Reply to comments by Anonymous Referee #2:

We are pleased that our approach to compare balloon-borne ozonesondes with MOZAIC is appreciated. Your recommendations will certainly help to improve the manuscript. Please find below our responses to your comments.

Comment (Page 7071, lines 3-6): Have there been any biases between the Arosa data and satellite data as well as satellite-satellite biases? I assume that any effects could be neglected. However, a comment would be appropriate.

Response: We agree that a comment is appropriate since we assumed that the effects are negligible. We plan to add the following sentences to the manuscript (after line 6 on page 7071):
"For the calculation of the correction factor (CF) the height difference between Payerne and Arosa is taken into account using only ozonesonde measurements above the Arosa height. Strong horizontal ozone gradients between Payerne and Arosa can occasionally occur, but are expected to cancel out in the mean values (see also Jeannet et al., 2007).

Arosa total ozone columns, together with many other ground-based Dobson stations, have recently been used by Labow et al. (2013) for comparison with reprocessed SBUV, BUV and SBUV-2 data. Typically, their time series' agree within 1% over the past 40 years (see, for example, their Figures 6 and 7). Over the last decade, the bias even approaches zero."

References:
Labow, G. J., R. D. McPeters, P. K. Bhartia, and N. Kramarova (2013), A comparison of 40 years of SBUV measurements of column ozone with data from the Dobson/Brewer network, J. Geophys. Res. Atmos., 118, 7370–7378, doi:10.1002/jgrd.50503.

Comment (P 7075, l 8): Why were matches between the same aircraft excluded?

Response: The “self-matching“ approach was applied to find the optimal match criteria and to estimate the accuracy of the methodology. The reason for excluding matches of the same aircraft was that otherwise we ended up comparing a lot of neighboring observations of the same aircraft, which does not provide information on the accuracy of the trajectories: the individual MOZAIC (1-min averaged) observations of the same aircraft are separated typically by less than 15 km (i.e. MOZAIC's horizontal resolution) and a few tens of Theta. Since the match radius is typically larger (r>=50 km and dTheta>= 0.25 K), a lot of observations were only compared to their adjacent observations (upstream and downstream). The comparison of such neighboring MOZAIC observations, however, does not provide an estimate of the accuracy of the trajectories. Hence such matches were excluded.

Comment (P 7075, l 19-20): I understand that the data base for comparisons r < 50 km and dTheta < 0.25 K is too small. But is there any good reason to believe that the real errors are higher that in the proposed optimal ranges? I recommend to expand the optimal ranges to r < 100 km and dTheta < 1 K.

Response: Thank you for this comment. The expansion of the optimal ranges to r<100 km and dTheta < 1 K will certainly increase the sample size. We chose the optimal match criteria to be in the middle of the allowed range r=50-100 km and dTheta=0.25-1K. We compared the Payerne/MOZAIC results using r<=75 km, dTheta<=0.6 K with the results using r<100 km and dTheta < 1 K to check the differences and the probable improvement. Despite the relatively small change to the error bars, it was felt that it was unnecessary to repeat the entire analysis with the larger optimal ranges since the sample size increased only to a very small extent (6%, or 2018 ascents instead of 1899).

Comment (P 7075): Another Match criteria used e.g. by Rex et al., J. Atmos. Chem., 1998 is to exclude trajectories whose cluster trajectories diverge too much. Since some kind of cluster trajectories are already calculated (see section 3.4.1) I wonder, why this criteria wasn't used? That criteria should be able to sort out additional outliers.

Response: For the Payerne-MOZAIC comparison we indeed used the trajectories (the central and the displaced trajectories as explained in section 3.4.1) to check for the loss of coherency of the trajectories. We applied several criteria to exclude trajectories whose cluster trajectories were too divergent. However, our results showed only very little improvement (in terms of reducing the uncertainty/error bars in the Payerne-MOZAIC comparison). Very tight criteria reduced the sample size dramatically while soft criteria hardly influenced the differences (Sonde-MOZAIC) and their associated uncertainties. Additionally, in our case, the weighting of the matches, which gives more weight to MOZAIC matches that are closer to the soundings, may reduce the effectiveness of such cluster criteria.

Comment (P 7079, l 24+): The differences between scaling and not scaling should be discussed some more. At least any knowledge about height dependent biases of BM sondes should be mentioned in earlier and later times. Any column scaling will be scaling with respect to the stratospheric ozone column where we have the majority of the ozone. If we have different biases in the stratosphere and in the UTLS we would expect differences as those reported.

Response: Dual flights at Payerne during the OZEX campaign (Stübi et al., 2007; see for example their Figure 8) showed that unscaled BM sondes measured about 5-8% lower ozone mixing ratios in the UTLS than ECC sondes, while above 12 km altitude, BM sondes measured approximately 12% less ozone. To compensate for the smaller total ozone column, a single correction factor (CF) is applied, which may, however, distort the shape of the profile. Since the correction depends primarily on stratospheric ozone, a certain percentage of the high BM ozone mixing ratios in the UTLS may be traced back to the different bias in the stratosphere and the UTLS. However, the differences in the deltaO3 (=2(Sonde-MOZAIC)/(Sonde+MOZAIC)) between 1994-1998 and 1998-2002 cannot be explained by this scaling, which is around 1.10 for both periods.

As recommended by the reviewer we will discuss this issue further in the paper. On page 7071, after line 3, we plan to add the following: "… *estimate the ozone column above burst altitude. Any errors in the column measurement therefore can be carried over into the whole profile, in particular since (unscaled) BM sondes tend to underestimate ozone in the stratosphere more strongly than in the UTLS (see Stübi et al, 2008, their Figure 8)."*

We also plan to modify the paragraph after line 3 on page 7080: *"... The fact that scaling of BM sondes changes the sign of the bias has also been noted by De Backer et al (1998) and Stübi et al. (2008). De Backer et al (1998) also proposed an alternative normalization procedure, which was evaluated by Lemoine and De Backer (2001) against SAGE-II data. Generally, the scaling has been introduced to correct for the low bias of the BM sondes' column, but this clearly has a strong impact on UTLS ozone measurements. BM-ECC dual flights at Payerne during the OZEX campaign (Stübi et al., 2008) showed that (unscaled) BM sondes underestimate ozone compared to ECC sondes by approximately 5-8% in the UTLS, and by 12-15% in the stratosphere (15-25 km). Since a single scaling factor, which depends primarily on the stratospheric ozone content, is applied to the whole profile, the higher bias of the BM sondes in the stratosphere is carried over into the UTLS. Thus, if scaling is applied to BM sondes, some portion of DeltaO3 certainly arises from higher biases of the BM sondes in the stratosphere. The differences in the deltaO3 between 1994-1998 and 1998-2002, however, cannot be explained by the scaling, which is around 1.10 for both periods.*

The ECC mean difference profiles between the two periods are rather similar, both showing mean deviations < 5% (ECC overestimate MOZAIC), in accordance with the JOSIE results (Smit et al., 2007). The normalization does not strongly affect the ECC performance since the correction factor CF is typically around 1.0."

Minor comments and typos

Many thanks for pointing us to typos and spelling errors, particularly in the references, which we have checked carefully. Additionally, as recommended, we add some text to the caption of Fig. 6 (P 7094) explaining the numbers corresponding to the color code.