We thank you for your review. Your comments and suggestions have certainly helped to improve the manuscript. Please find our responses to your comments below.

OVERALL COMMENTS

Comment (1): First, the Abstract does not convey clearly the 'framing' and motivation of the study. The reasons for conducting the comparisons are not given. For example, a conclusion is reached about trends being derivable from the ECC sondes but no statement is given about the reliability of the MOZAIC instrument. Statements like 'close agreeement' are not defined. Presumably the values referred to are better than the 20% discrepancy quoted for Brewer-Mast instruments. This is but one example of vagueness that does not make for a convincing conclusion. The less than clear Abstract reflects deficient organization in the paper. A series of scientific questions is not addressed but rather a recitation of statistics that are not prioritized is given.

Response: *In the revised manuscript, we modified the Abstract to convey the motivation of our study more clearly. Staufer et al. (2013) and Logan et al. (2012) recently raised concerns about the long-term stability of MOZAIC measurements based on NOXAR-Payerne/MOZAIC-Payerne and MOZAIC-Zugspitze comparisons, respectively. This forms the motivation of the paper because we want to see if this feature can be seen at stations other than Payerne. Second, the conclusion about trends has been revised. Due to the discrepancies between most ozonesonde launch sites and MOZAIC during the mid-nineties, we conclude that neither ozonesondes nor MOZAIC are reliable enough for robust long-term trend analyzes and need further examination. Third, vague statements like "in close agreement" have been avoided and replaced by "considerable differences of up to 25%" and "smaller differences of 5-10%".*

Comment (2): Second, analysis of the vertical sonde-MOZAIC results does not place the results in context of previous work. This detracts from the value of the study because no interpretation is supplied. For example, ozone comparisons of the sonde measurement with coincident ground-based and satellite measurements constitute an important part of establishing quality and reliability of the sonde data at any given station (cf Comment below on P 7102, Lines 10-18). There are no references to such studies for the stations used although literature exists on this topic, in a string of Ozone Assessments (the latest one dated 2010) and references therein, if not in more recent literature. Few or no details are given about some of the technical issues to which the authors refer, eg impact of sonde sensing solution in the ECC sonde measurement, high background currents, instrument type. In lab studies and in evaluations of sonde data at some of the stations included in the present study, there are definite biases in the sonde measurement. Examples of station bias applicable to the tropical sites used here, for example, are given in Thompson et al., 2007; 2012. However, none of the quality assurance and technical results, that are referenced in Smit et al in the papers (2007, 2011) or in Deshler et al, JGR, 2008) are applied to interpretation of the results shown. An obvious question for the reader is: why are some stations closer to MOZAIC results than others? Are these offsets what one would expect based on biases displayed by the sonde technique used at the station? That is at the heart of the

data quality assurance goal of the paper.

Response: T*o account for these deficiencies we added a new paragraph in the sonde description section (Section 2.1) of the revised manuscript, where the typical biases of the different ECC sensor types based on lab and field studies, particularly JOSIE (Smit et al., 2007) and BESOS (Deshler et al., 2008) are explained in more detail. In Section 2.1, the impact of sonde sensing solution on the different ECC sensors and the biases between same ECC sensors but processed with different background current corrections have been explained in more detail since they form the basis for discussing possible differences among different ECC launch sites (see discussion of DeBilt and Legionowo in the conclusions section of the revised manuscript). Additionally, we added, as suggested, references to paragraph (P 7102, Lines 10-18) and discuss the findings of several studies.*

In the revised manuscript, the results for the tropical stations are placed in the context of previous intercomparisons with coincident satellite measurements. The discussion of Scoresbysund and Sodankylä places our findings in context with the previous works of Smit et al. (2007) and Deshler et al. (2008). In the conclusion and summary of the main findings of our study we again place the results for the different stations in context with results from lab and field studies of Smit et al. (2007) , Deshler et al. (2008), and Smit et al. (2011).

Comment (3): Third and finally, the conclusion about "trends" is not supported by the evidence presented in the paper. Yes, interesting and valuable comparisons are given and the research methods are good, but without interpretation other than the figures given, there is not enough information to conclude that the sondes are 'ready for trends analysis,' as implied by the last sentence in the Abstract. Indeed, as mentioned below, data consistency at some stations appear to be better than at others where speculative discussion is given (useful) but no conclusions can be drawn. That message in itself is a more suitable, Abstract-worthy conclusion than what is written in the present Abstract.

Response: *We agree with the reviewer that there is not enough information to conclude that the sondes are 'ready for trends analysis'. Therefore, as explained in the response to overall comment #1 we changed the abstract accordingly.*

SPECIFIC COMMENTS:

Comment Page 7102. Line 6, after Smit et al, 2007; should add the first reference above, Thompson et al., 2007.

Response: *It has been added in the revised manuscript.*

Page 7102. Lines 10-18. The authors refer to some of the types of studies mentioned above that are needed to give perspective, ie 'space-borne, ground-based or other airborne...' but with no references or discussion of

what those reports imply for the data used in the present study. This should be amplified and clarified.

Response: *We extended this paragraph. In the revised manuscript we give references to previous studies and discuss parts of the findings that are important four our present study.*

Page 7102. The statement on line 17 that quality is most important in the UTLS is not unreasonable. However, it is somewhat misleading. The biases among techniques shown in the various JOSIE studies appear throughout the sonde profiles, including the troposphere (Smit et al and Deshler et al, references). Thus, biases *do* matter in the interpretation of the MOZAIC comparisons, that are mostly below 12 km.

Response: *This statement is removed from the paragraph. Since the paragraph was modified as explained in the comment above, we felt it was out of context.*

Page 7102. Lines 25 following. So what changed in 1998? Sonde instrument or something with the MOZAIC instruments or protocol? The reader cannot discern what changes have taken place over time and why. Pointing out that these uncertainties exist is a strength of the paper. However, the interpretation and conclusions are less definite as a result. FOR EXAMPLE - later in the paper, (page 7122, lines 15-25), the authors argue that problems with the sondes are indeed not likely to be the sole reason for discrepancies with the MOZAIC. *This is an important result of the paper and should be included in the Abstract. Is not the "conclusion" for SI2N readers that both sondes and MOZAIC need further examination before either one is 'robust' and reliable enough for trends?* Reviewer recommends highlighting this finding in Abstract and Conclusion.

Response: *We extended and modified this paragraph in the revised manuscript. Due to changes in the structure of the revised manuscript, the findings of Staufer et al. 2013 serve as the motivation for our study. Thus, we aim to find out whether the differences that Staufer et al. (2013) found for Payerne-MOZAIC are a unique feature of the Payerne time series, or, if they are also visible at other ozonesonde sites. This motivation is now highlighted in both the abstract and the introduction of this paper (see last paragraph of Section 1). The fact that both sondes and MOZAIC need further examination is highlighted in abstract as well.*

Page 7103. The second paragraph needs some correcting or clarification. Recommend deleting the sentence about 'led to confusion...'. Better wording. 'Led some groups to change their technique. (Remove word 'Additionally'). NOAA sites... [this is ok, but NOAA also changed instrument and technique at Pacific tropical stations it operates: Samoa, Hilo, Fiji, San Cristobal]. The latter stations, because they are not near MOZAIC routes, are not included in the present paper.

Lines 25-26. Recommend stating ... "After 2004 mainly ES sondes were launched but the 1.0% solution strength was retained (see meta-data at WOUDC). Recently the ozonesonde data user community

has been addressing how to account for changes in radiosonde instrumentation that have accompanied ozonesonde changes at a number of stations in the past 5 years or so [Stauffer et al., 2013*]. The radiosonde changes propagate to each ozone measurement, but mostly above 100 hPa; newer radiosondes mostly affect ozone data after 2009. Thus, these influences are neglected here."

Response: *Thanks for your suggestions and recommendations. We have included them in the revised manuscript.*

Page 7104. The first paragraph is about CF then ends with reference to background current and SOPs. ?? Is it out of place or unfinished? The idea is picked up again in P7121.

Lines 10-19. Information on background currents is given. Are background currents available to the authors? If so, how were they used? This paragraph does not relate to the rest of analysis in the way it is written. In a Table there is reference to background current 'increasing' or 'decreasing'. What does that mean for the analyses?

Response: (1) *The reference was out of place.* (2) *In the revised manuscript we extended the background current information to better inform the reader which launch site background currents are available and that its influence can be analyzed (due to the limitation of the Lagrangian technique it is only DeBilt, Uccle, Payerne and Legionowo, partially Sodankylä.). In the table the references to the respective treatment is given since we discuss their influence at DeBilt and Legionowo site (Section 3.1.2, 3.1.3 and Section 4).*

Page 7104. Like NILU, NASA Goddard has also collected campaign data that is archived in WOUDC. A number of the North American stations from which statistics are taken in this paper have additional data due to the Intensive Ozonesonde Network Study (IONS) experiments (eg http://croc.gsfc.nasa.gov/seacions). It would be appropriate to acknowledge it here by adding to line 25: "The Intensive Ozonesonde Network Study (IONS) experiments (Tarasick et al., 2010; Thompson et al., 2011) over North America have operated in 2004, 2006, 2008 and 2013. This has agumented regular launches at Boulder, Hunstville and Wallops as well as most of the Canadian stations listed in Table 1. The data are archived at NASA/Langley and WOUDC."

Response: *We agree that such a statement is appropriate. It is included in the revised manuscript.*

Pages 7105-7106. MOZAIC observations. Here is where details on quality assurance, calibration over time, etc, are lacking. More references and facts are required. How does the reader know whether or not this 'standard reference' is constant over time? Actually the authors seem to doubt that it is. (See comments above).

Response: *In the revised manuscript we do not give more details on the quality assurances since they*

remained the same since the project started and all technical details are given in Thouret et al. (1998). However, we considered that it is appropriate to explain to the reader that MOZAIC is often considered as a standard reference because of its regular calibration, but that recent analyses (Staufer et al. (2013), Logan et al. (2012)), however, cast some doubt on this long-term stability.

Page 7114 (first para). The literature concerning the ES vs SP differences has been based heavily on JOSIE and other tests since 2000. Maybe these instrument issues would not have affected changes in 1997.

Response: *OHP data are difficult to analyze regarding ES vs. SP differences because only 2 years of SP data are available during the MOZAIC period. Additionally, the large discrepancies between sondes and MOZAIC in the mid-nineties may also impact the results.*

Pages 7114-7119. Compared to the discussions of mid-latitude European sondes, these sections, although important to report, suffer from (1) fewer statistics; (2) long trajectories that imply considerable uncertainty in matches; (3) instrument changes that are not easy to interpret. Drawing conclusions from these sections can be misleading.

Recommend in the **Summary & Conclusions** that a distinction is made between stations with a lot of history and literature (mostly in Europe) and those where the analysis methods, in addition to instrument issues, only leads to uncertain conclusions.

Examples... (A) very well stated on page 7115, lines 5-10. What to make of the Boulder comparisons?

(B) page 7116, lines 10-15. The discussion of Japanese data really underscores
limitations of technique, with > 50 hr trajectories!

[C] page 7119 - about the tropical stations, ie Paramaribo. In SHADOZ analysis (eg Thompson et al., 2007) Paramaribo alone of a dozen or so sites, did not agree well with TOMS overpass nor with the co-located Brewer. A positive bias compared to all other tropical stations in the stratospheric segment of the profile was striking (see Figures at below taken from SHADOZ soundings and submitted for publication in 2011.) In 2012, as presented at an ozonesonde workshop (part of the O3DQA WMO/SI2N activity), a re-processing of Paramaribo was carried out by KNMI and total ozone offsets with satellite and spectrometer were nearly eliminated (also attached, as published).

Evidently the data used in Staufer et al (downloaded in 2010) reflect high-Paramaribo ozone archives. The authors should inquire of KNMI about these data, when and how re-processing was carried out. To the Reviewer's knowledge, details of the data change have not been published, making it difficult to draw conclusions about this very valuable dataset.

Response: *Mid-latitude European stations are, due to their location and large number of ascents and thus the high number of matches, favored by our approach. A distinction between results from European stations and non-European stations is therefore appropriate. In the revised manuscript, we discuss this issue in more*

detail in the Methodology section (Section 2.3) when we explain the uncertainties of the method. We make the reader aware of this problem before we present the results for the individual stations (paragraph before section 3.1). As suggested by the reviewer in the summary and conclusion section of the revised manuscript, a distinction is made between stations with a lot of history and literature, i.e., MOHp, Payerne, Uccle, DeBilt and Legionowo, and the non-European stations.

Ad [A]: Boulder is a very interesting station since the sensing solution strength changed twice. However, unique solutions are used and NOAA's own data processing techniques may account for these changes. Since the Boulder comparison suffers from (1) few statistics and (2) long trajectories, conclusions can be misleading but it appears that, according to our analysis, the Boulder UTLS ozone series is not affected by changes in solution strength (see Conclusions in the revised manuscript).

Ad [B]: In the revised manuscript, the limitation of the technique is discussed at the beginning of Section 3.

Ad [C]: Thanks for the information on Paramaribo. We contacted KNMI for further information. Paramaribo data were re-processed in 2012 with a constant background current and the pump flow correction of Komhyr (1986) is used instead of Komhyr et al. (1995). We extended the paragraph about Paramaribo by referring to the Thompson et al., (2012) paper and the re-processing.