General:

In my opinion, the paper is well written and demonstrates the advantages of having (and using) civil aircraft-based tropospheric observations for evaluating satellite data products. In the following I have major comments (eg a suggestion for performing further analyses) and minor comments (minor clarification in the text of the manuscript).

Major comments:

1: the authors calculate mean and discuss small details of the differences in the mean values (or biases). My question here is, if these differences in the baises are statistically significant. Maybe the authors could add the standard errors of the mean and mention in the discussion to what extend their obserserved differences (eg seasonal differences in the biases) are really significant. Mybe they are highly significant, because of the large number of independent observation that are compared, but this significance is not mentioned in the text.

2: Concerning specific humidity, the authors mention that the uncertainties (and differences in the bias they observe) are the larger the higher the specific humidity values. I think this is well understandable, and it might be useful to analyse also the relative uncertainty and biases of specific humidity. Maybe then other details become visible.

3: Maybe my most important comment, but at the same time a comment whose consideration would require most work: the authors analyse dependencies on the bias/quality of the satellite data with respect to the instrument/observing geometry (Figs. 2 + 3) and radiative or atmospheric conditions (Fig. 4 + 5). Given the large number of very good collocation they have, I was wondering whether the analyses on performance for different atmospheric condition could be further detailed. Personally, I think it could be interesting to investigate the satellite data performance for different categories of vertical layering. How is the performance for a well mixed vertical troposphere (relatively weak tropospheric humidity decrease with altitude, also relatively low temperature gradient) if compared to a highly stratified layering (exceptional humid boundary layer and at the same time a dry free troposphere, large temperature gradients). I think, this could give interesting insight into the data reliability; however, I also understand that the authors in this paper maybe first want to show the general advantages of using the AMDAR and WVSS-II data instead of only using the operational radiosonde data.

Minor comments:

Page 3, line 61: the authors might also think in adding other civil aircraft atmospheric observations like those from IAGOS.

Page 2, line 62: better write here AMDAR and WVSS-II, because you only mention at page 5 that you use AMDAR for both datasets.

Page 2, line 63 - page 3, line 67: please check, there seems to be repeating information.

Page 6, line 169-172: maybe mention that the IASI vertical resolution of the respective temperature and humidity product is good enough to use the IASI data without information on the vertical resolution (remote sensing averaging kernels).

Fig. 6: also related to my major comment 2: It seems that even the specific humidity relative error increases with specific humidity. At 10 g/kg, it is -1/10=-10%, and at 20 g/kg, it is -3/20=-15%. Maybe this could also be discussed in some way or the other.

Fig. 7: bias much smaller than std. What about the standard error of the mean? Is it much smaller than the std? So are these bias patterns significant? I have the same questions on significance of the bais differences for Figs. 2-5 (see my major comment 1).