Response to Anonymous Referee #2

General comments

The paper by Cimini et al. presents an interesting approach to derive continuous Mixing Layer Heights (MLH) from multi-frequency, multi-angle microwave brightness temperatures that are calibrated to MLH from lidar. This technique avoids the loss of information involved in first retrieving temperature and humidity profiles and then applying radiosonde diagnostics.
In summary, the manuscript is well written and presents a novel concept for possible application to several observations. A weakness of the concept is its pure statistical calibration and the paper lacks information on the physical processes behind. I this respect I have three main issues which need to be taken into account before final publication:
We are grateful for the positive feedback and the constructive comments below. We revised the manuscript accordingly. Our replies are shown in red hereafter, while modifications to the text are highlighted in yellow within the revised manuscript.

Main issues

1) I am strongly missing some physical explanation what information from the set of microwave brightness temperatures (Tb) contributes most strongly to the MLH estimate.
Such understanding is not only important because of scientific curiosity but also to further improve the method. Because a simple regression algorithm is used to retrieve MLH from Tb with lidar MLH as "truth" this should be quite simple to identify by systematically reducing the observation vector y. For example it is rather interesting to know if the K-band (humidity information) improves the MLH estimation.
Other questions are: What is the contribution of the low elevation angles – particular of the transparent ones where inhomogeneity can cause problems? What is the effect of cloud sensitive channels? What information causes the extrapolation of MLH to lower altitudes (cf. page 4982, line 21)?
We agree with the reviewer that little discussion was provided concerning the physical explanation behind our approach. The physical basis of the proposed approach is that the observed Tb carry mixed information on temperature, humidity, and virtual potential temperature profiles. Features in the vertical gradient of these profiles may be associated to MLH. Thus, the information content of Tb is exploited in a statistical sense to estimate the scalar variable MLH.
Initially we had performed sensitivity tests and had chosen the combination of frequencies/angles that provided the best rms difference with respect to our reference truth. Therefore, we are prone to say that overall the observations at low elevation angles and K-band channels do provide useful information, despite the problems related to inhomogeneity.
To answer the reviewer's questions, we report the most significant results of our initial sensitivity analysis as follows: Comparing to the adopted configuration (14 channels and 6 elevation angles), other configurations have shown to increase the RMS error and decrease the correlation coefficient (CC) with respect to the lidar reference. In particular, we found that excluding K-band (up to 30 GHz) channels leads to a 7% RMS increase and 3% CC decrease. Cloud sensitive channels (up to 53 GHz) contribute by 13% in RMS and 5% in CC. Off zenith angles contribute by 12% in RMS and 4% in CC, though low elevations angles contribute only marginally (~1%). Thus, low elevation angles (<19°) could be avoided without a significant loss of information to make the method more robust against atmospheric inhomogeneity.
These information have been added to Section 1 and 3 of the revised manuscript.
Concerning the last question ("What information causes the extrapolation of MLH to lower altitudes?"), we admit the sentence in Section 4 was not clear. **We modified it as follows**: "This demonstrates that the proposed algorithm, establishing a linear relationship between Tb and MLH, is able to go beyond the

information provided by the reference lidar estimates to reach values below the lidar overlap limit. This feature is probably due to the relative high information content of MWR data on temperature profile in the lower few hundred meters."

In particular I have my doubts about the statistical significance of training the regression for each month and worry that overfitting takes place. Waht is the physical reason behind selecting months (not seasons)? For an operational application anyway an algorithm based on prior data needs to be available. Therefore a robust algorithm is necessary and requires an understanding which information contributes to the retrieval result (see above).
Based on understanding a further separation by phenomena (like day/night or cloud/no cloud) could make sense but a separation in respect to months does not make sense to me.
The author need to give information on how strongly do the results change when all months are used at the same time?
Monthly or seasonal training is quite common for inversion techniques based on MWR data (Westwater et al. 1993). The physical reason consists in choosing an *a priori* set close to the conditions under analysis, which helps the problem linearization. For example, the operational Atmospheric Radiation Measurement (ARM) MWR retrievals change the set of coefficients every month. Using an unique training from all months (i.e. semiannual), the performances degrade to MD=-23.7 m, STD=668.5 m, RMS=668.9 m and CC=0.54.
These information have been added to Section 4 and Table 2.
As described in Section 3, training is already performed separately for night- and day-time. More information on day/night as well as cloud/no cloud separation are given in the response to another comment below.
Finally, we agree with the reviewer that for operational application prior data extending for a full year need to be available. This is part of ongoing work.

2) The manuscript does not mention the effect of clouds – in particular boundary layer clouds. A large fraction of all data presented is certainly affected by that. For the STRAT2-D algorithm I assume that MLH height is coupled to cloud base? What is the effect for the microwave radiometer (see above)? What happens in broken cloud situations that are characteristic for well developed boundary layers? What effect has the averaging interval on that?
We agree with the reviewer that the effect of clouds was completely missing in the original manuscript.

Concerning STRAT-2D, the reviewer is correct. STRAT-2D determines the locations of three key aerosol gradients and the cloud base height with a temporal resolution of 10 minutes. Then, it selects one of these layers using the newly developed attribution technique (Pal et al. 2013) and attributes its height to the MLH.
We mention this issue in the revised Section 2.3.

Concerning the cloud effect on MWR, this is definitely an interesting point and it surely deserves discussion. Following the reviewer's suggestions, we have divided our data set into clear and cloudy periods. We used the liquid water path (LWP) measured by the MWR, assuming clear sky if the retrieved LWP is less than 20 g/m2. This value is taken as a typical MWR uncertainty for LWP estimates (Westwater, 1993). As one may expect, the performances for MLH estimates turn out to be better in clear sky than in cloudy conditions (by ~30% in RMS and 24% in CC).
We also tried to use a different approach, making a clear/cloudy (instead of day/night) separation of the training set. With respect to the results in Table 2, the performances degrades only slightly (3% RMS increase and 4% CC decrease). This gives credit to the review's impression that a cloud/no cloud separation may be as meaningful as the day/night separation we adopted.
Finally, broken clouds are present in one of the time series depicted in Figure 4 (from Julian day 201.5 to 202, i.e. July 20th 12:00 to 24:00 UTC). The calculations above were made on the 1-hour averaged data as well as on the original 10-min resolution data sets, with no significant difference. Thus, we assume the averaging interval

has an overall marginal effect up to 1-hour. A deeper analysis on broken cloud situations would be beyond the scope of this paper.
We have added the above information in the revised Section 4 and 5.

3) The authors should comment on the value of MLH retrievals during night. Some publications argue that "..these methods are most applicable to the daytime or convective boundary layer and not the night-time or stable boundary layer. At night, surface diagnostic methods are a good proxy for the depth of the stable boundary layer." (Schmid and Niyogi, 2012). Being provocative one could also say that based on the presented results (Fig. 7) the nocturnal MLH is just 200 m – an estimate which fits into the uncertainty range of all the different estimates. Is there at least one convincing example that MLH retrieved from Tb provide important information for applications like air quality which can not be seen by lidar due to the overlap effect?
We agree that surface diagnostic may be used as a proxy for the depth of the stable boundary layer. The sentence quoted by the reviewer refers to previous work (Vickers and Mahrt, 2004; Steeneveld et al., 2007). Vickers and Mahrt (2004) compare height of stable boundary layer estimated from tower (up to 60 m) and aircraft (up to 150 m) measurements with predictions based on surface measurements. They conclude that "the existing formulations perform poorly". Steeneveld et al. (2007) found that the existing diagnostic equations underestimate the stable boundary layer height and proposed a new formulation which appears to reduce this effect for very shallow boundary layer.
We have added some of the information above to Section 1.
However, our aim is to propose an approach based exclusively on MWR data, which could be used during both stable and unstable conditions.

A convincing example of the application to air quality is given in the D panel (lower right) of Figure 4; for the case at Julian day 293 (i.e. October 20 00:00 UTC) the MWR-based MLH estimate is lower than lidar estimate and it is also evidently closer to the radiosonde Rbn estimate. In fact, MWR and Rbn estimates are below 100 m, while lidar and PTU are above 300-350 m. This difference is important for air quality purposes as the more shallow is the MLH, the higher is the threat of dispersed pollutants. In such a case, assuming a plume generated by a 100m tall chimney, the different MLH values may lead either to lofting or fanning of the plume (Stull, 1988).
These information have been added to Section 4.

Additional points

Abstract:

For people who only read the abstract the location of the study (Sirta or "a typical mid-latitude site") needs to be mentioned.
Agreed.

„The proposed method provides results that are more consistent with radiosonde estimates than MLH estimates from MWR retrieved profiles. – where is it shown?
In Section 5 of the original manuscript (page 15, line 25), we write "Estimates from MWR Tb and lidar agree within 200 m with radiosondes Rbn, while estimates from MWR profiles are consistently lower by 150 m. Thus, when compared to MLH estimated from MWR profiles, the proposed method provide results that are more consistent with radiosonde Rbn estimates during both day and night."
These considerations are deducted from Figure 6 and 7; more details are given at the last paragraph of Section 4 (from page 13, line 26 on).

2.1 Radiosonde data:

- The authors provide the number of radiosondes available but should also give here – in the beginning – the time interval used in this study.
Agreed.

- The threshold for the bulk Richardson number is taken from the "grey literature" with different values for day and night. The literature is rich in different numbers and I wonder why the authors didn't choose a more common one like 0.25 used also with ERA-Interim (van Engelen and Texeira, Journal of Climate, 2013). The authors should give at least an estimate on the sensitivity of the threshold. An problem for radiosonde determined MLH is the vertical resolution as several older studies use only low resolution profiles. In fact Schmid & Niyogi (2012) exploit variability of high resolution profiles for a new method to derive planetary boundary layer.
We agree with the reviewer that open literature should be preferred. We have changed the text accordingly as follows.
According to the literature, threshold for the bulk Richardson number is generally set between 0.10 and 0.40 (Sørensen et al, 1996). Here we set it to 0.22 or 0.33 for day and night radiosondes, respectively for unstable (Vogelenzang and Holtslag, 1996) and stable (Wetzel, 1982) conditions. However, it shall be noted that the value of the threshold has modest impact on the estimated MLH. In fact, our results change only slightly (27 m mean difference) if we set the threshold to 0.21 (Menut et al., 1999) as adopted by Haeffelin et al. (2012).
These information have been added to Section 2.1.
Finally, unfortunately we were not able to find the paper mentioned by the reviewer: "van Engelen and Texeira, Journal of Climate, 2013". Please, provide full reference if all possible. In any case, the mean difference with respect to our results is even smaller (5 m) if we set the threshold to 0.25 as proposed in that paper.

2.2 MWR data:

- The authors should specify Tb uncertainty – I know it is tricky – but at least they need to provide the numbers, which resemble the measurement error (Eq. 1).
We agree with the reviewer that information on Tb uncertainty was missing. Typical Tb noise level is within 0.5 K (Rose et al., 2005). However, systematic differences with respect to radiosonde-based simulations may account for several degrees (Löhnert and Maier, 2012). Since our regression approach is trained with actual measurements, Tb systematic and random errors are inherently accounted into the process, contributing to the overall performances given in Table 2.
We have added the above information to Sections 2.2 and 3.

- spelling Lönhert = Löhnert
Agreed. Thanks for spotting this typo.

2.3 Lidar data

This section should mention temporal and vertical resolution and the overlap height
We agree that the resolution issue should be clearly discussed in the manuscript. The STRAT-2D layers are determined with a temporal resolution of 10 minutes. While doing this, it uses the raw lidar data obtained with a temporal and spatial resolution of 30 s and 15 m, respectively. The full overlap of the lidar transceiver is reached at a height of around 200 m (e.g. Royer et al. 2011).
We have added information on the lidar resolution in Section 2.3.

- First the authors provide the comparison of the "original" STRAT2-D algorithm with 53 radio soundings (Tab. 1) while later an improved STRAT2-D algorithm (after Pal et al., 2013) is used. It would be good to know if the improved algorithm improves the rather poor skill during night in Tab. 1. Or is this a feature of the difficult estimation of MLH from soundings?
We agree with the reviewer. We modified Table 1 to present the results of the improved STRAT-2D algorithm (from Pal et al., 2013) because our study uses MLH Lidar retrievals based on the Pal et al. (2013) method. The results of Haeffelin et al. (2012) are removed since they are not relevant here.

- I am surprised that the morning transition is just taken at 11 UTC. In my opinion one strength of lidar observations is that the raise in MLH from sunrise onward can be observed rather well. From the statistics presented in the paper I can not infer how well this is done by MWR. As Fig. 2 illustrates the transition zone (several hours) is most demanding and I wonder how results change if this times are excluded in the analysis.
We agree with the reviewer that lidar observations can observe rather well the morning transition. As discussed in Section 3, the training of the proposed MWR-based method is performed separately for night- and day-time. In order to separate regimes of nocturnal and daytime boundary layer (thus, night- and day-time periods), we choose the times for monthly mean morning and evening transitions, as estimated by the diurnal cycle of the stability index based on the near-surface micro-meteorological measurements of Obhukov length scale (see Pal et al. 2013).
Time series in Figure 4 show few diurnal cycles in different weather conditions, where the MWR estimates seem able to follow the lidar estimates during the morning and evening transitions.
We have changed the text in Section 3 to make this more clear.

3. Methodology

- p4979 line 23: "accepts non-unique solutions" sounds strange. Maybe "leads to" ?
Agreed.

- p4980 line 7: "consists of .. state vector includes MLH.."
Agreed.

- Here the aouthors mention 10 min bins while later only hourly values are given.
As discussed in Section 3 (page 10, line 19), temporal co-location is obtained by averaging the state and observation vectors in 10-minute bins. Data at 10-minute resolution are reported in Figures 1-4.
Hourly averages are computed to reduce the effects of temporal and spatial collocation of the MWR and lidar instruments and to match the radiosonde fly-time. These data are reported in Figure 5 and Figures 6-7, respectively.

- The algorithm is trained separately for day and night. Why don't you show the results also separately in Tab. 2. - especially when considering the large differences in Tab. 1.
Alternatively, you could just color the dots for nighttime in blue in Fig. 5 and give the separate results in the lower right corner. They hopefully all lie in the lower left corner.
Agreed. **We have modified Figure 5 to differentiate day and night estimates**. As correctly guessed by the reviewer, night time estimates mainly lie in the lower left corner.

4. Results

p4983,line 25: MLH itself does not depend on climatology.
Agreed. "Climatology" has been replaced by "mesoscale and synoptic forcing".

p4984,line 12: The authors might want to mention that RS estimates are only representative for one “slanted” profile that might be by pure chance related to an especially strong eddy while the lidar estimates are for temporal averages.
We agree with the reviewer. To make it more general, we state that part of the differences in Figure 6 shall be attributed to different temporal and spatial sampling of radiosonde and remote sensing instruments.

5. Summary

p4985, l23: consistently
Agreed. Thanks for spotting this typo.

- I find it a bit dangerous to talk about seasonal statistics as the considered time interval covers just a few months which might strongly be affected by certain weather types.
We agree with the reviewer that the dataset available to us is not climatologically significant as it covers less than one full year. **We have added this information in Section 5**.

Fig. 5: Please enlarge the letters.
Agreed. Figure 5 has been modified to enlarge fonts.