Reviewer #3, Matthias Schneider

We received this reviewer's comments during the pre-review phase that determined if this paper should be fully reviewed for AMT. We addressed the reviewer's valuable comments at that stage and wrote a point-by-point response describing how we implemented those comments between the AMTD and AMT submissions. We apologize that it was not clear to the reviewer if their points had been addressed in the publicly-posted version of the manuscript. Below, we are reproducing our response to the reviewer from that stage.

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).

We thank the reviewer for the time spent on evaluating this work and determining its suitability for further review for AMT.

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.

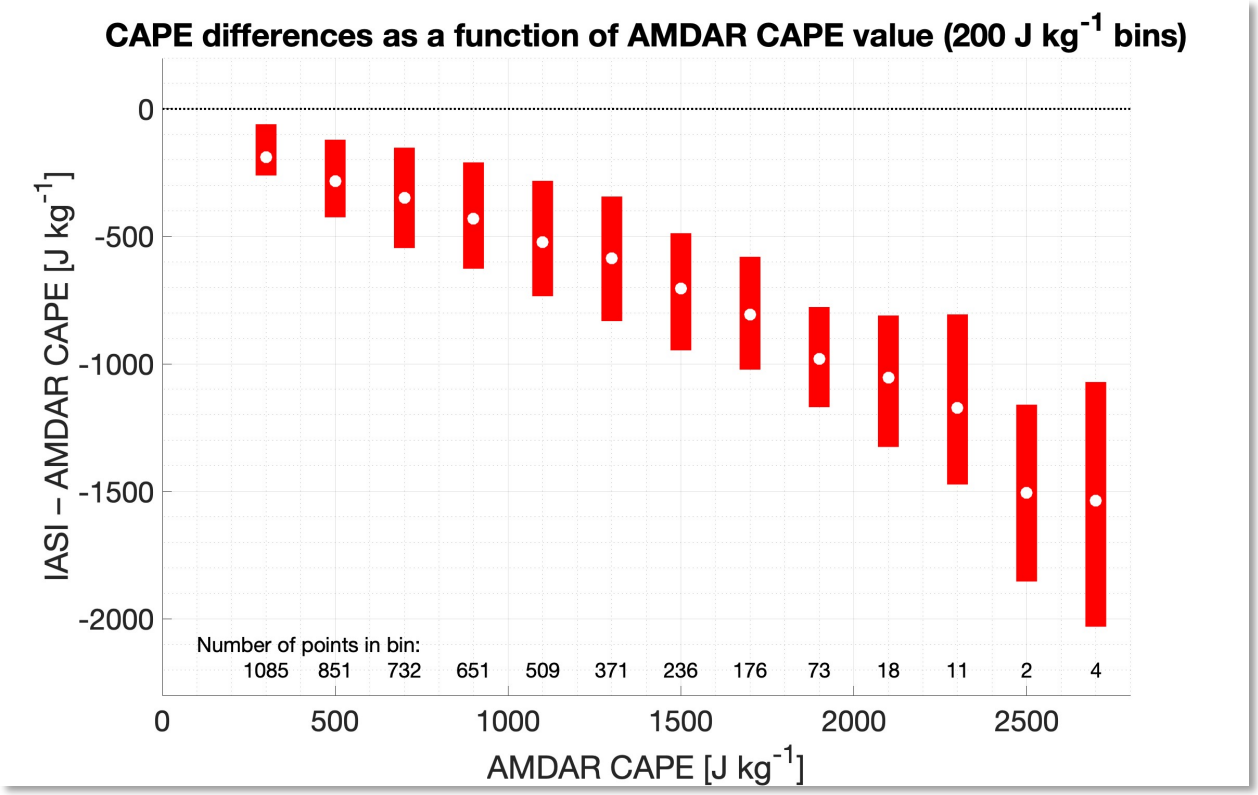
The reviewer is correct in that the large number of observations means that even the smallest differences are statistically significant. For example, Figure 2 notes the differences between the two IASI instruments. While the difference curves have almost identical shapes, there are hundreds or thousands of observations in each bin. As a result, a two sample t-test for the difference of mean indicates that the mean temperature differences that are statistically significant at the 95% confidence interval at every height, despite the biases being different by less than 0.01 K at certain heights. While the other stratifications shown here may have smaller bins, they also feature larger differences. We have added this discussion to the paper.

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.

The challenge with evaluating the relative uncertainties is that many of the observations contain very small amounts of water vapor as aircraft spend most of their time at cruising altitude where absolute water vapor content is small. By far the majority of observations have a specific humidity of 1 g/kg or less, and thus most of the time the small differences are amplified when evaluating from a relative perspective as we have to divide by numbers much less than one. Therefore, we chose to focus on absolute uncertainties.

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.

The intent of this paper is to show the general suitability of airborne observations for evaluating satellite-observed thermodynamic profiles. We share the reviewer's interest in further stratification of the data in order to evaluate the satellite performance in different observation types. In fact, we are currently conducting work that shows an increasing underestimate in convective available potential energy (CAPE) with increasing CAPE values, likely due to an increasing low level dry bias in more unstable environments. A teaser of that work is shown below. This and related analyses are beyond the scope of the current paper, the aim of which is to describe and demonstrate methodology.



Minor comments:

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

The focus of this paper is on the thermodynamic products from satellite and their validation. The IAGOS aircraft are much fewer in number and focus more on atmospheric composition, which we are not evaluating at this time. We have slightly modified the wording in this paragraph to stress that we are evaluating thermodynamic profiles. Overall, we feel that the inclusion of IAGOS brings more confusion than clarity to this discussion.

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.

Thank you for suggesting this change which increases readability. We have made it.

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

That is correct, and we have edited these sentences to omit the repetition.

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).

As this is a paper devoted to observational techniques, we feel that making claims about the specific attributes of the data for assimilation may be beyond the scope of what this paper is addressing. Regardless of the true vertical resolution of an instrument, the information content is still coming from a layer of the atmosphere instead of a specific height, and averaging kernels help ensure that the observations are properly distributed.

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.

When we are talking about relative error in this sense, we are referring to the fact that at high altitudes, the absolute values in observed water vapor are very small. Therefore even small absolute differences can manifest themselves as large relative differences when the baseline value is much less than 1 g/kg.

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).

Since the standard error of the mean is simply the standard deviation divided by the square root of the number of the observations, it goes to zero with an increasing number of observations. For the bins with a non-zero number of observations, the median number of temperature observations in a bin is approximately 2800 and some bins have well over 10^5 observations; moisture observations are roughly one order of magnitude smaller in number. As a result, SEM values for this figure are on the order of 0.05 K (0.1 g/kg) or less, much smaller than the uncertainty as represented by the standard deviation.