics related to dwell time of each compound and the length of the MID cycle. I could not find how much total time per compound is spent for each height (see earlier comment 21). Thus, it is unclear if the detection limits you report are just to characterize instruments or do you care to reflect the actual detection limits for the reported ambient data accounting for the exact number of samples (cycles) to average.

When averaged data was used, detection limits were calculated for the respective time as well.

24) P3765 L10. It seems to me that you are describing here the white noise distribution, not the PTR-MS statistics distribution. When you average the data, the averaging filter is also a noise filter, so Poisson noise is reduced or eliminated with averaging. There is also a component related to true ambient variability of a compound at short time scales, but the precision must necessarily be much lower at short time scales.

We are not sure if we fully understand the comment. We agree that Poisson noise is actually reduced with increasing dwell time as the ratio of $\sqrt{I_{counts}}$ and I_{counts} reduces with increasing number of counts. We are not taking in to account the effect of the true ambient variability to the average concentrations, as most of the comparison was done with three instruments measuring from the same inlet, experiencing the same variations.

25) P3765 last sentence: "Thus, the primary ion signal uncertainty is less than 1 % and it was neglected.". This is surprising. Even if overall the m/z 21 signals were similar, when you normalize to primary ions you introduce Poisson noise which, as mentioned above, will be different at different dwell-times and/or averaging times. Thus, I think you should consider including uncertainty in primary ion signal unless you did not normalize or you sufficiently pre-averaged the m/z 21 signal.

m/z 21 was pre-averaged. See also answer 22.

26) How many dilution steps of the standards were done and what was the duration of each step? Did the steps look well equilibrated (particularly methanol?)? Given the nature of the article I am surprised not too see any figures showing the actual calibration curves or steps from the instruments. These would be extremely useful (e.g. in SI).

Only one dilution step is used for the PTR-MS calibration. However, the linearity of the calibration has been tested.

27) Uncertainty of the PTR-MS measurement. This section seems a little bit vague. Not only the precision is important but also the accuracy. How about other things that impact uncertainty? For example, SEM voltage optimization if not done regularly can lead to decrease of sensitivities. Performance of MFCs in the dynamic calibration system should be assessed regularly (were those checked for leaks, crosscalibrated with an accurate flow meter?). What about line losses? Was the calibration done through the 100-m line? Finally, it is certainly interesting to see the variabilities in sensitivities, but the real question is how frequency of calibrations and zero air impact these uncertainties in terms of accuracy?

The SEM voltage is regularly optimized and the flows are checked with a flowmeter during each calibration (see SI). The line losses have been studied by Kolari et al., (2009). We added short description about the line losses to the text.

28) P3769 L14 "The concentrations measured with different instruments had temporal discrepancies, as all of the instruments had different sampling intervals. PTR-MS1 measured several compounds sequentially, each with an integration time of 2 s, which lead to a 1 min resolution." Do you mean you only got one point per one minute? In other words your 1 min data effectively contains only 2 s of the actual data?

Yes, during one hour PTR-MS1 was sampling each compound 43×2s.

29) P3769 L17 "The ambient concentrations were measured 43 times during each hour, after which the background was sampled 11 times." Do you mean that each compound was sampled effectively for 43*2s, so ~1.5 min per hour, or is it even less because you were switching the heights?

Yes. PTR-MS1 was measuring from one height only.

30) P3770 L20 "A constant ratio was assumed for the sensitivity and its uncertainty". Interesting, so you suggest that uncertainty of sensitivity is the same directly following the calibration and in a few days after the calibration? Again, I think frequent calibrations can be the key to achieving high accuracy.

Yes this is the assumption we made.

31) Which sensitivities (MCM or ACM?) were used to derive concentrations for PTRMS1 and PTRMS2 in Figure 2 and 3?

All PTR-MS calibrations were performed using ACM.

32) Figure 4. Why did you remove the outliers? How did this operation affect the mean? Can you add the mean values to the box plots?

The outliers were only removed from Figure 4 (to make the figure more clear) not from the data. The mean values were added to the box plots.

33) Conclusions: "Thus, it can be easily estimated that e.g. any emission measurement of methanol has an uncertainty of 50–100 % due to the sensitivity of the instrument used. The results of this study show that when doing long-term measurements of ambient air, occasional comparison measurements are needed to validate the measured concentrations, even if the instrument is calibrated regularly. ". I agree

with the recommendation of occasional cross-calibration exercises when conducting long-term measurements, but it is surprising that the authors generalize about so high uncertainty in methanol measurements. This probably could be linked to metal (and/or Viton) surfaces of a particular MFC and/or equilibration time, and I think more discussion of these issues could be interesting for readers.

We believe that 50-100 % is a realistic estimate for the bias between different measurement systems and it is based on the results of this study. By comparing two identical systems with identical calibrations, done preferably with the same calibration system and calibration gas we probably get smaller bias. However, in reality in our scientific community the measurements are conducted in the field by different investigators using different methodologies, and this will lead to surprisingly large biases between measured concentrations.

We have now included a section with questions arising from our intercomparison experiment, and we include the reasons for large biases into this list.

Technical:

34) P3755 L9: Duplicated sentence: "This paper presents correlations. . ..".

Corrected

35) Multiple places: "PTR-MSs" or "GC-MSs". Might consider "PTR-MS instruments" or "GC-MS instruments/machines/systems/etc.".

Corrected

36) P3760 L4 "pumped" should be "drawn".

Corrected

37) P3760 L24 remove dot.

Corrected

38) Figure 2, the traces look in many places unclear – the markers are big and obscure the lines. You could consider all lines (e.g. color-coded) instead of markers and lines.

We replotted the figures.

39) Figure 5, can you add goodness of fit and slopes?

We have now added slopes and goodness of fit to Figure 5.

Anonymous Referee #2

General Comments:
The study by Kajos et al.. seeks to address the reliability of ambient measurements of aromatic and oxygenated VOCs when different instruments are employed for measuring air in the same environment through a field intercomparison exercise. In their experiments conducted in a boreal forest site, the authors intercompared ambient data acquired using two proton transfer reaction mass spectrometers and two GC-MS systems. Such a study is laudable as it is never easy to undertake an intercomparison exercise during field measurements and the outcome of such studies is valuable for highlighting

technique specific strengths and limitations under ambient conditions. While the manuscript is generally well written, I feel the manuscript needs to be strengthened in several aspects before it can be published in AMT. Several questions/concerns arise in the present version pertaining to the robustness of the results and the validity of the conclusions for practitioners of these analytical techniques.

Major Concerns:

1) Calibration data of the individual instruments are not presented and discussed in detail. In an intercomparison of techniques, the raw measured data acquired during calibrations often holds clues to instrumental issues, in particular if the calibration experiments are carried out under relevant ambient regimes encountered at a site (e.g. at different relative humidities and ambient temperature). The authors mention that calibrations were performed atleast thrice during their month long study. It would be instructive to present and analyze the data from the calibration experiments in the paper. How long were the lines conditioned with flow of VOCs during the calibration experiments? How do chromatograms measured by the two GC-MS systems compare while measuring from the same VOC standard and for the common ambient periods? Are the peaks always equally well separated by both GC-MS systems on occasions where there was poor agreement? The QA/QC employed by the operators of the different instruments for both the calibration experiments and ambient data need to be documented in more detail in the paper for the reader to have confidence that all known precautions were followed and the results are really specific to instrumentation and not operator skills/know-how.

We have described the calibration and QA/QC procedures in the supplementary information.

2) For periods where there is better agreement between instruments and periods where the agreement was not good, were the ambient air conditions significantly different in terms of humidity and temperature? Perhaps a trace of the ambient temperature and RH could be added to Fig 2 and Fig 3. Water vapour is a common problem for several instruments and for VOCs such as methanol and benzene measured with the PTR-MS the effects can be very different on the sensitivity as also acknowledged by the authors. Were calibration experiments performed over different RH ranges (0% to 90% RH) specifically to test for the magnitude of such effects in different instruments?

We studied the discrepancies between the instruments against temperature, relative humidity, wind speed and wind direction. The results of this analysis have now been added to the manuscript. Temperature and relative humidity were added to figures 2 and 3. We did not test the effect of relative humidity to the calibration of different instruments.

3) Toluene discrepancy ; the authors state that being an anthropogenically emitted VOC they do not expect toluene variability to be high over short time scales in the forest where they were measuring. I am not entirely convinced that this is true all the time even in a remote biogenic setting. A number of studies have shown that under stress, many plants including Scots pine (a major tree species at their site) can emit benzenoid compounds such as toluene (see for example: Heiden et al. GRL, 1999). As the sampling time varied between the different instruments, this could explain part of the discrepancy

This is a good point. However, Rantala et al. (2015) reported that during April and the biogenic emissions of are still low at the site. We have observed low emissions in the beginning of May as well. Thus, we do not think that the discrepancy is caused by local the biogenic emissions. This was added to the manuscript as well.

4) The use of r values for interpreting correlation/ agreement between instruments is a bit misleading to me. For example, an r value of 0.5 for acetaldehyde reported by the authors is still interpreted as a correlation, while to me r of 0.5 is actually only an r2 of 0.25 , which implies only 25 % of the data co-varied (which is more lack of a correlation than any correlation ! I would suggest use of r2, even though it may not look better.

We decided to leave the correlation values as they are. We have calculated linear correlations and we do not think that it is misleading as it R^2 can be easily calculated from the R values.

5) Fig 5 has very poor resolution in the web version, and much of the information was obscured so it should definitely be made more legible.

The bad resolution of the figure 5 was caused the format of the discussion paper and should be better in the final paper.

6) An outcome of a paper of this kind should be to inform the community about detailed technical aspects and potential artefacts intrinsic to different types of instrumentation so that lessons can be passed on to the community at large. This aspect is really in short measure in the current version of the manuscript and I would encourage the authors to collectively strengthen the same by focusing on suggestions made above and in the other review.

This is a very good comment and we have added recommendations based on findings of our study. Additionally we have listed some open questions arising from this study.