Reviewer 2

Interactive comment on "Atmospheric boundary layer dynamics from balloon soundings worldwide: CLASS4GL v1.0" by Hendrik Wouters et al.
The paper presents a significant advancement in producing both useful accessible boundary layer data from radiosondes, and a nice marriage with a simple ABL model to produce continuous ABL data constrained by analyses, with open-source software.

We would like to thank the referee for providing their review of the manuscript, and we are very glad regarding the appreciation of the software's potential. We also appreciate the comments, especially the suggestions related to possible sources of model biases. We provide a point-by-point answer below. The changes to the manuscript are provided as quoted text, which will be included in the next revised version of the manuscript.

I was not able to fully run the CLASS4GL software myself. On a Mac using MacPorts, the PyYAML was not available and I downloaded directly from the website - there were issues recognizing the CLoader option - apparently a version inconsistency. But I will follow through as I would like to use this tool.

Thank you for testing software! The goal is to have a platform-independent software, so we strive to make it work on all platforms including Mac systems. As a solution on Mac, we would like to suggest to try either a Python environment with anaconda (as explained on https://class4gl.eu/?page_id=105) or Pycharm+homebrew. In case of pycharm+homebrew, these are the needed steps to install the CLoader module:

```
brew install libyaml-dev
pip install pyyaml
```

Please note that '`brew install libyaml`' (so without '`-dev`') will not work. A similar solution may exist in case of your current Python environment using MacPorts. The CLoader is required to read yaml files 10–100 times faster, but depends on modules written in the C language.

Regarding the manuscript, I suggest only minor changes are needed (editorial and regarding content), as outlined below:

P3 L14: Use "automates" instead of "automises". Likewise on P4 L7.
P4 L3: Change "dirunal" to "diurnal"

Thanks for identifying the typos. They will be corrected in the revised manuscript.

P5, L10-11: It is a common assumption that the heat, moisture and momentum content of the ABL are perfectly mixed, but of course there will be mean vertical gradients, especially near the entrainment zone and the surface. In other words, the gradients here are a little weaker than for a well-mixed ABL, which may be compensated by other parameter choices. What would be the effect of specifying more realistic but still simple tails (e.g., exponential or even linear) of theta, q *and V at the top and bottom of the ABL? This will relate to comments below regarding apparent biases.*

It is true that the ABL model considers a perfectly-mixed ABL with values of potential temperature, specific humidity and wind speed that are constant throughout the ABL, whereas the entrainment zone is represented as a jump between the ABL values of and the free atmosphere values, and the surface layer as a analytic profile between ABL values and the surface values. Other gradients within of the ABL are not explicitly represented. The Monin-Obukhov similarity theory is employed for calculating analytic surface layer profiles and the gradient transport in surface layer in an implicit way as a replacement for a more explicit representation. For the entrainment zone, the heat entrainment ratio (β) of 0.2 (the ratio of heat entrainment to heating through the surface layer) is considered and the additional entrainment by wind shear, based on observations and large eddy simulations (Vilà-Guerau de Arellano et al., 2015). More realistic tails at the top and bottom of the ABL are not yet considered in the ABL model, hence, it is not possible to quantify their effect and the possible associated biases in the model. Therefore, one would require a dedicated study for which one needs substantial changes to the ABL model formulations. We are aware about the model limitations and associated uncertainties, and about the need for more research employing more realistic profiles. This will be mentioned more explicitly in the revised manuscript as follows:

"The use of the mixed-layer equations implies that the turbulence inside the ABL is not explicitly solved, and assumes that the potential temperature (θ), specific humidity (q) and wind components are homogeneous within the ABL. This assumption tends to be supported by the efficient turbulent mixing under convective conditions (Bauer, 1908). At the top of the ABL, the entrainment of heat and moisture is parameterized by a jump of θ, q and wind components over an infinitesimally small height, which are initialized with a constant lapse rate with height in the overlying free atmosphere. Entrainment flux is calculated as a fixed fraction (0.2) of the buoyancy flux, for which one also adds the entrainment flux driven by shear. An important feature of the model is the possibility to represent the subsidence coupled to the entrainment process at the inversion zone (Vilà-Guerau de Arellano et al., 2015). The surface–atmosphere turbulent exchanges for momentum, heat, and moisture **in the surface layer** are calculated considering their aerodynamic resistances. These are calculated in an iterative way assuming constant values for aerodynamic roughness lengths, while applying correction factors for non-neutral stratification of the atmospheric surface-layer (Paulson, 1970) according to the Monin–Obukhov similarity theory (Monin and Obukhov, 1954). **It should be kept in mind that more realistic profiles with explicit ABL gradients for temperature, humidity and wind speed – especially at the top (entrainment zone) and bottom (surface layer) of the ABL – are not yet considered by the model. In order to tackle these limitations and associated uncertainties, more research is needed employing more realistic profiles.**"

P7 L4-10: Please state how many (or what percentage) of the 42,000 profiles are excluded for each reason (lacking both 00 and 12UTC soundings vs. non well-mixed profiles? The first seems a hard criterion, but exactly how well-mixed is that criterion and what if it is relaxed?

The criterion for a well-mixed profile is that the root mean square error of the profile measurements in the boundary layer is lower than 1.5°C. This information will be added to the manuscript. In addition, we will provide the statistics on the reasons of profile retainment for each filtering step. Therefore, paragraph 2.2 will be revised and it will read in the revised manuscript as follows (additional information is indicated in bold):

“2.2 Automated balloon data mining

Global data of weather balloon soundings are taken from the Integrated Global Radiosonde Archive (IGRA; Durre et al., 2006) which is maintained under the auspices of the National Oceanic and Atmospheric Administration (NOAA). The IGRA archive is routinely updated and currently includes more than 2700 stations covering major global climate regions. The CLASS4GL sounding database is additionally supplemented with data from intensive radiosonde campaigns from HUMPPA (Williams et al., 2011), BLLAST (Pietersen et al., 2015) and GOAMAZON (Martin et al., 2016) – see Tab. 1 and Fig. 2. Other sources of vertical profile data (from e.g., aircraft, satellites, other observation campaigns or long-term operational soundings) may be considered in future applications of the framework. As described above, CLASS requires morning sounding profiles for initialization and afternoon profiles for validation to enable a mechanistic interpretation of the diurnal ABL evolution.
[new paragraph]
All balloon sounding profiles **(~15 million profiles)** are pre-processed first **by calculating the bulk mixed-layer properties:** ~~An estimation of ABL properties is obtained for the selected profile pairs.~~ ~~First, t~~The mixed-layer height (h) is assessed as the height at which the Bulk Richardson number (RiB) exceeds a critical value (RiBc). We adopt the estimates for RiBc provided by Zhang et al. (2014) : RiBc = 0.24 for strongly stable boundary layers, RiBc = 0.31 for weakly stable boundary layers, and RiBc = 0.39 for unstable boundary layers. The uncertainty range of h (used below) is determined from its interval corresponding to the RiBc range [0.24, 0.39], for which the interval is further extended to the nearest sounding records above and below. Second, the mixed-layer potential temperature (θ), specific humidity (q), zonal wind (u) and meridional wind (v) are calculated as their average values recorded within the mixed-layer. The capping inversion is estimated by a linear extrapolation of the two lowest sounding measurements above h, for which its lapse rate for potential temperature ($\gamma\ \theta = d\theta/dz$), specific humidity ($\gamma\ q = dq/dz$) and wind components ($\gamma\ u = du/dz$ and $\gamma\ v = dv/dz$) are calculated. The jump values at the h for potential temperature ($\Delta\theta$), specific humidity (Δq) and wind components (Δu and Δu) are estimated from the difference between the values of the capping inversion at h and the values within the mixed-layer.

Afterwards, morning--afternoon profiles are selected that meet a series of selection criteria: the morning profiles, ie. profiles before 12 h local time, are selected first and they amount to ~ 6 million profiles. Here, the selection of suitable morning soundings **(and the subsequent afternoon soundings after 12 h)** ~~balloon sounding (morning–afternoon) pairs~~ is ~~largely~~ based on the timing of these soundings (a): Morning **(**and afternoon**)** sounding profiles ideally should be acquired after sunrise and before sunset, respectively. However, routine sounding launches happen synchronously on a daily basis at 0 h and 12 h UTC, whereas launches at intermediate timings (3 h, 6 h, 9 h, 15 h and 18 h UTC) are rare. As a result, many launches, especially those at 0h UTC in Europe and Africa, often happen several hours before sunrise. Since the net exchanges near the surface for heat, moisture and radiation are generally low at the end of the night, the atmospheric profiles tend not to change dramatically before sunrise (unless the synoptic situation changes), being often representative for the time the convective ABL starts to emerge (van Stratum and Stevens, 2018). As such, in order to maintain a high number of soundings in our analyses, launch times within 3 h prior to sunrise are still allowed here. For these soundings, the ABL simulation starts at sunrise, assuming that the change in the atmospheric profile since the balloon launch time is negligible. Furthermore, **(b) only those soundings are retained with more than seven measurements in the vertical below 3000 m (72% of the morning soundings), (c) for which the uncertainty of the mixed-layer height is lower than 150 m (26% of the morning soundings), (d) for which the ABL is sufficiently well-mixed (ie. for which the root-mean square deviation of the temperature from the estimated mixed-layer average is lower than 1.5°C; 92% of the morning soundings). We also (e) set the morning lower temperature limit to 278 K in order to minimize the chance of freezing temperatures during the course of the simulations (70% of the morning soundings). The next criteria is that (f) an afternoon sounding can be found with the same criteria as the morning sounding except regarding the uncertainty of the mixed-layer height (which is the case for 24% of the filtered morning soundings). Here, the afternoon radiosonde profile** on the same day needs to occur between local noon and 1 h before sunset, and at least 4 h after the morning sounding for allowing a sufficiently large timespan of the model simulations ~~model initialization in the morning~~**. We also require that (g) all ground parameters are available (8.7% of the filtered sounding pairs). The above criteria lead to 21,826 profile pairs from 134 stations**. Finally, the current version of CLASS is only capable of representing growing mixed layers. Therefore,an observed mixed-layer growth of 40 m h^{-1} is considered as a lower limit (an last filtering of 85% of the profile pairs), **which leads to 18385 profile pairs from 121 stations**. It should be noted that these criteria can be flexible and may be reconsidered according to the intended application, since there is an obvious trade-off between sounding quality and amount of data being retained."

P7 L12: Change "says" to "days".

Done

P7 L13-17: Are there clear discrepancies between the behavior and/or statistics of gap-filled (model) versus observationally driven results? I assume you have looked at this - a caveat might be warranted here.

At this stage, we have only implemented the possibility of extracting profiles from continuous/gap-filled datasets (reanalysis, satellite-based products, and Earth system models...). The discrepancies between *gap-filled versus observationally driven ABL model results have not been done yet. We agree that this is an important caveat, so the user should be warned that such a validation is needed as soon as a gap-filled dataset is employed. Hence, the following text will be added to the revised manuscript:*

*"*This alternative to the use of sounding data holds great promise for spatially-explicit climatological ABL studies and multi-annual trend assessments. **It should be noted that ABL model simulations using continuous/gap-filled datasets may deviate from those using**

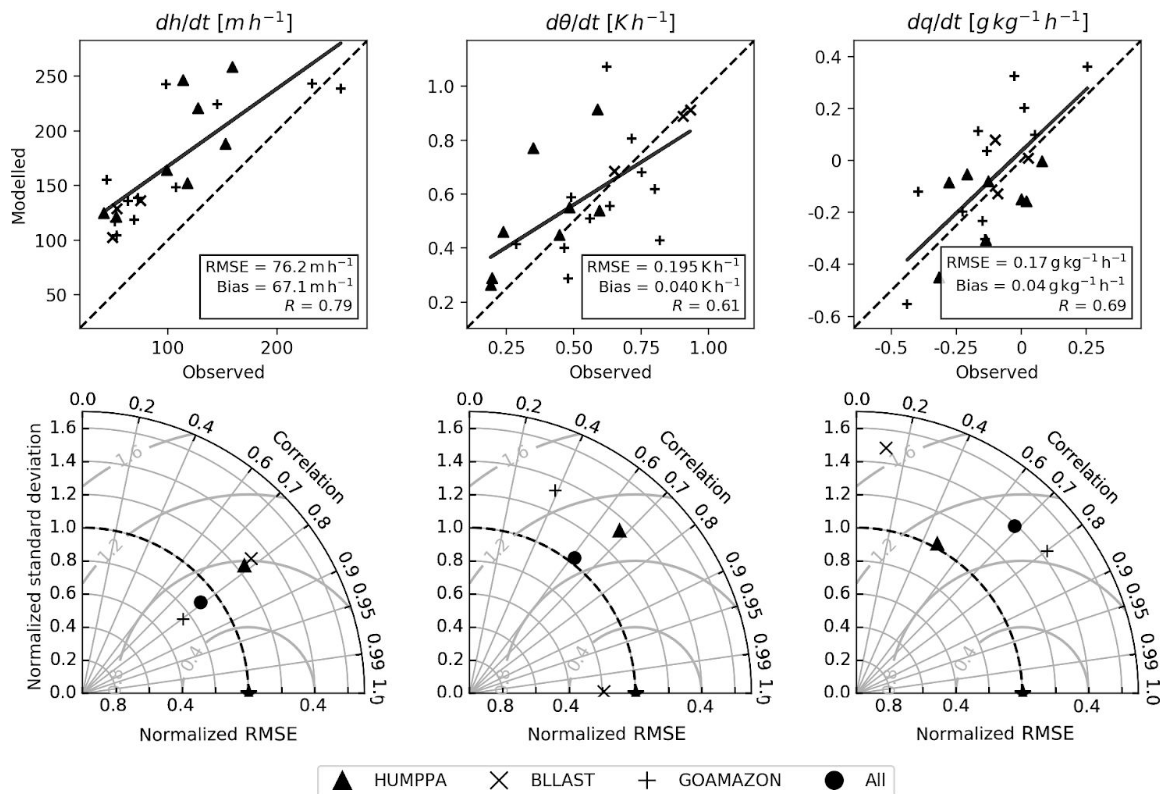
the observations. Hence, an additional validation is needed as soon as such datasets are employed, in which one should compare the gap-filled datasets and the observations and the error propagation on the ABL model simulations. Such an in-depth evaluation against the available sounding pairs can be done using the ~~default~~ presented framework~~ based on the balloon sounding data~~."

P10 L8: Change "reassure" to "assure".

Done.

Figure 3: There is only a circle (All) for dq/dt - not the other rates. Is something missing?

They were indeed missing. All symbols will appear in the revised manuscript, as shown in the figure below:



P12 L6: Here I start wondering about the sources of biases and if you have been able to examine them. For dq/dt, a positive evaporation bias, excessive low-level moisture flux convergence (in the boundary conditions) or too little entrainment of dry air could each explain this. Has it been investigated? Is it likely a problem with the model or forcing data?

We also expect that the bias have multiple origins, so both the ABL model (physical concepts) and its forcing data (convergence/advection, evaporation bias, cloud cover...) but also model tuning parameters (eg., entrainment ratio) and errors in the sounding observations used to initialize and validate the model, and all of these possible errors should be investigated in the further development of CLASS4GL. We will make this clear by adding the following text in the revised manuscript as follows:

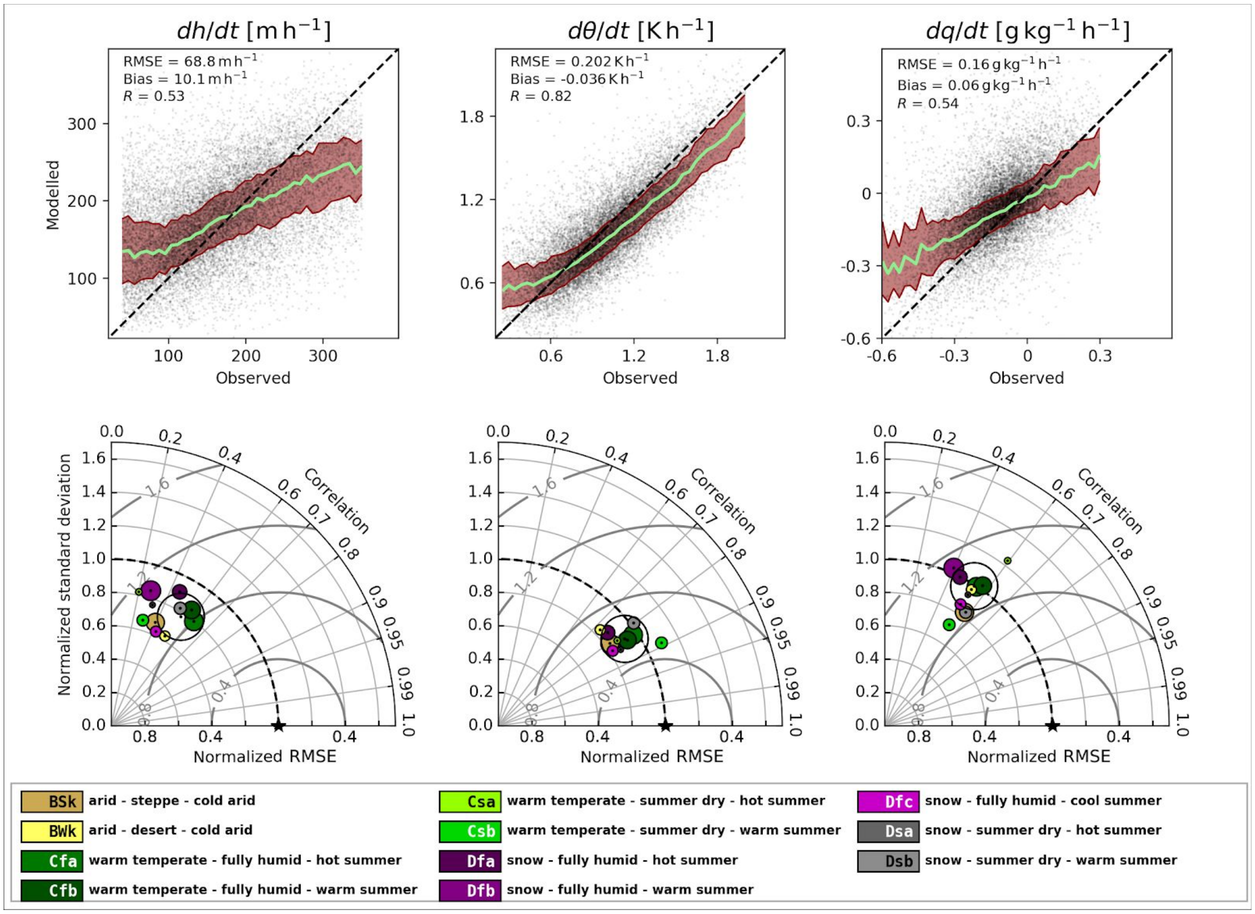
“**The bias is expected to come from multiple origins, including the ABL model and its physical concepts, the forcing data (convergence/advection, evaporation bias, cloud cover…, see Table 1), model tuning parameters (such as the entrainment ratio, see Table 2) and errors in the sounding observations used to initialize and validate the model. All these possible error sources should be investigated in further development of CLASS4GL**”

The sources of biases from forcing data is supported by improved model statistics when using 1-hourly values (ERA5) instead of daily values (GLEAM) for evaporative fraction for the campaigns (see figure above) and the global results (see figure below, which is discussed further below). Particularly, the bias of dq/dt for the campaigns is now 0.04 g kg^{-1} h^{-1} (see figure above), whereas it was much higher in the previous results 0.17 g kg^{-1} h^{-1}. The new results will also be included in the revised manuscript. Using the updated results as shown with the figures above, we have modified the text accordingly, eg.,:

“However, we could identify common model limitations over the three campaigns, see Fig. 3. ~~,~~ ~~particularly an overall positive bias in dq/dt, and a too low variability in the tendencies~~. **This includes an overall (slight) positive bias in dh/dt (dθ/dt and dq/dt), and a under(over)estimation of its (their) variability** as indicated with **a** normalized standard deviation **different from** ~~of less than~~ 1 in the Taylor plots (Fig. 3)**.“**

Fig 4 and associated text: If the heating and moistening rates are converted to J/kg/h by multiplying by C_p and lambda_v respectively, we get that the heating bias is -52 J/kg/h but the positive moistening bias is 175 J/kg/h.

Using the 1-hourly ERA5 data as forcing for evaporative fraction as discussed above, we now get an overall better model performance for the IGRA global soundings (see figure below): particularly, the global negative bias in dβ/dt has been reduced from -0.052K/h (-52 J kg^{-1} h^{-1}*)* to -0.036K h^{-1} (-36 J kg^{-1} h^{-1}), and the positive bias in dq/dt has been reduced from 0.07 g kg^{-1} h^{-1} (or 175 J kg^{-1} h^{-1}) to 0.06 g kg^{-1} h^{-1} (or 150 J kg^{-1} h^{-1}). Pearson correlation coefficients have slightly improved (0.54/0.79/0.52 -> 0.53/0.82/0.54), and all RMSEs are slightly lower as well (72.3 m h^{-1} / 0.211 K h^{-1} / 0.17 g kg^{-1} h^{-1} -> 68.8m h^{-1} / 0.202 K h^{-1} / 0.16 g kg^{-1} h^{-1}).



These results will be updated in the revised manuscript, and the text and discussions will be modified accordingly.

The discrepancy is 123 J/kg/h – ...

With the updated results mentioned above, the discrepancy has now been slightly reduced from 123 J kg^{-1} h^{-1} to 114 J kg^{-1} h^{-1}.

… again there could be multiple sources of this. First thought is net radiation, but excessive ground heat flux from the soil, advection (convergence) or entrainment could all be reasons. Any idea about the source of this net energy bias?

P13 L25-29: Related to above, a nice speculation on causes, but atmospheric models including reanalyses tend to have too much surface net radiation due to cloud errors and lack or proper representation of aerosol effects. R_Net or the input ERA-I radiation should be validated against independent data (e.g., the available CERES data) as a sanity check.

In line with the discussion above, we expect that the source of the net energy bias have multiple origins, including the ABL model and its physical concepts, the forcing data, model tuning parameters and errors in the sounding observations. We agree with the suggestions of the referee that this could be especially due to biases in the net radiation (which is calculated by the model by prescribing the cloud cover), the ground heat flux, the entrainment rates, and/or the prescribed advection. All this will be discussed in the revised manuscript as follows:

"In addition, the overall modelled range in dh/dt, dθ/dt and dq/dt agrees well with the observed range, with departures from the standard deviation of the observations below 22% – see Taylor plots in Fig. 4. There is also a systematic underestimation of the variability for dh/dt, dθ/dt, but not for dq/dt. **The negative bias in the temperature tendency and the positive bias in the humidity tendency leads to an overall net heat bias of 114 J kg^{-1} h^{-1}. Similar as for the results in the campaigns, it is expected that such global biases have multiple origins, including biases in the net radiation (which is calculated by the model by prescribing the cloud cover), underestimation of ground heat storage to the soil, the entrainment rates, and/or the prescribed advection*. Further research should investigate possible errors related to input datasets and validate them against independent data (e.g., the available CERES data could be used to evaluate the net radiation).*** "