Authors' Reply

The authors would like to thank the referees and the Executive Editor for their constructive feedback, that helped in improving the manuscript. In the following, all revision comments are addressed and the resulting edits are included in the following way: The comments are repeated and responses are given below. Changes made in the manuscript are indicated in blue. Figure numbers with "R" correspond to figures in this reply. In the attached manuscript, red and blue indicate removed and added text, respectively.

Reply to Anonymous Referee #1

Comment: This manuscript uses simultaneous observations from a Ka-band radar and a visible/near-IR lidar aboard a high-flying aircraft to evaluate the distribution of cloud- top and cloud-base heights in shallow cumulus in two simulations with varying grid resolution. The study is focused on winter-time conditions near Barbados, site of the NARVAL experiment, where shallow (warm) clouds and precipitation dominate. After discussing how cloud tops may be determined from lidar and radar observations, and how cloud base might be identified in radar returns, the authors describe how synthetic observations are constructed from regional simulations with the Icon model at "storm- resolving" kilometer-scale horizontal resolution and "large-eddy model" at 300 meter resolution and a more complicated treatment of microphysics. The observations show a bi-modal distribution of cloud top heights in the more sensitive lidar observations, one corresponding to well-developed shallow convection and the other attributed to very shallow clouds (as diagnosed by the small difference between the lifting condensation level and cloud-top height) which is better reproduced in the higher-resolution configuration. Sorting results by liquid water path illustrates a range of model errors with roots in some combination of microphysical and dynamical processes. The 1.25 km configuration, unsurprisingly, does not perform very well especially in thin clouds.

The manuscript is sound: all steps along the way are described in detail and appear technically correct, and the comparison between observations and model results is careful. The manuscript would be most improved by providing more context, motivation, and narrative structure, so that the inferential path readers are asked to follow becomes more clear. This might be accomplished in extensive if minor revision; the authors might also choose to undertake a more thorough re-thinking.

Response: We hope to satisfy this comment through modifications and responses to your individual comments as outlined below and also following Referee #2.

Comment: Readers will be especially grateful for a scientific motivation as to why it's important to look at the distributions of lidar- and radar-derived cloud top and cloud base heights. What model deficiencies does such a comparison highlight? What hypotheses might be tested by examining these statistics? Why have the authors chosen to strike this particular balance of details (e.g. with forward simulation of lidar and radar reflectivities) and abstraction (identifying cloud top and cloud base)? Additional explanation of the motivating ideas would be very welcome. The authors might also use a hypothesis to help prune away a little extraneous material.

Response: We realized, that a better motivation could help the reader. Thus we extend the introduction and explain: "To assess vertically resolved cloudiness and shallow convection, we compare the vertical cloud boundaries. The exact location of the cloud boundaries is the major parameter determining the heating rate profile. Further, the cloud top height is an indicator of the convective activity and therefore allows assessing the model physics with observations indirectly. While cloud fraction at cloud base is rather robust among model assumptions, cloud fraction near the trade inversion varies strongly (Vogel et al., 2020). Clouds near the inversion are often very thin (O et al., 2018) and therefore the problem of instrument sensitivity (e.g., Stubenrauch et al., 2013) can provide different answers about their exact vertical placement."

Comment: As a related point the final section is more speculative than is satisfying. Readers will recognize that the scale of these simulations makes it difficult or impossible to produce variants. They might nonetheless expect the answer to a question or at least evidence for or against a hypothesis. The authors suggest, for example, that the inability of the storm-resolving configurations to reproduce observations in Figures 5 and 6 is due to microphysical choices or possibly vertical resolution. They

neglect the possibility that the horizontal resolution is key, which one might argue based on ideas about the scale of buoyancy production. Is it possible to distinguish between these explanations on the basis of the observations, or might the observations be interrogated differently to provide this insight?

Response: We investigated the effect of horizontal resolution by looking at another, coarser domain from the SRM and LEM
r50 outputs. They are the SRM at 2.5 km and LEM at 600 m grid spacing. The differences between the two SRMs or two LEMs are significantly smaller than those between the 1.25 km SRM and 600 m SRM (see Fig. R1). However, the different spatiotemporal sampling of the model data might have to be considered here as well. We make this more clear in the manuscript, provide Fig. R1 in the new appendix C, and consider the influence of horizontal resolution in the conclusions as well.

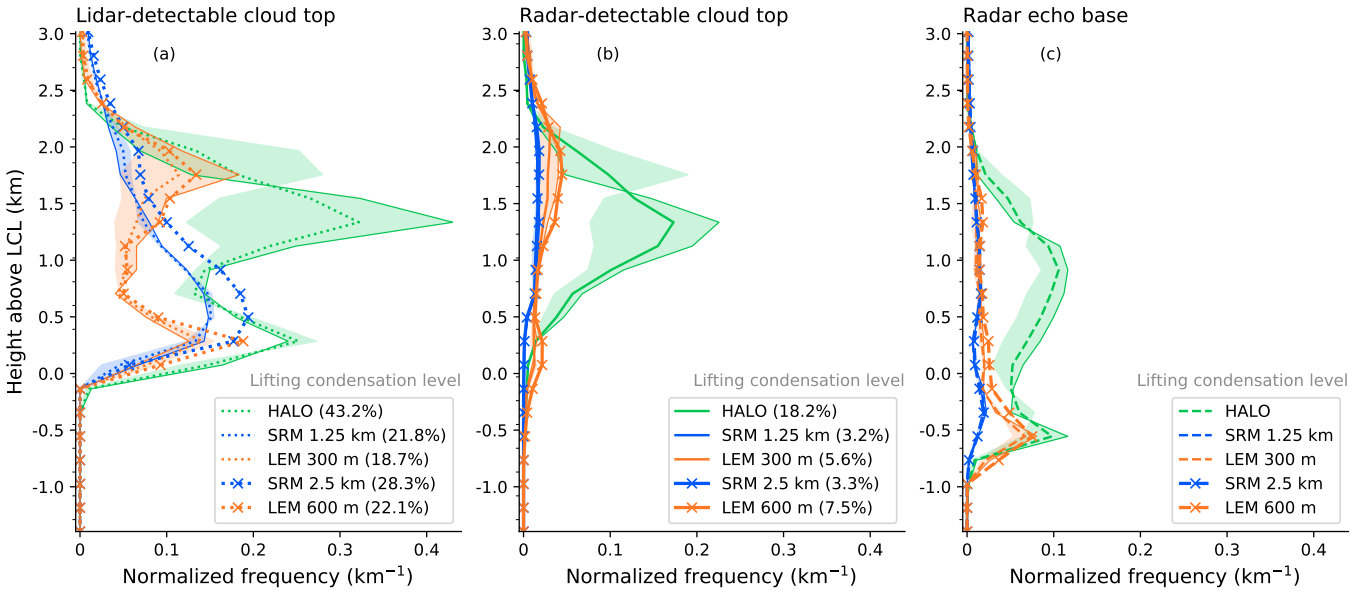

Figure R1. This figure extends Fig. 5 by including data from two additional domains from the SRM and LEM outputs at 2.5 km and 0.6 km grid spacings, respectively. For description see Fig. 5.

r55 **Comment:** The authors combine a sophisticated and careful calculation of synthetic radar and lidar signals, for which differences in the microphysics schemes used SRM and LEM simulations are likely germane, with a very simple analysis based on masking and stratification by liquid water path. How did the authors decide on this approach?

Response: The approach was chosen based on the available observational dataset. As mentioned in the reply to the second comment, we wanted to emphasize the necessity to account for different sensitivities of cloud observations. As the lidar signal
r60 is often attenuated shortly below cloud top, we choose to constrain our analysis to the height of the lidar-detectable cloud top. To then unify the analysis of the radar data with the lidar data, similar masking was applied to the radar data considering that the radar signal is fully attenuated only rarely. This is reflected in introduction of the manuscript (ll 85– 90 in the version below)

Comment: Similarly, what motivates the assumption to ignore rain in computing lidar returns (line 200)? It's true that rain
r65 is quite unlikely to affect the lidar returns but such an assumption seems inconsistent with the level of detail used in other parts of the calculation.

Response: The cloud top height estimation is not expected to change from the raindrop scattering based on the assumption of an optically thin raindrop class. However, we reconfigured CR-SIM to also consider raindrops in the lidar calculation. The related cloud top statistics are shown in Fig. R2 and confirm our assumption, i.e., the rain does not affect the lidar forward
r70 simulation. We clarified the sentence in the revised manuscript as follows. "As raindrops near cloud top are optically thin,

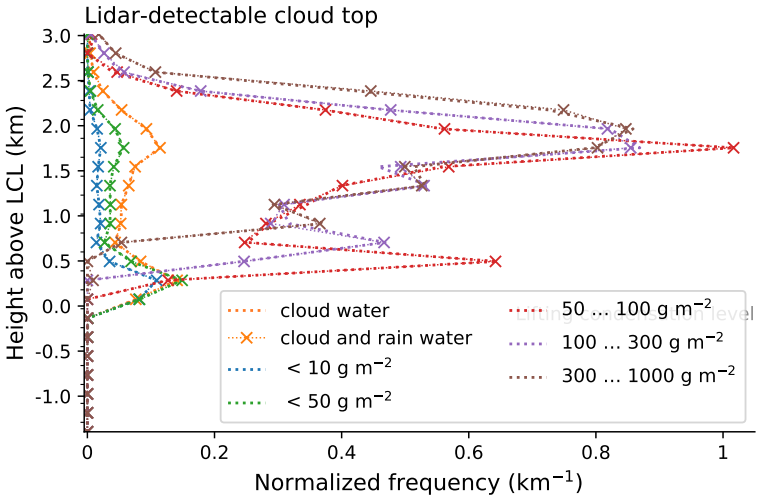

Figure R2. Lidar-detectable cloud top hight distributions from the LEM dataset testing the sensitivity of the lidar forward simulation towards raindrops. Thick are from forward simulations considering cloud droplets only (Like Figures 5a and 7a, d, g, j, and m). Thin lines with x markers are from forward simulations considering cloud droplets and raindrops. Colors denote different LWP regimes.

if present in the models at all, we decided to simplify the forward simulation of the backscatter lidar and thus ignore their contributions."

Comment: The substance of the comparison lies in figures 5 and 6, which show the distribution of cloud tops detected by
r75 lidar and the tops and bases detected by radar from the observations and the two simulations. Figure 6 elaborates on Figure 5 in sorting by liquid water path. One wonders if other analyses might also be informative, especially a look at the joint distribution of lidar and radar cloud top or the joint distribution of radar top and base, depending on the questions motivating the comparison.

Response: Figure R3 shows respective joint distributions. The direct comparison of lidar and radar cloud top and base heights
r80 provides an interesting view on the data, even though it limits the analysis to the cases that provide a radar signal. The clouds observed from HALO are often either about 200 to 400 m thin (along the diagonal) and at about 1 km above the LCL or they precipitate (on the left) with similar cloud top heights. The SRM dataset shows a similar mode of thin clouds at generally lower heights, and precipitating clouds with deeper cloud tops compared to the observations. In the LEM, the non precipitating clouds develop with a wider range of depths. The precipitating clouds however, have a similar top height like in the SRM.
r85 Furthermore, there are a couple of observed cases (right of the diagonal in Figures 6 d-f) in which the lidar-detected cloud top heights are below the radar base heights. These signals relate to lateral raindrop transport out of the precipitating core with a patch of cloud beneath. Such cases also occur in the LEM dataset and less frequently in the SRM dataset. The smaller grid spacing in the LEM could be responsible the higher likelihood for lateral transport of raindrops into a neighbor grid cell in the LEM compared to the SRM.

r90 We included Fig. R3 as Fig. 6 and this discussion into the manuscript.

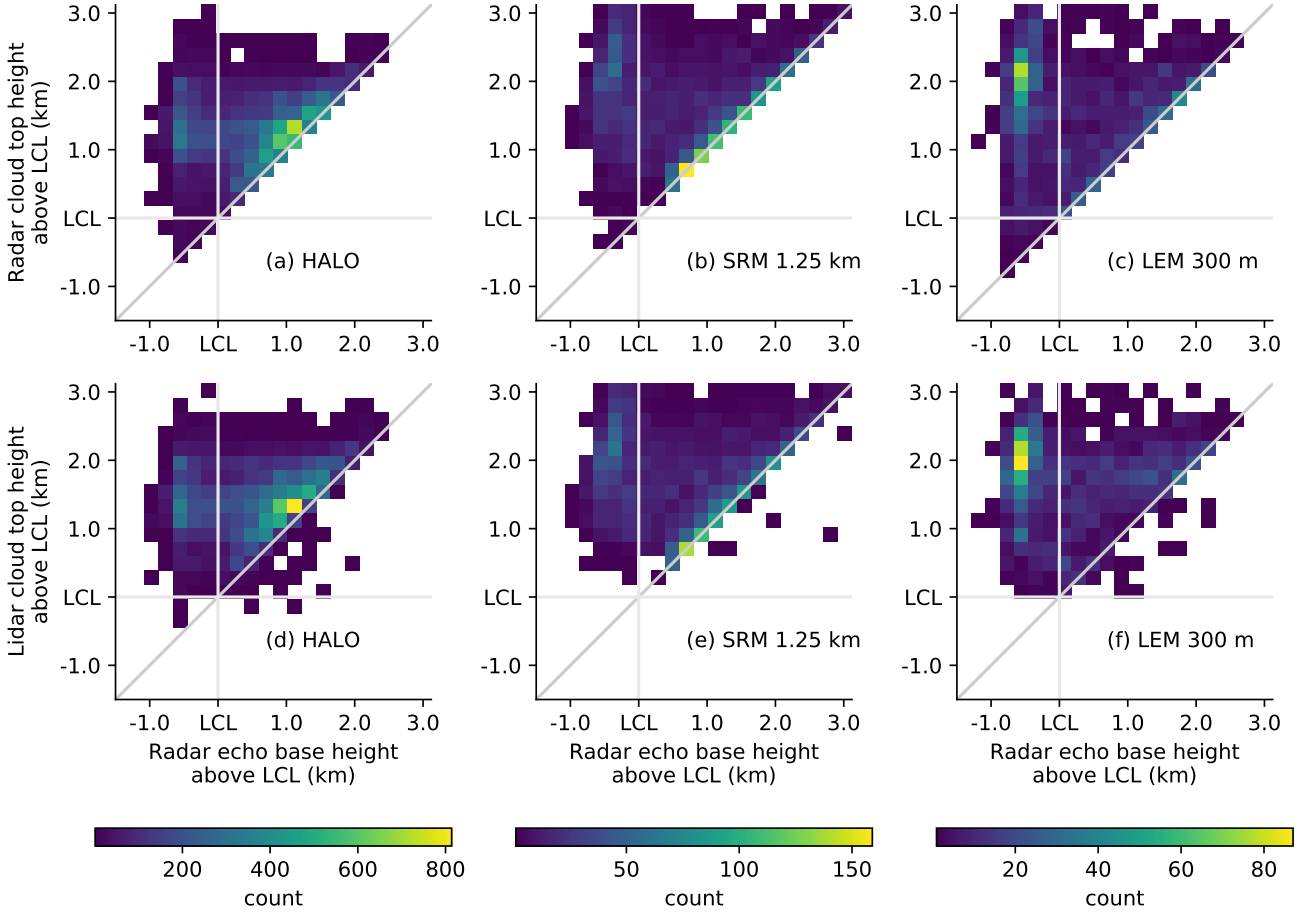

Figure R3. Relation of radar-detectable cloud top height (a-c) and lidar-detectable cloud top height (d-f) above LCL versus radar echo base height above LCL. Based on HALO observations (a, d), 1.25 km SRM dataset (b, e), and 300 m LEM dataset (c, f).

Comment: Section 5.3, which primarily describes Figure 6, would be enhanced by focusing on interpretation of the figure in lieu of description.

Response: We agree with the reviewer and reorganized section 5.3. Furthermore, we added an additional figure (Fig. 8) which
r95 summarizes the key cloud features as a function of LWP. The figure shows the relation of key cloud features as a function of LWP more easily. This figure also shows the influence of model configuration (SRM or LEM) versus horizontal resolution by including the outputs of the SRM and LEM at two additional resolutions and supports the discussion.

General minor points:

r100 **Comment:** The referencing is quite heavily biased towards contributions from Germany. The use of sorting analyses in terms of column water vapor is due to Bretherton et al. 2005 (doi:10.1175/JAS3614.1); the sensitivity of cloud fraction to observing system has been discussed since the 1980s and covered exhaustively by Stubenrauch et al. 2013 (doi10.1175/BAMS-D-12-00117.1), etc. A more balanced and complete view would benefit readers and encourage the first author to read widely.

Response: The reviewer is right: We performed a more detailed literature review and included additional references. Further, we removed the sentence related to water vapor sorting from the manuscript as it provided superfluous information, which could be misinterpreted.

Comment: The manuscript's use of language would benefit from polishing by one of the senior authors. There are more than a few typographic mistakes and a relatively large number of not-quite-standard English constructions which distract unnecessarily.

Response: We performed a more thorough proof reading and revised typos also considering the comments from referee #2.

Comment: The American Meteorological Society, at least, prefers "liftING condensation level" to "liftED".

Response: We adjusted the manuscript to the AMS recommendations and updated the text and all figures accordingly.

Comment: Specific minor points

The sentence spanning lines 49-50 is vague, provocative, and probably unnecessary.

Response: We changed the related sentence to "A reason to initiate the NARVAL expeditions was to extend satellite observations. This can test which cloud variables are sufficiently resolved from space and which characteristics benefit from higher spatial resolution in respect to shallow cumulus clouds." and hope this motivation is clearer and less provocative.

Comment: Section 2, describing the observations, is roughly 55 lines long. Almost half of these and one of two figures is devoted to a discussion of radar sensitivity and the definition of a threshold. Is this point important enough to warrant this level of attention?

Response: We realize, that this discussion might distract the reader. Because the radar sensitivity aspect is important we moved the details and derivation of the threshold into an appendix and only kept the main message in section 2.1).

Comment: Line 153: The line linking the parameterization suite to NWP is perhaps distracting. The authors are careful to describe differences in the microphysical approach in the SRM and LEM later, and to motivate why these might be relevant to the evaluation. This text might raise more questions than it answers.

Response: We removed the distracting and irrelevant mention of NWP and shortened the sentence to "The SRM is run without a convection parameterization." To highlight an important difference of the LEM, we add a link to the applied Smagorinsky scheme for turbulence.

Comment: Line 175: "forward simulations" of what?

Response: We specified the section title and changed it to "Radar and Lidar Forward Simulations".

Comment: Line 182: That forward operators need relevant state information is well-understood. Is there another point here?

Response: In addition to the numerical output fields of the model, assumptions of the model like the drop size distributions have to be considered in the forward simulation. To clarify this, we added the explanation "This means that the same PSD shape and parameters have to be used in the simulator as assumed in the model".

Comment: Line 193: It does seem a bit odd that the synthetic lidar observations are created with a Radar Simulator while the synthetic radar observations are created with yet another package, but maybe there's nothing to be done about that.

Response: We added a clarification that the lidar backscatter was "simulated using the lidar simulation capabilities of the Cloud Resolving Model Radar Simulator".

Comment: Line 199: The relevance of water on a telescope to observations from a platform high above the clouds is not obvious.

Response: We clarified the relevance of this statement by appending the sentence "Furthermore, as the airborne lidar is not affected by liquid collection on the telescope during raining conditions, there is no need to account for such effect" by "as it

would be for a ground based lidar".

Comment: Line 237: Is the finding that precipitation drops do not extend to the very highest reaches of shallow cumulus
r155 novel?
Response: True, this is unsurprisingly but different from the forward simulation as pointed out later in the manuscript. We noted this observational finding to relate it to the forward simulations at a later point of the manuscript.

Comment: Line 259: A better motivation would be useful here. Readers will not expect a case study to be representative of
r160 the entire data set.
Response: We changed the introduction of the subsection to "After introducing and discussing the approach in a case study, all observations and simulations are jointly analyzed in this section.".

Reply to Anonymous Referee #2

r165 **Comment:** Summary: This study uses remote sensing observations from an airborne field campaign over the tropical Atlantic ocean upstream Barbados to assess two sets of cloud-resolving ICON simulations, one at 1.25 km grid spacing (SRM) with one-moment microphysics and the other at 300 m grid spacing (LEM) with two-moment microphysics. The model– observation comparison is based on forward-simulated model output to mimic what the aircraft radar and lidar would see given the atmospheric state and the microphysical assumptions in the model. The authors find the LEM to reproduce the observed bi-
r170 modal cloud top height distribution seen by the lidar, while the SRM fails to represent the upper mode. Stratifying the results into different LWP classes shows that also the LEM model has significant deficiencies in its representation of the radar- and lidar- detected cloud top and base height distributions. This is a nice study that fits well into the scope of GMD. The use of the forward simulation gives interesting new insights about the deficiencies of cloud-resolving simulations in representing shallow cumulus clouds. My main comments regard a more thorough comparison of the representativeness of the selected LEM and
r175 SRM profiles, and an analysis of the uncertainty of the forward-simulation and the sensitivity of the results to the microphysical model assumptions. My general comments are detailed in the following, as well as more specific comments and typographical suggestions.

GENERAL COMMENTS:

Comment: 1. Comparability of selected LEM and SRM profiles:
r180 Vial et al. 2019 showed in their Figure 5 that the 1.25km SRM has a larger cloud cover than the 300-m LEM, especially due to larger contributions from clouds with cloud tops > 1.3km. So I'm surprised that your results here are so different. This might be due to the different microphysical assumptions, but could also be due to the different domains and days used for the SRM vs. the LEM.
Response: In our Fig. 5 (Fig. 5 of the initially submitted manuscript) the 1.25-km SRM has a total projected 2D cloud coverage
r185 of 21.8 %, which is slightly higher than the cloud covered of the 300-m LEM (18.7 %). These numbers are slightly different from the ones presented in Fig. 5 by Vial et al. (2019). Vial et al. denote the total cloud covered as 28 and 20 % for the 1.25-km SRM and 300-m LEM, respectively. This means, their difference is larger, which is likely due to the reduced sample (domain, days) and the fact that we are analyzing day-time data only.
To better quantify these sampling effects, we checked the exact choice of domain and dates used in our analysis by selecting
r190 subsets from the SRM dataset. To clarify this in the manuscript we added Fig. C3 into appendix C and explain the sampling effect as follows. " Figure C3 investigates the exact choice of domain and dates used in the analysis by taking subsamples from the 1.25 km SRM dataset. First the SRM 1.25 km dataset is restricted to those points that are near the LEM meteogram locations. The statistics of different cloud tops and bases in the different spatial subsets (Fig. C3) seems quite robust. Thus, we conclude, that the meteogram locations are in principle able to represent the cloud behavior of the full domain. Further,
r195 we restrict the SRM dataset to the four days for which also LEM output is available. This SRM subsample (also in Fig. C3) indicates a limited development of a deeper stratiform of lidar-detectable cloud, which is, however, not as prominent as in the

observational or LEM datasets (Fig. 5).” Thus the conclusions from our original analysis hold.

Comment: For the LEM, it seems that you are using data from only 10 grid points on 4 days, all sampled at the same latitude. Due to the high temporal resolution of the meteogram output this may give you a lot of profiles, but they will all be highly (auto)correlated. The LEM thus samples much less variable conditions than the SRM. To allow for a more robust and fair comparison of the LEM and the SRM, a comparison of the cloud fractions and/or cloud top height distributions of the LEM and SRM for the same domain and the same days should be made. This would establish how representative the meteogram data is.

Response: See previous response.

Comment: As the necessary input for the forward-simulator is available only for the meteogram points of the LEM, it might be difficult to use the model output of the full LEM domain to do the forward simulation. You could just use the mean and variability of the parameters from the meteogram points to constrain the forward-simulator, which can then be applied to the entire domain.

Response: Unfortunately, no 3D output of rain water content from the LEM is available, such that the analysis of radar-detectable cloud and precipitation and not be extended to the full LEM domain. However, we hope that the representativity analysis in Fig. C3 can answer this question sufficiently.

Comment: Additionally, to understand how much of the forward-simulated SRM-LEM differences come from the different microphysics, I find it important to first show a comparison of the cloud-top height distributions of the two models without using the forward-simulator. This should also be compared to a best-guess observational cloud-top height distribution, either from the lidar alone or from a combination of the lidar and radar-detected clouds. For the lidar, you mention that clouds with liquid water content exceeding 10^{-7} kg/kg are detected. So you could apply this same threshold to the LEM and SRM simulations (for the SRM, also the sub-grid cloudiness will have to be taken into account).

Response:

Forward operators are important to translate sensors threshold into model variables. For example, in the case of one-moment lidar forward simulations, one can read from Fig. 2a that the observational threshold of BSR > 20 corresponds to a model threshold of $q_c > 10^7$ to $10^{-6}\,\mathrm{kg\,kg^{-1}}$. As most of the grid cells show higher lidar backscatter ratio values (color shading in Fig. 2a), such q_c threshold identifies almost all grid cells which would be also lidar-detectable as derived from the forward simulator setup. Likewise radar detection rules can be estimated for the one-moment microphysics from Fig. 2c suggesting the signal detection criterion: $q_r > 10^{-7}\,\mathrm{kg\,kg^{-1}}$ (upper left branch of the distribution) or $q_c > 8 \times 10^{-4}\,\mathrm{kg\,kg^{-1}}$ (lower left).

As expected, the lidar cloud layer detection from the forward simulators can be approximated with the given thresholds to a very high degree of agreement as shown in Fig. R4a. In case of the one-moment SRM microphysics, the threshold-based radar approximation also provides good estimates of the full forward simulation (Figures R4 b and c), once appropriate thresholds are found using the forward simulator. In contrast, the more complex two moment microphysical scheme of the LEM shows that simple thresholds on the model hydrometeor water contents are not sufficient to represent the cloud statistics one could get from more advanced forward simulators. As this analysis shows, there is no unique threshold for cloud definition. Using a threshold for the water load is as arbitrary as any other definition. However, the forward operator approach ensures a fair comparison between observations and models in the same quantity space. To not confuse the reader, we would like to avoid a further excursion from the initial study by discussing these water load thresholds in the manuscript.

Regarding the sub-grid cloudiness, we now explain: “An appropriate analysis of the cloud top height distribution of the sub-grid cloudiness in the SRM would require an assumption on the horizontal overlap of sub-grid clouds within a model column. To circumvent an assumption on this, the influence of the sub-grid clouds is analyzed in terms of the CF profile.” This analysis is already covered in appendix B, but we added the motivation above to clarify why appendix B uses a slightly different metric than the main part.

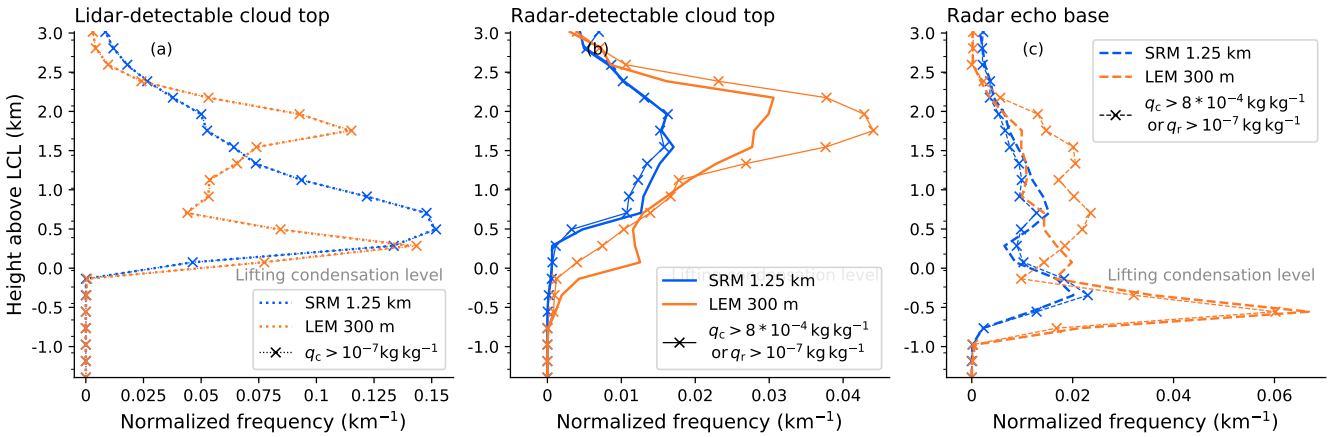

Figure R4. Comparison of cloud top and base heights derived from forward simulations and direct thresholds on model output variables. Thick lines denote distributions from forward simulations, thin lines with x marker denote distributions from direct thresholds. (a) Lidar-detectable cloud top height. (b) Radar-detectable cloud top height. (c) Radar echo base height.

Comment: Apart from showing the frequency distribution as in Figure 5, it might also be worth comparing the cumulative distributions as e.g. in Medeiros et al. 2010 (Figure 7) or van Zanten et al. 2011 (also Figure 7), with the lowermost level representing the total cloud cover.

r245 **Response:** We thank the reviewer for the good idea and added the corresponding plots (Fig. C1 in the appendix) which nicely summarize the results.

Comment: Other questions regarding the simulations are:

- Are there any spin-up issues at the beginning of the simulations and is it feasible to use the LEM simulations already from

r250 12 UTC on, i.e. just 3h after initialization? - In the appendix you mention that the SRM uses a diagnostic cloud scheme in addition to the prognostic cloud scheme. This is an important detail that should already be mentioned in section 3. Furthermore, could you describe how the 'prognostic' cloud scheme works? Is it just a simple saturation adjustment?

Response: The SRM was initialized at 00 UTC on each day. We use data starting from 12 UTC after granting the model 12 hr for spin-up. The LEM was initialized at 09 UTC using outputs from the SRM. The coarsest LEM simulation has the same

r255 resolution as the SRM such that nested LEM should be usable after a shorter spin-up time. Thus, we chose to grant the LEM 3 hr of spin-up. We added details about the spin-up time to the manuscript.

Regarding the prognostic scheme we add that "It is assumed, that turbulent perturbations distribute the total water content in a probability density function (PDF) of rectangular shape centered around its prognostic grid-box mean. The supersaturated part of this PDF is then interpreted as diagnostic cloud cover and $q_{c,dia}$" in the appendix B. Further, we mention the diagnostic

r260 scheme in Sect. 3 now.

Comment: - Difference in vertical resolution: In L420 you mention this as a potential reason for the underrepresentation of inversion cloud in the SRM. I guess the 1.25km SRM version that was used to drive the LEMs should have 150 levels – so in case this model output was saved, you could use this SRM version to verify whether the vertical resolution is indeed the reason

r265 for the reduced anvils. Otherwise you could try to better understand the influence of the horizontal resolution on the anvil cloud amount by comparing the 300m-LEM to the next coarser LEM nest.

Response: The LEM was directly forced with the 75-level SRM. Thus, we cannot analyze this point directly. However, we include data from the next coarser LEM nest with 600 m grid spacing in the analysis. The statistic of that dataset is much more similar to the 300-m LEM, than to the 1250-m SRM. This can also be seen in the new Figures 8 and C2 in the appendix as also

r270 motivated in third comment by referee #1.

Comment: 2. Uncertainty and sensitivity of forward-simulations to model assumptions

I'm not very familiar with forward simulators, but I feel that it would be important to analyze and discuss the uncertainty of the forward simulations, and how this might influence the results. You mention that the forward simulator has to be configured
r275 such that the PSD used in the forward-simulator matches the PSD of the model as good as possible. I assume that there is some uncertainty involved in this process, and it would be good to show or discuss this more explicitly.

Response: This process involves very little uncertainty, as it is about implementing a PSD which matches the one assumed by the model in the forward simulator. To make this more clear and dispel doubts we removed the addendum "as possible" from the manuscript. The related sentence is now: "The forward simulator has to be configured such that the PSD used to simulate
r280 hydrometeor characteristics matches the PSD assumed in the atmospheric model accurately."

Comment: I would also appreciate if you could show somewhere what the variability of the input fields for the scale parameters are in the LEM, i.e. how variable the number concentrations are. This information can also help constrain a forward-simulation using the entire model domain of the LEM.

r285 **Response:** In order to better illustrate the variability we added the quartiles of N_c and a plot for the q_r – N_r relation from the LEM output. Further, we extended the discussion of that figure and the sensitivity of forward-simulation. The last paragraph of Sect. 4 related to the LEM is now: "The median of N_c is $3.8 \times 10^8\,\mathrm{kg}^{-1}$ with an interquartile range between 2.0 an $6.8 \times 10^8\,\mathrm{kg}^{-1}$ for grid cells containing any cloud water. This means N_c is mostly larger in the LEM than in the SRM (fixed to $N_c = 2 \times 10^8\,\mathrm{kg}^{-1}$, Tab. 1), and that cloud droplets are smaller for the same q_c in the LEM." ... "Different to cloud water,
r290 the radar reflectivity of rain is in many cases amplified in the two-moment scheme compared to the one-moment simulation, such that also some grid cells with lower q_r are above the radar detection threshold. This indicates in general larger and fewer raindrops in the LEM than in the SRM for the same q_r as also depicted in Fig. 3."

Comment: Also, I think that you could learn more about the potential deficiencies in the model microphysics by playing
r295 around with the forward simulator and feeding it with slightly adjusted input microphysics parameters. What would have to be different in the microphysics to render the simulations more comparable to the observations, given the simulated mass mixing ratios? You could try to understand how a slight change of the fixed parameters of the SRM one-moment scheme would influence the radar-detectable cloud fraction. Given that the droplet radius is so important for the radar-detectability, Figure 3c might look very different if you'd just fed the forward-simulator with slightly different number concentration parameters.
r300 For the LEM, You could also prescribe the mean number concentrations of the LEM as fixed parameter to mimic what a one-moment microphysics scheme would do.

Response: We concur with the reviewer that instrument simulators generate "synthetic" measurements that we can compare against real observations, thus, offers an alternative methodology for "adjusting" model microphysics. In this framework, observations can be a powerful constraint, however, at the same time, the model should contain enough information (i.e. mi-
r305 crophysical moments) to facilitate the "adjustment".

To better explain the relationship between microphysical quantities, we added to the manuscript that "For a fixed water mixing ratio, a change by a factor of α in the number concentration ($N_c' = \alpha\, N_c$) will result to a change in the radar reflectivity in dBZ from Z to $Z' = Z - 10\log_{10}\alpha$. Thus, if we double the N_c ($\alpha = 2$), the Z will reduce by 3 dB. By the same token, if we change the hydrometeor diameter by a factor of α ($D' = \alpha\, D$), then the radar reflectivity will be $Z' = Z + 30\log_{10}\alpha$".

r310 Thus, as the reviewer suggest, if we attempt to match the observed radar reflectivity with the simulator output, the preference will be to adjust the size of the droplets if the difference is large. If the difference is small (1–2 dB), we can not use the radar reflectivity to adjust this properly in the model since both parameters (N and D) can address small differences. In addition to the direct impact on radar reflectivity, a change in droplet number or size would also influence the dynamics of the model as N is a prognostic variable. This could change the overall model cloud statistics significantly, which would be an interesting topic
r315 for further research.

Comment: A more thorough analysis of the uncertainty and sensitivity of the forward-simulations would render the manuscript scientifically more interesting, and should allow you to make your discussion in Section 6 more robust and less speculative.

Response: We realize that many readers are not so familiar with forward operators and therefore added some of the discussion of the above points into the manuscript.

Comment: SPECIFIC COMMENTS:
- Definition of cloud modes / types (e.g. L263-275): Please better define what you mean with 'thermal driven' mode resp. 'shallow convection' mode. You could also use well-established classifications or definitions such as the 'forced, active and passive' categories of Stull 1985, or the definitions from the cloud atlas of the World Meteorological Organization that were used in Vial et al. 2019 JAMES.
Response: We appreciate this suggestion and moved our previous definition from Sect. 5.3 to the introduction of Sect. 5. There we added the following. "To ease the following discussion, we define three layers in which the lidar and radar signals occur. Every signal below LCL is in the 'precipitation' layer. Typically, only the radar base is in this layer. Clouds with their tops within 600 m above LCL are called 'very shallow clouds' following the definitions by Vial et al. (2019). Vial et al. defined this mode in terms of an absolute top height below 1.3 km which corresponds to a similar height considering that the LCLs in the dropsonde, SRM, and LEM datasets in this study have typical heights of 720 ± 135, 763 ± 144, and 777 ± 121 m, respectively. Cumulus humilis is a typical representative of these very shallow clouds but in principle this class contains also small parts of deeper but slanted clouds. More active clouds can grow deeper than these very shallow clouds until they encounter the trade inversion and are forced to form a lateral outflow which is often perceivable as a stratiform layer. Stratiform remnants of such shallow convection can last for hours and thus much longer than the original convective core (Wood et al., 2018). We summarize all cloud signals above LCL + 600 m as 'stratiform' mode, acknowledging also contributions from active cores."
We adjust the subsequent discussion in the manuscript accordingly. By following the definitions by Vial et al. (2019), we reduced the vertical extend from the lower cloud mode from LCL + 1000 m to LCL + 600 m and thus adjusted cloud fractions in corresponding statements.

Comment: - Cloud-top height detection: I think it is never explicitly written whether you only consider the first-detected highest cloud-top height, or whether you also consider 2nd or pot. 3rd cloud-top heights in case of multilayered cloud scenes. Please mention this explicitly.
Response: We clarified that only the uppermost cloud tops and lowest bases are considered in the second paragraph of Sect. 5.1. It is stated now: "In the case of multi-layer clouds, individual layers could be hidden due to attenuation. Therefore only the uppermost cloud top and lowest base are considered."

Comment: - Referencing previous literature:
o The bi-modal distribution of trade cumuli in the vicinity of Barbados has been extensively studied by Nuijens et al. using data from the Barbados Cloud Observatory. Please refer to Nuijens et al. 2014 QJRMS in L56 and also later on in the manuscript. (Related to this, in L262 it would be good to mention that the 30% dominance of the upper mode vs. the lower mode is opposite when considering ground-based observations (see Nuijens et al. 2014).)
Response: We performed a more detailed literature review and included additional references. Related to Nuijens et al. (2014) we added after L56 from the initial submission: "In addition to spaceborne, also ground-based observations have been used to study the distribution of low-level cloud in the trades. Nuijens et al. (2014) analyzed cloud observations taken at the Barbados Cloud Observatory (BCO; Stevens et al., 2015) at the upstream eastern coast of Barbados at Deebles Point facing the Atlantic Ocean. For two years of ceilometer data of shallow clouds with tops below 4 km, they found that the shallow cloudiness is dominated by clouds near the lifting condensation level (LCL) with about two thirds of the shallow cloud coverage coming from clouds with bases below 1 km." Further, we added the following after L262 from the initial submission: "This is contrary to the ground-based impression from the same region but other period when the shallow mode was clearly dominating (Nuijens et al., 2014)."

Comment: o Observations from the CSET field campaign in the eastern pacific presented in O et al. 2018 GRL and Wood et al. 2018 JAS revealed the common occurrence of persistent thin outflow layers with very low droplet concentrations (the authors refer to 'veil' clouds and ultra-clean layers). It may be good to cite and discuss these papers in the context of the present results.

Response: Thanks for pointing this out. We follow these suggestions and could refer the observations of clouds in the outflow or stratocumulus mode with little LWP (below 10 and 50 g m^{-2}) to such veil clouds. The second paragraph of Sect. 5.3 is
r370 extended to: "It is remarkable that high cloud tops in the stratiform layer were often observed by the lidar under low LWP conditions (below 10 g m^{-2}). Such clouds likely correspond to thin 'veil' clouds frequently observed near the upper boundary layer, i.e., below the trade inversion, in the stratocumulus to cumulus transition by Wood et al. (2018) and O et al. (2018). They report on geometrically and optically thin clouds with low droplet number concentration (about 5 cm^{-3}) but relatively large droplets with radii ranging from 15 to 30 µm. Droplets of such sizes are large enough to provide a radar reflectivity above the
r375 detection threshold.

Extending the LWP class from 10 g m^{-2} to 50 g m^{-2} includes more additional lidar-detectable stratiform cloud coverage to the statistics than very shallow cloud coverage. This means even more veil clouds are included, which were estimated to have a typical LWP of about 25 g m^{-2} (Wood et al., 2018). In all cases with LWP < 50 g m^{-2}, the stratiform layer was observed about 1.5 times more often by the lidar than the layer of very shallow clouds, which is a bit more often than in the LEM and SRM
r380 (see also Fig. 8a)"

Comment: o The referencing for the first two sentences in the introduction should be improved.
Response: The references got mixed up. They should read: "The representation of low-level oceanic clouds contributes largely to differences between climate models in terms of equilibrium climate sensitivity (Bony and Dufresne, 2005; Schneider et al.,
r385 2017). Global atmospheric models with kilometer-scale resolution are considered the way forward in forecasting future climate scenarios (Satoh et al., 2019)."

Comment: - The LCL computation from the dropsondes: Can you say how many dropsondes are used to interpolate the LCL? And by how much they are separated in space and time on average?
r390 **Response:** We added: "Fifty dropsondes were released in total in the study area with a median separation along flight track of about 515 km (quartiles: 384 and 658 km, see also Fig. 1)."

Comment: - Differences between the western and eastern part of the domain related to cloud deepening: Not only is there a difference in the height of the upper mode, but also in the normalized frequencies of the upper mode, with the deeper western
r395 half having a reduced frequency compared to the shallower eastern half. This, and also the insensitivity of the lower mode, was also shown in LES of Vogel et al. 2020 QJRMS.
Response: We appreciate this note. In the revised manuscript, we relate our findings to Vogel et al. (2020) and supplement our statement in the third paragraph of Sect. 5.2: "This deepening probably caused the frequency reduction of the stratiform mode in the western half compared to the shallower eastern half. Such relation between deepening of the cloud layer and reduced
r400 formation of stratiform clouds was also shown in an LES study by Vogel et al. (2020)." [...] "In contrast to the deeper and stratiform clouds, the frequency and height of very shallow lidar-visible clouds is almost the same in the western and eastern parts, which also agrees with Vogel et al. (2020)."

Comment: - Section 5.3: The discussion of the results in this section should be better structured and more focused on the
r405 most important features. It is not always clear what is compared to what, and there is a lot of switching around between LWP categories, the observations and the different models. I also spotted a lot of typographical errors that should be corrected (e.g. L347 partial coverage; L363 that such a cloud doesn't need any contribution...)
Response: Following also a comment by Referee #1, we realized that the discussion in Sect. 5.3 was not fully comprehensible. Thus, we reorganized section 35 and included an additional figure (Fig. 8), which shows the relation of key cloud features as
r410 a function of LWP more easily. This figure also shows the influence of model configuration (SRM or LEM) versus horizontal resolution by including the outputs of the SRM and LEM at two additional resolutions. Furthermore, We did a more thorough proof reading and revised typos.

Comment: - Figure 1: This figure could be improved. Please zoom more into the area of flight operations (only showing
r415 e.g. 7° N to 20° N), make sure that all flight paths are visible and not overlapping, and add markers/crosses for the dropsonde locations.

Response: We improved Fig. 1 following the suggestions. The partial overlap of flight tracks cannot be excluded completely, as the flights were accurately flown along the same track using the autopilot. However, we used different line widths to make sure that all tracks are visible.

Comment: - Figure 5: What exactly does the cloud fraction in the legend refer to? Is it just the maximum cloud fraction? It would be nice to give the total projected cloud cover instead of a cloud fraction, as this would give a sense of the total cloudiness.
Response: The percentage in the parenthesis actually refers to the total projected 2D cloud coverage. To make this more clear, we added "2D" in several occasions and went through the manuscript to consistently use the terms "cloud coverage" for the projected 2D cloud coverage and "cloud fraction" for the vertically resolved profile of cloudiness.

Comment: - Figure A1: similar to the above, in the caption you mix cloud cover and cloud fraction, but I guess you mean the same thing.
Response: Both should be the same. Thus, we changed "cloud cover" to "cloud fraction" in the caption.

Comment: TYPOGRAPHICAL SUGGESTIONS: - L65 (and everywhere else): model data –> model output
Response: We followed this suggestion.

Comment: - L89: maybe add 'first phase of the NARVAL... Often referred to as NARVAL1 in other studies.
Response: We now mention NARVAL1 in parenthesis: "NARVAL-South (also referred to as NARVAL1)". However the term NARVAL1 is not used in the manuscript to not confuse with flights over the mid-latitude North Atlantic which where also part of the first NARVAL field experiment (Konow et al., 2019).

Comment: - L97: this sentence is a bit odd there and should be moved down after L104, maybe adj. it to 'The following subsections describe...'
Response: We followed this suggestion.

Comment: - L98: form –> from
Response: Corrected.

Comment: - L99: better than 20 g/m-2 and 10% (?)
Response: We now make clear, that this error is relative to the LWP itself, i.e. "The LWP retrieval from the microwave radiometer has a high accuracy, which is better than 20 g m^{-2} for LWP < 100 g m^{-2} and relatively better than 20 and 10 % of the retrieved LWP for LWP greater than 100 g m^{-2} and 500 g m^{-2}, respectively, as described by Jacob et al. (2019)."

Comment: - L107: radar –> reflectivity
Response: We added "reflectivity" in "the radar reflectivity is approximately proportional".
.

Comment: - L116: Ragged point –> Deebles point
Response: Corrected.

Comment: - L124: a clear frequency maximum?
Response: We clarified that the "frequency maximum" was meant.

Comment: - L206 & other instances throughout the manuscript: remove commas before 'that' ⟶ "it has to be noted that..."
Response: We removed the comma before "that" in several occasions.

Comment: - L212: by –> be
Response: We corrected "can only by reached" to "can only be reached".

Comment: - L255-257: could be omitted.
Response: We omitted these last two sentences.

Comment: - L285: what do you mean with shallow clouds here? (Please also see my specific comment on the definition of the cloud modes above)
Response: Now, we define cloud modes more clearly in the beginning of Sec. 5.

Comment: - L306: remove 'is'
Response: Removed.

Comment: - L316: simulated infrequently –> underrepresented
Response: We changed "simulated infrequently compared to observations" to "underrepresented compared to observations".
Comment: - L321ff: The thresholds for the LWP classes are different from Figure 6 (i.e. lower bounds not given)
Response: In the text, we included the lower threshold in the parenthesis to agree with that figure.

Comment: - L338: Ref to figure 5 and not figure 4
Response: Corrected.

Comment: - L383: liar –> lidar
Response: Corrected.

Comment: - L419: a gap of what? Please be more specific
Response: It is the "gap in the cloud top frequency distribution".

Comment: - L424: might be an...
Response: Corrected.

Comment: - L491: remove 'with'
Response: Removed.

Reply to Executive Editor Comment by Astrid Kerkweg

Comment: Dear authors,
in my role as Executive editor of GMD, I would like to bring to your attention our Editorial version 1.2:
https://www.geosci-model-dev.net/12/2215/2019/
This highlights some requirements of papers published in GMD, which is also available on the GMD website in the 'Manuscript Types' section:
http://www.geoscientific-model-development.net/submission/manuscript_types.html
In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:

> If the model development relates to a single model then the model name and the version number must be included in the title of the paper. If the main intention of an article is to make a general (i.e. model independent) statement about the usefulness of a new development, but the usefulness is shown with the help of one specific model, the model name and version number must be stated in the title. The title could have a form such as, "Title outlining amazing generic advance: a case study with Model XXX (version Y)".

r510 As you are using just one model, please state in the title that you are using ICON and mention its version number.

Additionally, please provide more details on the used codes in the "Code availability" section: provide the numbers or unique identifiers of the code versions used and make sure that these versions (because this is relevant to reproduce the results of this article) are permanently archived and accessible in the future. With regard to the ICON model code I understand, that you did not perform the simulations yourself, but from your provided references (Klocke et al., 2017, and Vial et al., 2019) it is hard
r515 to find out (if at all) which model version(s) have been used and how to access the code. Please provide this information upon submission of the revised version.

Response: Two fundamentally different versions of ICON were used in this study. The ICON storm resolving model (SRM), which is typical run for numerical weather prediction, and the ICON-based Large Eddy Model (LEM), used by the research community on a case study bases. To make this clear, we would like to change the title to "Multi-layer Cloud Conditions
r520 in Trade Wind Shallow Cumulus – Confronting two ICON Model Derivatives with Airborne Observations"[1]. Regarding the model versions, we specified that "The ICON SRM was run using revision '28436M' of the 'icon-nwp/icon-nwp-dev' branch (Klocke et al., 2017). The ICON LEM was run using the ICON release 2.3.00 (Stevens et al., 2019)." in the data availability section. If further details on the used model output are desired, please let us know, so we could make all ICON model output data we used available on a data archive.

r525 **Other minor changes**

We renamed the cloud cover variable clc to "cloud fraction (CF)" in the appendix B.

We changed the abbreviation of the autoconversion to the more commonly used AU.

r530 We renamed variable name for the droplet number concentration from q_{nc} to the more commonly used "N_{c}".

We updated the references.

[1] Correcting of our initial reply (https://doi.org/10.5194/gmd-2020-14-AC1), it should be "Derivatives".

Multi-layer Cloud Conditions in Trade Wind Shallow Cumulus – Confronting ~~Models~~ two ICON Model Derivatives with Airborne Observations

Marek Jacob[1], Pavlos Kollias[1,2], Felix Ament[3], Vera Schemann[1], and Susanne Crewell[1]

[1]Institute for Geophysics and Meteorology, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany
[2]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794-5000, USA
[3]Meteorological Institute, Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

Correspondence: Marek Jacob (marek.jacob@uni-koeln.de)

Abstract. Airborne remote sensing observations over the tropical Atlantic Ocean upstream of Barbados are used to characterize trade wind shallow cumulus clouds and to benchmark two cloud-resolving ICON (ICOsahedral Nonhydrostatic) model simulations at kilo- and hectometer scales. The clouds were observed by an airborne nadir pointing backscatter lidar, a cloud radar, and a microwave radiometer in the tropical dry winter season during daytime. For the model benchmark, forward operators convert the model ~~data~~ output into the observational space for considering instrument specific cloud detection thresholds. The forward simulations reveal the different detection limits of the lidar and radar observations, i.e., most clouds with cloud liquid water content greater than 10^{-7} kg kg^{-1} are detectable by the lidar, whereas the radar is primarily sensitive to the "rain"-category hydrometeors in the models and can detect even low amounts of rain.

The observations reveal two prominent modes of cumulus cloud top heights separating the clouds into two layers. The lower mode relates to boundary layer convection with tops closely above the ~~lifted~~ lifting condensation level, which is at about 700 m above sea level. The upper mode is driven by shallow moist convection, also contains shallow stratiform outflow anvils, and is closely related to the trade inversion at about 2.3 km above sea level. The two cumulus modes are ~~reflected~~ sensed differently by the lidar and the radar observations and under different liquid water path (LWP) conditions. The storm-resolving model (SRM) at kilometer scale reproduces the cloud modes barely and shows ~~the~~ most cloud tops being slightly above the observed lower mode. The large-eddy model (LEM) at hectometer scale reproduces better the observed cloudiness distribution with a clear bimodal separation. We hypothesize that slight differences in the autoconversion parametrizations could have caused the different cloud development in the models. Neither model seems to account for in-cloud drizzle particles that do not precipitate down to the surface but generate a stronger radar signal even in scenes with low LWP. Our findings suggest that even if the SRM is a step forward for better cloud representation in climate research, the LEM can better reproduce the observed shallow cumulus convection and should therefore in principle represent cloud radiative effects and water cycle better.

Copyright statement. TEXT

1 Introduction

The representation of low-level oceanic clouds contributes largely to differences between climate models in terms of equilibrium climate sensitivity ~~(Schneider et al., 2017)~~(Bony and Dufresne, 2005; Schneider et al., 2017). Global atmospheric models with kilometer-scale resolution are considered ~~as~~ the way forward in forecasting future climate scenarios ~~(Bony and Dufresne, 2005; Satoh et al., 2019)~~. (Satoh et al., 2019). The increased model resolution and better matching scales with measurements allow for a more direct observational assessment by comparing the present day representation in the models with atmospheric measurements and thus anchoring models to reality. Recently, Stevens et al. (2020) demonstrated the general advantage of high resolution simulations compared to typical climate models in terms of cloud representation using different versions of the ICOsahedral Non-hydrostatic model (ICON). The progress in such novel large-area high-resolution models and new capabilities of synergetic airborne measurements in the trades motivate the following guiding questions of this study. How do two cloud-resolving versions of the ICON model represent shallow cumuli in comparison to observations? What is an appropriate approach to ~~asses~~ assess the model clouds? How does the liquid water path (LWP) help ~~to interpret~~ interpreting differences between observed and simulated cloud structures?

Increased model resolution facilitates the model-observation comparisons. However, there are several other factors to be considered (Lamer et al., 2018). On the one hand, particle size distributions (PSDs) in models are typically represented by bulk and spectral microphysical schemes, or Lagrangian superparticles (e.g., Grabowski et al., 2019). Bulk microphysics schemes predict changes in condensate using one to three moments. These are usually the lower moments like particle number concentration and mixing ratio (Khain et al., 2015). On the other hand, radars and lidar, like those used in this study, observe different moments of the PSD. A backscatter lidar, for example, is primarily sensitive to the second moment, while a radar is sensitive to the sixth moment.

An objective definition of a cloud is required when comparing cloudiness in models with observations. If one asks different instrument operators to provide average "cloud coverage", one can get different answers, e.g., 19 to 46 %, for the very same scene as demonstrated by Stevens et al. (2019). This range is caused by different sensitivities due to different measurement principles and sampling methods by the remote sensing instruments involved and also affects global climatologies (e.g., Stubenrauch et al., 2013). To find a common definition, it is favorable to compare clouds in models and observations in terms of the same quantities (Bodas-Salcedo et al., 2011). Here, forward simulators can be used to simulate measurements as they would be recorded by a radar or lidar, based on the atmospheric state and assumptions in the model (Lamer et al., 2018).

The observations used in this study were recorded with the research aircraft HALO ~~(High Altitude and LOng range, Krautstrunk and Gie~~ (High Altitude and LOng range; Krautstrunk and Giez, 2012) which was equipped as a flying remote sensing cloud observatory during the NARVAL-South experiment ~~(Next generation Advanced Remote sensing for VALidation, Klepp et al., 2014)~~ (Next generation Advanced Remote sensing for VALidation; Klepp et al., 2014) in December 2013. A reason to initiate the NARVAL expeditions was ~~that satellites cannot provide sufficient resolution for multiple cloud variables .~~ to extend satellite observations. This can test which cloud variables are sufficiently resolved from space and which characteristics benefit from higher spatial resolution in respect to shallow cumulus clouds. For example, the spaceborne Cloud-Aerosol Lidar with Orthog-

onal Polarization (CALIOP) has frequently been used to investigate marine low clouds. Leahy et al. (2012) observed two modes of low clouds in the tropical Pacific trade wind, and reveal that CALIOP misses small clouds (< 1 km) and combines adjacent but separated clouds due to the CALIOP sampling rate. Leahy et al. identified modes at about 800 and 2000 m above the sea surface in tropical Pacific trade wind cumulus (15° S, 155° W) using two years of data. The shallower mode is considered to be formed by small cumuli with insufficient buoyancy to grow while convective clouds and detrained elements from thick clouds suppressed under subsidence form the deeper mode. Genkova et al. (2007) compared trade wind cumuli cloud top heights from passive optical spaceborne instruments. They ~~also observed bimodal distribution with~~ identified two cloud top modes at 650 and 1500 m above sea level in an area similar to this study (10°–20° N to 55°–65° W) from about 150 scenes between September 2004 and March 2005 using data from three different satellites~~but~~. However, they also found vertical biases of 250 to 500 m due to different retrieval approaches and spatial resolutions.

In addition to spaceborne, also ground-based observations have been used to study the distribution of low-level cloud in the trades. Nuijens et al. (2014) analyzed cloud observations taken at the Barbados Cloud Observatory (BCO; Stevens et al., 2015) at the upstream eastern coast of Barbados at Deebles Point facing the Atlantic Ocean . For two years of ceilometer data of shallow clouds with tops below 4 km, they found that the shallow cloudiness is dominated by clouds near the lifting condensation level (LCL) with about two thirds of the shallow cloud coverage coming from clouds with bases below 1 km.

Since active instruments are advantageous for observing cloud ~~heights~~vertical extents, the HALO instrumentation included an aerosol backscatter lidar as part of the WALES ~~(WAter vapor Lidar Experiment in Space, Wirth et al., 2009)~~ (WAter vapor Lidar Experi airborne demonstrator, and a cloud radar. The radar is one part of the HAMP ~~(HALO microwave package, Mech et al., 2014) together with~~ (HALO microwave package; Mech et al., 2014) while a microwave radiometer is the other. The latter provides the vertically integrated ~~LWP~~ liquid water content (LWP) (Jacob et al., 2019), which helps to approach the liquid water content which is a key quantity to describe clouds in models like the ICON. The direct observation of the liquid water content profile is difficult (Crewell et al., 2009), but the LWP can be used to estimate the water content when combined with estimates of cloud vertical extend by lidar and radar either in a simple average approach or more sophisticated as a profile (Frisch et al., 1998; Küchler et al., 2018). In addition, dropsondes were released regularly during the flights to probe the temperature and humidity profile. Compared to ground-based observations, the airborne remote sensing instruments, especially the microwave radiometer, have the advantage of not being harmed by precipitation or sea spray deposition on the instrument (Rose et al., 2005).

The observations are used to confront the simulations of winter season trade wind cumuli in the tropical west Atlantic Ocean. Such clouds are regularly subject in idealized ~~large eddy~~ large-eddy simulation (LES) studies ~~(e.g., Siebesma et al., 2003; van Zanten et al.,~~ (e.g., Siebesma et al., 2003; van Zanten et al., 2011; Bretherton and Blossey, 2017) due to their high relevance for the climate. As it is difficult for small domain LES models to generate realistic mesoscale cloud organization (Jeevanjee and Romps, 2013), we use simulations by Klocke et al. (2017) that were run on large domains (> 1500 x 900 km) with kilo- and hectometer horizontal grid spacings and were forced by numerical weather prediction output. Simulations with 1.25 km grid spacing were produced using the storm-resolving model (SRM) version of ICON, while simulations with 3 hectometer grid spacing were produced using ICON large-eddy model (LEM).

To ~~asses~~ assess vertically resolved cloudiness and shallow convection, we compare the vertical cloud boundaries. The exact location of the cloud boundaries is the major parameter determining the heating rate profile. Further, the cloud top height is an indicator of the convective activity and therefore allows assessing the model physics with observations indirectly. While cloud fraction at cloud base is rather robust among model assumptions, cloud fraction near the trade inversion varies strongly (Vogel et al., 2020). Clouds near the inversion are often very thin (O et al., 2018) and therefore the problem of instrument sensitivity (e.g., Stubenrauch et al., 2013) can provide different answers about their exact vertical placement. The use of instrument simulators applied to model output allows the different sensitivities of the instruments to be taken into account, as explained in the following.

As the backscatter lidar ~~is quickly attenuated completely~~ becomes completely attenuated by the presence of hydrometeors in a cloud quickly, lidar measurements and their forward simulations are considered for a cloud top height estimate only. The radar, however, can penetrate through the cloud and precipitation layers and thus provides estimates of cloud or precipitation base ~~height~~ heights in addition to cloud top ~~height~~heights. As shallow cumulus convection is not expected to trigger at the same time and place in a model and reality, a statistical approach is adopted here, in which the airborne observations are compared to their model counterpart for different LWP regimes. ~~The analysis in LWP space is similar to the studies in moisture space that first have been published by Schulz and Stevens (2018) for ground-based observations and by Naumann and Kiemle (2019) for airborne observations.~~ In the LWP space it is possible to study microphysical cloud processes like the transition from non-precipitating to precipitating clouds.

This paper is structured as follows: The observations and their sensitivities in Sect. 2 are followed by a brief description of the model ~~data~~ setup and output in Sect. 3. Then, the forward simulations are presented in Sect. 4 taking into account the instrument characteristics and specifications of model outputs. Finally, the model ~~data~~ outputs of ICON SRM and LEM are confronted with the airborne observations in Sect. 5 including the analysis in LWP space. A summary and conclusions are given in Sect. 6.

2 Observations

The airborne measurements were taken during the NARVAL-South (also referred to as NARVAL1) field experiment in the tropical Atlantic east of Barbados. The NARVAL remote sensing package (Stevens et al., 2019) recorded data during 8 research flights in the tropical domain south of 20° N from 10 to 20 December 2013. The flight tracks are depicted in Fig. 1. A total of about 22 000 km of HALO along track observations with about 91 thousand profiles were sampled at a frequency of 1 Hz from altitudes between 13 and 14.5 km. Further details of the experiment and flight ~~planing~~ planning are provided by Klepp et al. (2014) and Konow et al. (2019). In this study we use the backscatter lidar cloud top height~~time series~~, the radar reflectivity factor Z, liquid water path (LWP) retrieved from microwave radiometer, and the ~~lifted condensation level (LCL)~~ LCL estimated from dropsondes. The remote sensing lidar, radar, and microwave radiometer were installed in a near-nadir pointing direction under the fuselage of the aircraft.

~~This section briefly describes the measurement principles of the radar and lidar and the respectively used thresholds for cloud detection.~~ The LWP retrieval ~~form~~ from the microwave radiometer has a high accuracy, which is better than 20 g m^{-2} for LWP < 100 g m^{-2} and relatively better than 20 and 10 % of the retrieved LWP for LWP greater than 100 g m^{-2} and 500 g m^{-2}, respectively, as described by Jacob et al. (2019). The LWP is defined as the integral of all liquid in the column comprising cloud liquid and rain water. The LCL is derived from the dropsonde temperature and relative humidity (RH) measurements closest to the surface using the code by Romps (2017). The LCL measurement uncertainty is mostly affected by the RH measurement, such that an overestimation on the order of the calibration repeatability of 2 % RH (Vaisala, 2017) would result in an about 60 m lower LCL. The LCL from dropsonde releases is temporally interpolated to generate a continuous time series along the flight track. Fifty dropsondes were released in total in the study area with a median separation along flight track of about 515 km (quartiles: 384 and 658 km, see also Fig. 1). The following subsections describe the measurement principles of the radar and lidar and the respectively used thresholds for cloud detection.

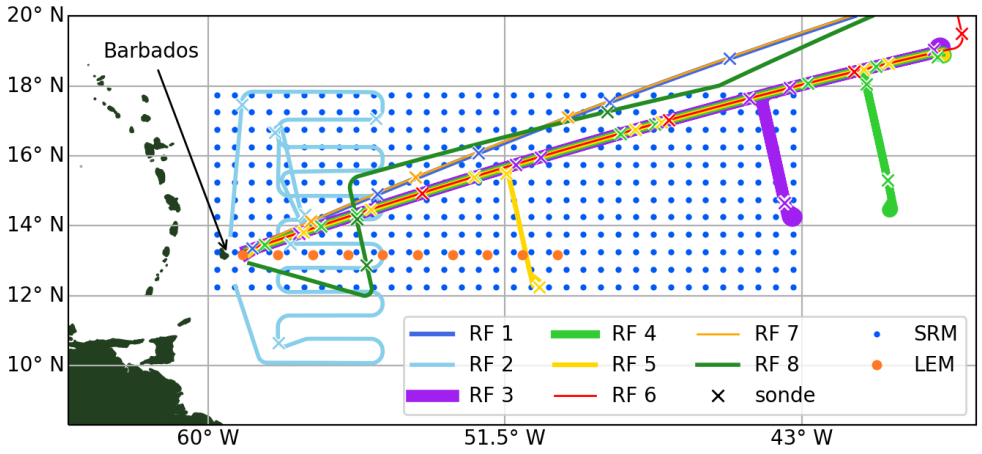

Figure 1. Map showing research flight (RF) tracks and the model columns, which are used in this study. The storm-resolving model (SRM, blue, original model grid spacing: 1.25 km) is thinned to a 0.5° × 0.5° grid. From the large-eddy model (LEM, orange, original model grid spacing: 300 m), ten meteogram outputs are used.

2.1 Radar~~Sensitivity~~

The radar reflectivity factor – short "reflectivity" – Z is measured by the HAMP radar at 35.5 GHz. In case of small spherical liquid droplets, the radar reflectivity is approximately proportional to the sixth moment of the PSD at a given range. This means that larger raindrops show a higher reflectivity than smaller cloud droplets given the same mass mixing ratio. The HAMP radar is calibrated following Ewald et al. (2019) and was operated at a vertical resolution of about 30 m with 1 Hz sampling. This sampling frequency corresponds to a surface footprint of about 136 m × 376 m at a cruising speed of about 240 m s^{-1}.

The instruments minimal detectable signal (MDS) in dBZ decreases with range r and is estimated by Ewald et al. (2019) as

$$\mathrm{MDS}(r) = -39.8 + 20\log_{10}\left(\frac{r}{5\,\mathrm{km}}\right). \quad (1)$$

According to this equation, the MDS in the shallow cumulus layer is about −32 dBZ when flying at 13 km. However, this does not include sensitivity reduction due to Doppler broadening caused by the aircraft motion (Mech et al., 2014). ~~To estimate the practical sensitivity limit, HAMP radar statistics are compared to~~ In this study, we use a sensitivity threshold of −20 dBZ by comparing HALO radar statistics with ground-based ~~measurements. The ground-based measurements were taken at the Barbados cloud observatory (BCO) at the upstream eastern coast of Barbados at Ragged Point (Stevens et al., 2015). The BCO radar operates at the same Ka-band frequency as the airborne radar, but has a better sensitivity due to a larger antenna and longer integration time (Lamer et al., 2015). Therefore, the lower MDS of the BCO radar offers the opportunity to asses the practical sensitivity limit of the HAMP radar.~~

~~A comparison can only be made on a statistical basis as the BCO and HAMP radars do not sample the same volume. To avoid statistical effects of the diurnal cycle identified by Vial et al. (2019), BCO data are only considered roughly during the time when HALO was flying, i.e., between 12:00 and 21:00 UTC (8:00 and 17:00 local time) on the 8 flight days.~~

~~The higher BCO sensitivity compared to HAMP is notable in the height-resolved reflectivity histograms in Fig. A1. The BCO radar frequently measures reflectivity signals down to at around with a clear maximum below for Z up to~~ measurements taken at the BCO as outlined in the appendix A. ~~Klingebiel et al. (2019) identify such weak signals at BCO below as originating from sea salt aerosols and only signals above are attributed to clouds. Clouds with reflectivity between the HAMP MDS ( ) and and within above sea level are observed in of the time at BCO but only rarely (<) by HAMP. Only clouds with a reflectivity higher than about are similarly or more often observed by HAMP than at BCO. Thus, we use as the practical cloud detection threshold of HAMP and use this value in the further analysis to define in the observations and forward simulations.~~

2.2 ~~LIDAR~~Lidar

The lidar system WALES supplements the HAMP radar with optical active remote sensing on HALO. WALES comprises a water vapor differential absorption lidar system (DIAL) at different wavelengths and a high spectral resolution lidar (HSRL) which measures molecular and aerosol backscatter at 532 and 1064 nm. The scattering of an emitted laser pulse on a liquid hydrometeor mostly follows the principles of geometrical optics as the wavelength is much smaller than the particle. Therefore, the back-scattered energy is in first order approximation proportional to the hydrometeor ~~diameter~~ cross section and thus to the second moment of the PSD (O'Connor et al., 2005). This means ~~,~~ that a backscatter lidar is more sensitive to the number of small droplets compared to a radar. Besides hydrometeors, also other aerosol particles like dust scatter the lidar pulse back. However, the aerosol signal is much smaller than the hydrometeor signal. Therefore, we follow Gutleben et al. (2019) and use a threshold of backscatter ratio (BSR) > 20 in the 532 nm channel to differentiate cloudy scenes from clear-sky or dusty scenes. As ~~hydrometeors~~ cloud droplets attenuate the lidar signal strongly, the WALES lidar is used only to detect cloud top height using that threshold. ~~The~~ This lidar top height is measured every second with a vertical accuracy of 15 m and the lidar footprint width at the surface is at about 22 m.

3 ICON-NARVAL model ~~data~~output

Two different versions of the ICON model were run to supplement the NARVAL experiment. The runs of the ~~so called~~ so-called storm-resolving model (SRM) and the large-eddy model (LEM) are described by Klocke et al. (2017) and Vial et al. (2019). The most important aspects relevant for this study of the SRM and LEM are summarized in this section.

3.1 ICON SRM

The SRM (Zängl et al., 2015) was run at 2.5 and 1.25 km horizontal grid spacing with a stretched vertical grid of 75 levels up to 30 km which has 12 and 22 levels below 800 m and 3 km, respectively. The domain spans the western tropical North Atlantic from 4° S to 18° N and from 64° W to 42° W ~~. The~~ (Stevens et al., 2019). The 1.25 km SRM is one-way nested into a coarser 2.5 km SRM which is initialized and nudged with lateral boundary data from the European Centre for Medium-Range Weather Forecasts (ECMWF). The SRM ~~uses physical packages that are similar to those used in operational numerical weather prediction codes, but does not use a convection parametrization.~~ is run without a convection parameterization. The cloud and precipitation microphysics are represented by a one-moment microphysics scheme (Baldauf et al., 2011) that predicts the specific water contents of five different hydrometeor classes including liquid cloud water (q_c) and rain (q_r). In addition, a diagnostic sub-grid scheme adjusts the cloud fraction and cloud water input to the radiation code. In the following however, the cloud water predicted by the prognostic equations is considered primarily, but the potential influence of the diagnostic scheme is discussed in appendix B. 17 modeled days from 10 to 28 December 2013 are used and cover the whole NARVAL experiment. More details on the model setup can be found in Klocke et al. (2017).

The model output is archived hourly. ~~This study only~~ The SRM was initialized at 00 UTC on each day. This study uses model output between 12:00 and 21:00 UTC ~~to avoid influence of a diurnal cycle . This is analogous to the ground-based data described in Sect. 2.1.~~ (08:00 - 17:00 local time) granting the model 12 hr of spin-up and avoiding diurnal cycle influence. The data are spatially subsampled on a coarser $0.5° \times 0.5°$ grid to reduce the computational effort while still conserving the variety of atmospheric profiles. A compromise of domain overlap between all available model ~~data~~ output and observations is achieved by limiting the SRM ~~data~~ output to the area of 12 to 18° N and 60 to 43° W as marked in Fig. 1. ~~The total number of analyzed SRM columns in this study is 97920.~~ If not specified differently, the term "SRM" relates to the 1.25 km grid-spacing. The 2.5 km SRM output is only considered to discuss the influence of model resolution.

3.2 ICON LEM

The LEM (Dipankar et al., 2015; Heinze et al., 2017) with 300 m grid spacing was run in a multi-step nested setup, including a 600 m LEM nest, forced with the SRM . ~~This means,~~ The LEM was initialized at 09 UTC of each simulation day using the outputs from the SRM. This means that the LEM also has a realistic, non-idealized ~~setup~~initialization. The LEM vertical grid also reaches up to 30 km ~~but~~ and has 150 levels with 14 and 37 of them below 800 m and 3 km, respectively. The LEM physics package differs from the SRM configuration. The LEM uses a Smagorinsky scheme for turbulence, but the most important difference for this study is that the microphysics are represented by the two-moment scheme of Seifert and Beheng (2001).

This scheme predicts the hydrometeor number concentrations in addition to the specific water contents and thus provides ~~q_{nc} and q_{nr}~~ N_c and N_r for liquid cloud water and rain, respectively.

In contrast to the SRM, the LEM was only run for the six days of research flights 2 to 6 and 8. However, the full hydrometeor state including rainwater and the number concentrations were only archived for four of the runs in the form of ~~so called~~ so-called "meteogram output". This means that hydrometeor profiles are available with high temporal resolution (every 36 s) but only at 12 model columns. Such meteogram output was saved for the days of research flights 4, 5, 6, and 8. The ten model columns east of Barbados are used for this study and are also marked in Fig. 1. The LEM ~~data are~~ output is also limited to the time between 12:00 and 21:00 UTC~~. The total number of analyzed LEMcolumns in this study is 37030.~~ , after 3 hr of spin-up. The term "LEM" relates to the 300 m grid-spacing simulation if not specified differently. The 600 m LEM, however, is only considered for the assessment of horizontal resolution.

4 Radar and Lidar Forward Simulations

Forward simulators, also called forward operators, can simulate how the remote sensing instruments presented in Sect. 2 would perceive a scene provided by an atmospheric model. A forward simulator requires input like model variables and the knowledge about the microphysical assumptions employed in the atmospheric model. The basic variables are temperature, pressure, layer height, and humidity for each model level in a column for a 1D vertical forward simulation. The variables describing the hydrometeors depend on the microphysical scheme. Typically, these include mass mixing ratios (e.g., q_c or q_r) of different hydrometeor classes. The forward simulator has to be configured such ~~,~~that the PSD used to simulate hydrometeor characteristics matches the PSD assumed in the atmospheric model ~~as accuratelyas possible~~accurately. This means that ~~for~~ the same PSD shape and parameters have to be used in the simulator as assumed in the model. For models with advanced microphysical schemes, also the variables describing those aspects of the PSD are important input parameters for the forward simulation and need to be saved during the model run. In the case of the ICON LEM, the two-moment scheme by Seifert and Beheng (2001) uses the particle number concentrations as additional variables.

As this study focuses on the tropical shallow cumulus below freezing level, we confine the following description and analysis to precipitating and non-precipitating liquid hydrometeors, which are the raindrops and cloud droplets in the ICON microphysical schemes. Both ICON models ~~which are used in the subsequent Sect. 5 assume PSDs with~~ assume modified Gamma distributions as PSDs. The number concentration ~~N~~ $N(D)$ of spherical drops with diameter D can be described as

$$N(D) = N_0 \, D^{\mu} \exp(-\Lambda \, D^{\gamma}) \tag{2}$$

with the scale parameters N_0 and Λ and the shape parameters μ and γ. These parameters are either fixed or derived from the input variables as described in Tab. 1.

The lidar BSR is forward simulated using the lidar simulation capabilities of the Cloud Resolving Model Radar Simulator ~~(CR-SIM, Oue et al., 2020)~~(CR-SIM; Oue et al., 2020). The code has been slightly modified such that the configuration for the two-moment ICON microphysics can be used for one-moment microphysics following the relations in Tab. 1. Though CR-

Table 1. Configuration of modified Gamma distribution (Eq. 2) for liquid hydrometeors in ICON one and two-moment microphysical schemes.

scheme	hydrometeor	N_0	μ	Λ	γ	additional constrain
one moment (SRM)	cloud droplets	$f(q_c)$	8	$f(q_c)$	3	~~$q_{nc} = 2 \times 10^8\,kg^{-1}$~~ $N_c = 2 \times 10^8\,kg^{-1}$
one moment (SRM)	raindrops	$8 \times 10^6\,m^{-4}$	0	$f(q_r)$	1	
two moments (LEM)	cloud droplets	~~$f(q_c, q_{nc})$~~ $f(q_c, N_c)$	8	~~$f(q_c, q_{nc})$~~ $f(q_c, N_c)$	3	
two moments (LEM)	raindrops	~~$f(q_r, q_{nr})$~~ $f(q_r, N_r)$	2	~~$f(q_r, q_{nr})$~~ $f(q_r, N_r)$	1	

SIM can also simulate radar reflectivity, the Passive and Active Microwave ~~TRANsfer package (PAMTRA, Mech et al., 2020)~~ TRAnsfer package (PAMTRA; Mech et al., 2020) is used to forward simulate the radar as it offers a higher degree of flexibility.

The lidar forward simulations are used to detect the hydrometeors layer top and not for quantitative retrievals or estimates. Furthermore, as the airborne lidar is not affected by liquid collection on the telescope during raining conditions, there is no need to account for such effects~~. Thus,~~ , as there would be for a ground based lidar. As raindrops near cloud top are optically thin, if present in the models at all, we decided to simplify the forward simulation of the backscatter lidar and ~~assume that the raindrops are optically thin and~~ thus ignore their contributions. Therefore, the BSR is primarily a function of q_c as shown in Figs. 2 a and b. ~~One could further imagine that a raining cloud is always topped by small droplets contributing to q_c and that the lidar pulse hence would be scattered back by those cloud droplets, which would very likely have a BSR > 20, anyway and be thus identified as cloud top. However, this is not always true as some grid cells in ICON LEM with enough rainwater to generate a radar signal $Z >$ were simulated above or horizontally attached to a precipitating cloud (e.g., Fig. 4b, at 20:11). Therefore, it has to be noted, that the cloud amount reported below for the simulated lidar clouds slightly underestimates the cloudiness and cloud top height seen by a real backscatter lidar.~~

The approximated proportionality of the radar reflectivity to D^6 makes Z especially sensitive to larger raindrops. Therefore, q_r (and ~~q_{nr}~~ N_r) has to be considered in addition to q_c (and ~~q_{nc}~~ N_c) when simulating the radar signal. The size difference between cloud droplets and raindrops produces a two-modal relation between the total liquid water concentration $q_t = q_c + q_r$ and Z as it can be deduced from Figs. 2 c and d. The mode along a line of low q_t corresponds to grid cells that predominantly feature rainwater. In this mode, even low amounts of liquid water in the rain category produce a reflectivity that can only ~~by~~ be reached by cloud droplets with a three to four orders of magnitude higher cloud water content. Grid cells with such high q_c and no q_r align in a second mode parallel to the rain mode. A mixture of cloud and rain water accordingly results in an intermediate Z which populates the space of $q_t > \approx 10^{-5}\,kg\,kg^{-1}$ between the two main modes in Figs. 2 c and d. By setting the radar threshold to −20 dBZ, hardly any cloud-only grid cells in the lower right high-q_t mode can be detected by the simulated radar. This means ~~,~~ that in the ~~forward simulated~~ forward-simulated dataset, all lidar-detectable hydrometeors are from the ICON cloud category while the radar-detectable hydrometeors have to contain at least a small amount of water from the ICON rain category.

The ICON LEM uses a two-moment scheme including ~~q_{nc} and q_{nr}~~ N_c and N_r. Therefore, the forward simulation broadens the relation between the water content and the ~~forward simulated~~ forward-simulated signals (compare Figs. 2 a to b and c to d).

~~Especially the radar reflectivity of rain~~ The median of N_c is ~~amplified~~ 3.8×10^8 kg^{-1} with an interquartile range between 2.0 an 6.8×10^8 kg^{-1} for grid cells containing any cloud water. This means N_c is mostly larger in the LEM than in the SRM (fixed to $N_c = 2 \times 10^8$ kg^{-1}, Tab. 1), and that cloud droplets are smaller for the same q_c in the LEM. Thus, also the forward simulated signals are lower, such that slightly higher amounts of cloud water are required for a cloud to be detectable in the LEM than in the SRM. For a fixed water mixing ratio, a change by a factor of α in the number concentration ($N_c' = \alpha N_c$) will result to a change in the radar reflectivity in dBZ from Z to $Z' = Z - 10 \log_{10} \alpha$. Thus, if we double the N_c ($\alpha = 2$), the Z will reduce by 3 dB. By the same token, if we change the hydrometeor diameter by a factor of α ($D' = \alpha D$), then the radar reflectivity will be $Z' = Z + 30 \log_{10} \alpha$. Different to cloud water, the radar reflectivity of rain is in many cases amplified in the two-moment scheme compared to the one-moment simulation, such that also some grid cells with lower q_r are above the radar detection threshold. This indicates in general larger and fewer raindrops in the LEM than in the SRM for the same q_r as also depicted in Fig. 3.

5 Model – observation comparison

Observations and forward simulations of the SRM and LEM runs are used to ~~asses~~ assess the vertical structures of the shallow clouds by focusing on the boundaries sensed by different instruments. In the following, shallow clouds are analyzed in terms cloud top heights estimated from lidar and radar measurements as well as the radar echo base height. All heights in the different scenes are set in relation to the theoretical cloud base of an adiabatic thermal-plume-driven boundary layer cloud by setting the height in relation to the ~~lifted condensation level (LCL)~~LCL. First, a case study with example scenes from the observations and the LEM illustrates the approach. The case study is followed by the statistical analysis of the full datasets and the analysis stratified in the liquid water space to identify differences in microphysical processes.

To ease the following discussion, we define three layers in which the lidar and radar signals occur. Every signal below LCL is in the "precipitation" layer. Typically, only the radar base is in this layer. Clouds with their tops within 600 m above LCL are called "very shallow clouds" following the definitions by Vial et al. (2019). Vial et al. defined this mode in terms of an absolute top height below 1.3 km which corresponds to a similar height considering that the LCLs in the dropsonde, SRM, and LEM datasets in this study have typical heights of 720 ± 135, 763 ± 144, and 777 ± 121 m, respectively. Cumulus humilis is a typical representative of these very shallow clouds but in principle this class contains also small parts of deeper but slanted clouds. More active clouds can grow deeper than these very shallow clouds until they encounter the trade inversion and are forced to form a lateral outflow which is often perceivable as a stratiform layer. Stratiform remnants of such shallow convection can last for hours and thus much longer than the original convective core (Wood et al., 2018). We summarize all cloud signals above LCL + 600 m as "stratiform" mode, acknowledging also contributions from active cores. To limit the analysis to shallow clouds, an upper limit is set to 4 km above sea surface.

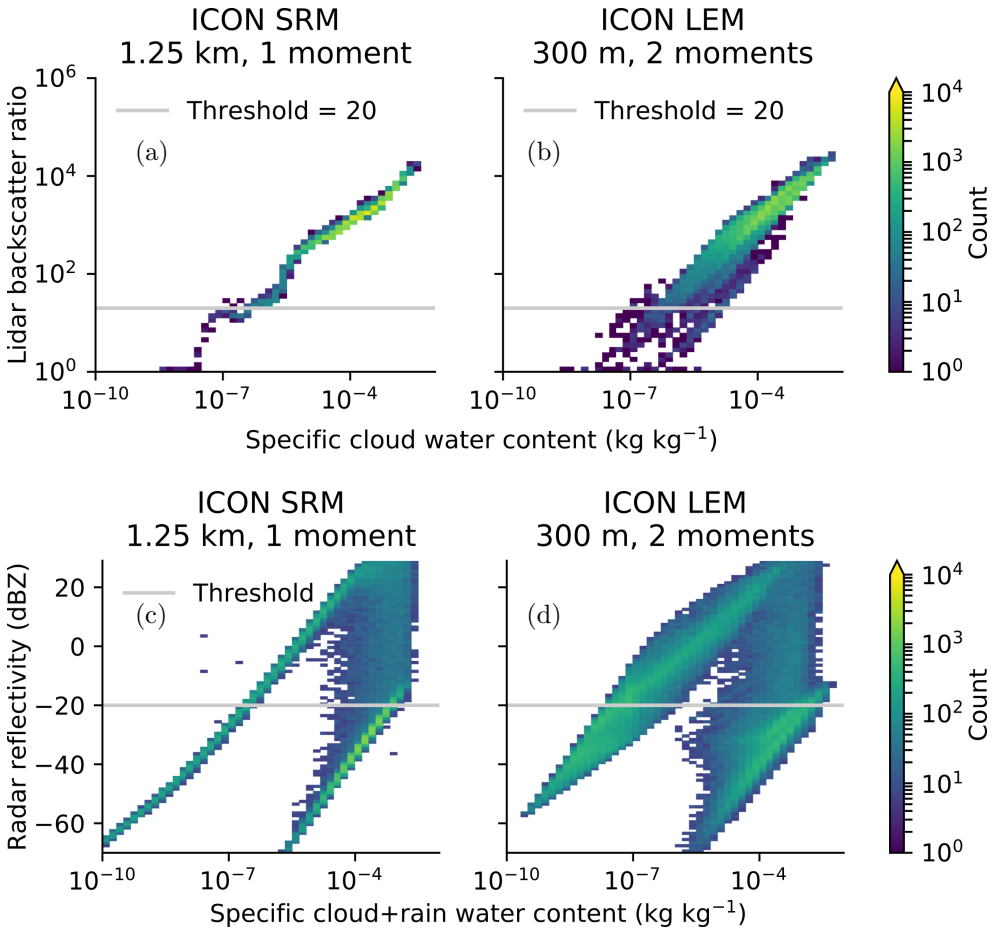

Figure 2. Simulated lidar and radar signals as a function of hydrometeor contents. CR-SIM and PAMTRA simulate the observable lidar and radar signals from drop size distributions in the one-moment ICON SRM and two-moment ICON LEM microphysical models. Signals are simulated without attenuation as they would be sensed at cloud top.

5.1 Case study

An example scene observed from HALO during research flight 5 is depicted in Fig. 4a. Here, several very shallow clouds close to the LCL were observed first, followed by a precipitating cloud with stratiform shallow anvil outflow. The shallow clouds were only detected by the lidar, whereas the precipitating cloud was detected by both the lidar and the radar. However, the lidar detected cloud top heights about 50 to 100 m, i.e., up to three radar range gates, above the ~~upper most~~ uppermost recorded radar echo. Also, a larger part of the ~~outflow~~ stratiform layer was visible to the lidar. Thus, we conclude ~~,~~that the precipitating shallow cumulus has a thin layer of very small droplets on top which are only seen by the lidar due to its higher sensitivity (compare Fig. 2).

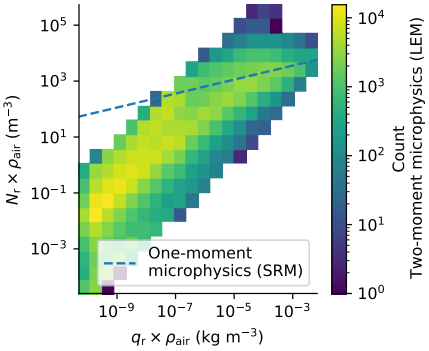

Figure 3. Relation of volumetric raindrop number concentration and mass mixing ratio for one- and two-moment microphysics.

A joint standard grid for the radar and lidar observations and forward simulations is used to facilitate additional analysis. A grid spacing of seven radar range gates is chosen, so that histograms are calculated as counts in 210 m high bins normalized by the bin width and the total number of cases in the total dataset. The histogram statistics in the right part of Fig. 4a summarize the detected cloud layers in the scene. The integral over the histogram equals the 2D shallow-cloud cloud ~~fraction~~ coverage detected by the respective sensor. In the particular scene depicted in Fig. 4, the lidar sees a cloud in about 73 % of the time, while the radar cloud ~~fraction~~ coverage is about 46 %. Note that the histograms depict the vertical distribution of detected cloud tops or base heights in a column and are therefore different from profiles of vertical cloud fraction. In the case of multi-layer clouds, ~~one layer is hidden by the other. To limit the analysis to shallow clouds, an upper limit is set to above sea surface~~individual layers could be hidden due to attenuation. Therefore only the uppermost cloud top and lowest base are considered. The histogram in Fig. 4a reveals the separation of the radar echo base into ~~non raining drizzle in the outflow~~ large non-raining droplets in the stratiform layer and precipitation that falls out of the cloud base at LCL. Note ~~,~~ that the lowest usable radar range bin is at about 100 m above the sea surface to avoid any surface clutter artifacts.

Figure 4b displays an example time series from ICON LEM which also includes precipitating clouds (beginning of the time series) and a few very shallow thermal-driven clouds (in the end). The cloud tops seen by the lidar and radar are mostly in the ~~upper~~ stratiform mode about 2 km above the LCL. The peak of the radar cloud top heights is about 400 m above most of the lidar cloud tops. This order is contrary to the observed case study ~~. The~~ and probably caused by evaporation of cloud droplets at cloud top as the higher reaching radar signal originates from grid cells at cloud top containing only rainwater but no cloud water. This can be seen by the pixels with a radar reflectivity signal above the lidar cloud top height, e.g., at 20:11. Here, raindrops might be transported out of the cloud core by wind sheer or turbulence. As only a few thin lidar-only-visible clouds near LCL are present in this scene, the mode of ~~lower~~ very shallow clouds is not ~~very pronounced~~ much pronounced in this example.

~~Two short scenes illustrate the information content gained by analyzing the vertical distributions of lidar- and radar-detectable cloud top, and radar-detectable cloud base heights. More sound findings on the relative occurrence of upper- and lower-mode clouds and their typical heights can be gained by applying this method to the full dataset~~

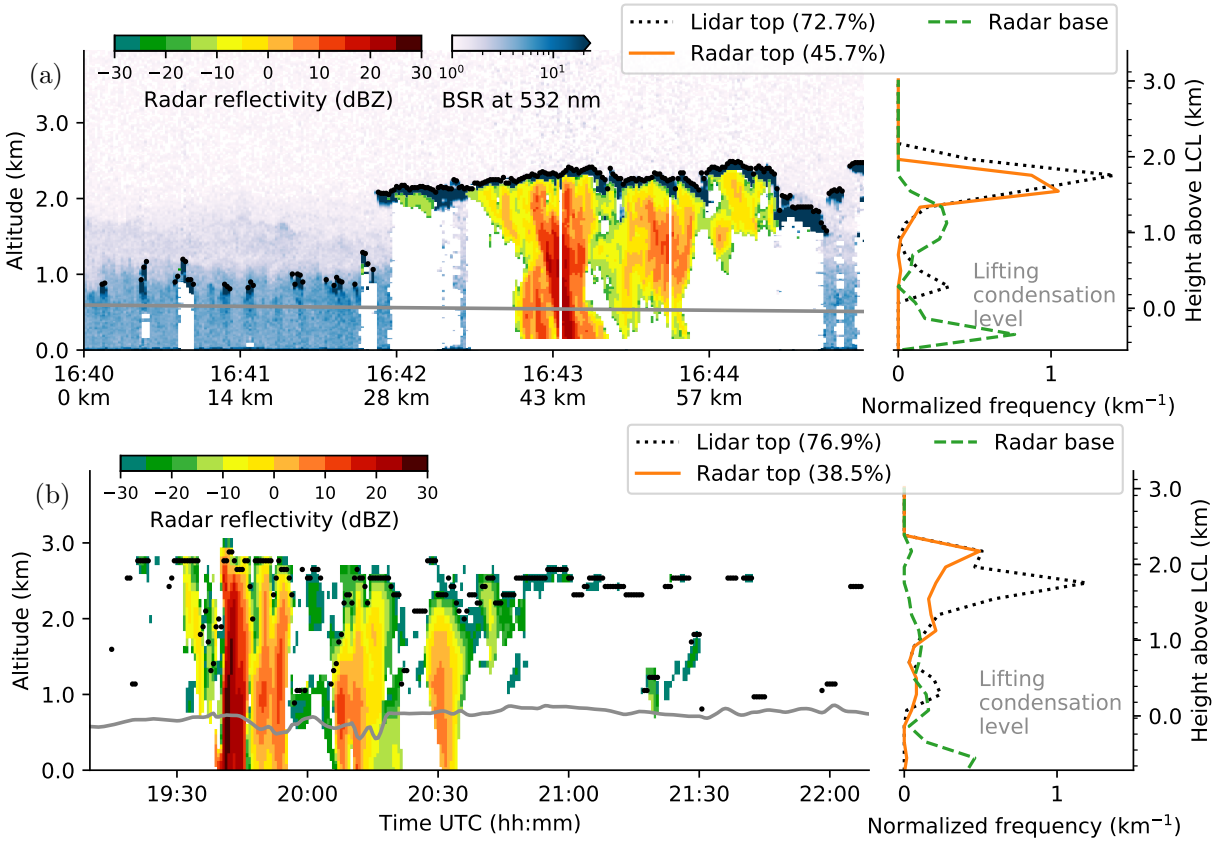

Figure 4. Case study time series of observed (a) and modeled (b) radar reflectivity, lidar-detectable cloud top height, ~~lifted~~ lifting condensation level (LCL) and their vertical distribution. Shallow 2D cloud ~~fraction~~ coverages detected by lidar and radar in each dataset ~~is~~ are given in each legend. Observations (a) are from research flight 5 on 2013-12-15 and also include lidar backscatter ratio (BSR) plotted below the reflectivity. Model simulation is from an ICON LEM meteogram station on 2013-12-16. The vertical distributions are normalized by the number of time steps in each scene.

5.2 Cloud statistics

~~To investigate whether the findings of the case study~~apply ~~generally~~After introducing and discussing the approach in a case study, all observations and simulations are jointly analyzed in this section. The histograms of the observed lidar cloud top heights (Fig. 5) reveal ~~, similar to the case study,~~ two modes of cloud top heights, similar to the case study. While the lower ~~one is~~ mode of very shallow clouds is centered at about 300 m above LCL, the ~~upper one is~~ stratiform mode is centered at

about 1.3 km above LCL. Frequency wise, ~~the upper modedominates over the lower mode by about~~ stratiform clouds (and active cloud cores included in the stratiform mode) were observed more than twice as often as very shallow clouds, when they were not hidden beneath such a stratiform cloud. This is contrary to the ground-based impression from the same region but other period when the shallow mode was clearly dominating (Nuijens et al., 2014). The lower mode of very shallow cumulus clouds on top of the well mixed boundary layer (Stevens et al., 2017) is very likely to be thermal driven and hardly produces any precipitation. The radar, however, observes in principle just one mode of top heights with its maximum at about 1.3 km above LCL, consistent with the upper lidar mode. But, similarly to the example in Fig. 4, the distribution is shifted slightly towards lower top heights than the lidar-visible cloud top distribution. Overall, the lidar sees clouds more than twice as often as the radar (43.2 vs. 18.2 %) due to its higher sensitivity that even responds to low cloud water contents of about 10^{-7} kg kg^{-1} (compare Fig. 2).

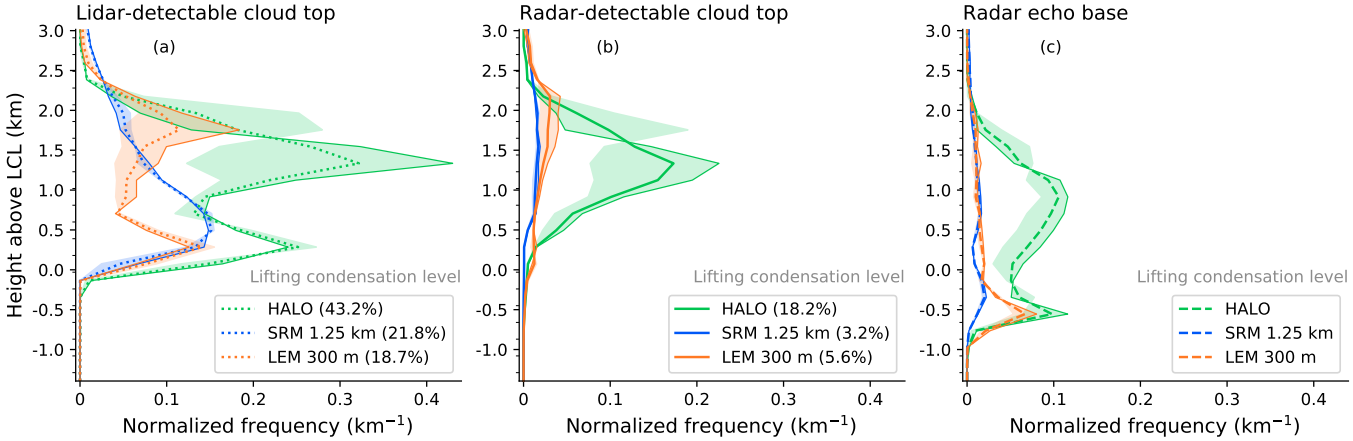

Figure 5. Cloud boundary statistics on all observed and ~~forward-simulated~~ forward-simulated lidar and radar signals: (a) lidar cloud top, (b) radar cloud top, and (c) radar echo base. Same thresholds for cloud detection are used for the observed and simulated lidar and radar signals. Height is in relation to the ~~lifted~~ lifting condensation level (LCL). Shadings depict western (bright edge) and eastern (dark edge) half of each dataset. The histogram bin edges are depicted as ticks on y-axis. ~~Shallow~~ Total shallow 2D cloud ~~fraction~~ coverage detected by lidar and radar in each dataset ~~is~~ are given in the ~~legend~~legends. This total cloud coverage can be derived from a cumulative version of this figure presented in the appendix Fig. C1.

We attribute the ~~upper~~ stratiform mode to shallow convection, precipitating clouds and their shallow anvil outflow. This interpretation is supported by the distribution of ~~reflectivity bases detected by the radar~~ radar echo bases. These bases are also bimodal with the upper mode about 400 m below the mode of radar top heights. ~~This upper mode of radar base heights~~ The upper radar base mode spans over the stratiform and very shallow layers. This upper mode is related to the outflow anvils and not-yet precipitating clouds in which the layer of radar-detectable hydrometeors is only a few hundred meters thick. The lower mode of radar base heights is below the LCL, i.e., in the precipitation layer and comprises clearly precipitating cases even if the precipitation occasionally evaporates before reaching the surface. The radar-detectable cloud depth distribution can also be

seen in Fig. 6 (a) as distance to the main diagonal. The joint histogram of radar base and top height confirms that the clouds observed from HALO are often either about 200 to 400 m thin (along the diagonal) and at about 1 km above the LCL or they precipitate (on the left) with similar cloud top heights.

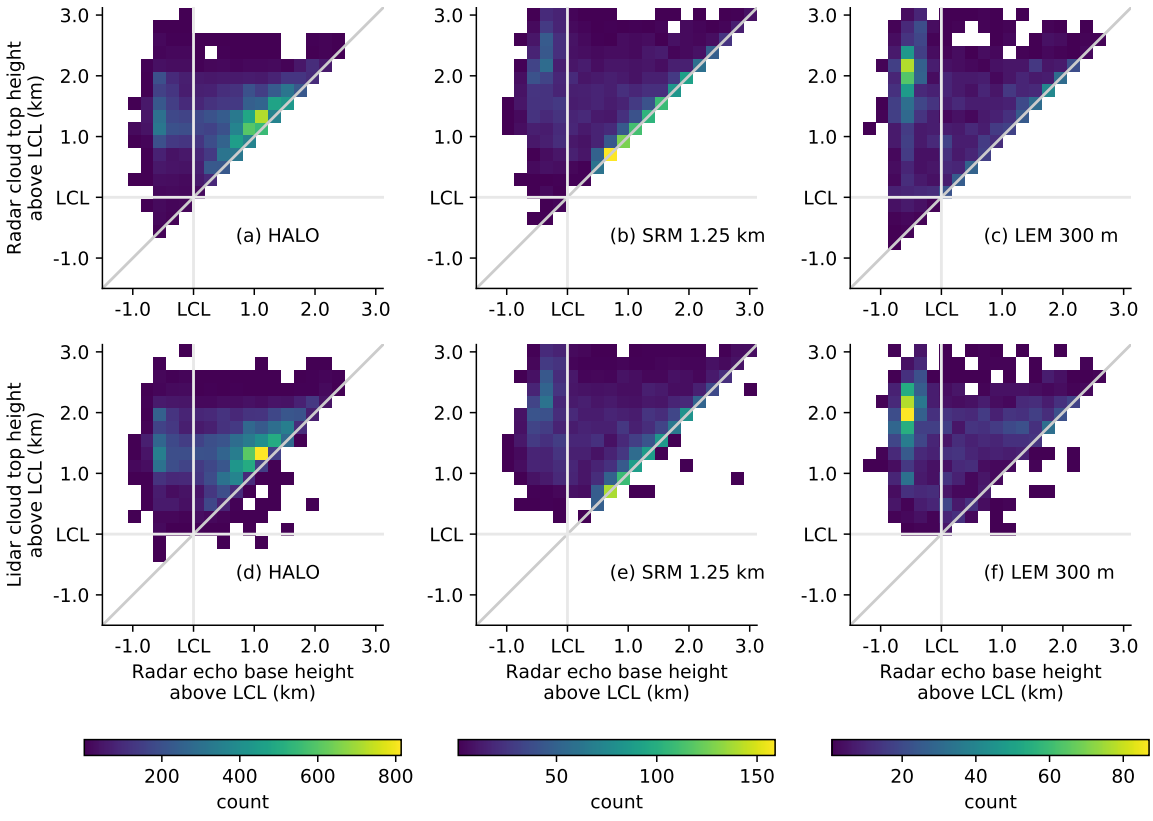

Figure 6. Relation of radar-detectable cloud top height (a-c) and lidar-detectable cloud top height (d-f) above LCL versus radar echo base height above LCL. Based on HALO observations (a, d), 1.25 km SRM dataset (b, e), and 300 m LEM dataset (c, f).

A deepening of the ~~shallow~~ cumulus cloud layer in accordance with a sea surface temperature increase is expected from the stratocumulus decks in the east tropical Atlantic to the cumulus regime in the west (e.g., Wyant et al., 1997). A temperature increase of about 2 K from east to west in the flight area motives a separation of our data by longitude. The deepening of the cumulus cloud layer can be seen in the HALO observations as the lidar and radar detect the ~~upper~~ stratiform mode about 400 m higher in the observations west of 51.5° W than east of it. ~~However, the frequency and height of the lower mode of the~~ This deepening probably caused the frequency reduction of the stratiform mode in the western half compared to the shallower eastern half. Such relation between deepening of the cloud layer and reduced formation of stratiform clouds was also shown in an LES study by Vogel et al. (2020). However, the relation seems opposite to the positive correlation between thin stratocumulus cloud fraction and planetary boundary layer depth observed with satellites on monthly timescales in the marine stratocumulus to

cumulus transition by O et al. (2018). In contrast to the deeper and stratiform clouds, the frequency and height of very shallow lidar-visible clouds is almost the same in the western and eastern parts, which also agrees with Vogel et al. (2020).

~~Better~~ More pronounced than in the case study, a bimodal distribution of cloud top heights is also present in the dataset of all available ICON LEM ~~data~~ output (Fig. 5). The ~~lower mode~~ mode of very shallow clouds behaves very similar to the observations. It has its maximum frequency at the same height and is also only detectable with the sensitivity of the lidar~~only~~. However, the frequency of this mode and the overall ~~shallow cloud fraction~~ total cumulus cloud coverage is only half the observed cloud ~~fraction~~ coverage (18.7 vs. 43.2 %). The height of ~~upper~~ stratiform mode is about 400 m above the observed ~~upper~~ stratiform mode of the whole ~~dataset~~ HALO dataset, but matches the ~~upper~~ stratiform mode of the western part of observations very well. This is in line with the fact ~~,~~that the LEM is only represented by meteograms in the western flight area. ~~The shallow clouds~~ Clouds are detected by the ~~forward simulated~~ forward-simulated radar in only 5.6 % of the LEM scenes compared to 18.2 % in the observations. In agreement with the observations, the radar cloud tops are mostly modeled in the ~~upper~~ stratiform layer of the LEM, but with the maximum higher than the lidar cloud tops – similar to the example discussed before. The distribution of the modeled radar signal base heights indicates ~~,~~that most clouds in the LEM are precipitating if they are visible to the radar. In contrast to the observations, non-precipitating radar-detectable clouds in the LEM develop with a wider range of depths as shown in Fig. 6 (c).

The ICON SRM represents the clouds rather differently than the LEM. The clouds visible to the lidar generally form one broad mode with the most frequent lidar cloud top heights around 500 to 700 m above LCL. The frequency of ~~shallow~~ cloud tops decreases with altitude until they disappear at 2.6 km above LCL, which is similar to the other two datasets. The clear separation of cloud tops into two layers, however, is not evident in contrast to the observations and the LEM. While a double layer structure could be seen on individual days in the SRM data (not shown), ~~this is likely caused by the significantly varied altitude of the upper layer~~it does not show up on average due to the strong variation in altitude between different days. Radar-detectable clouds and precipitation are also modeled but only in about 3 % of the SRM scenes, which is much less than observed (18 %) and in the LEM (6 %). The radar top height distribution, however, has a similar shape ~~as~~ compared to the observed radar clouds. Even if less frequent, the relative distribution of radar signal base heights in the SRM is similar to the observations with one peak between LCL and LCL + 1 km ~~and~~ (spanning over the stratiform and very shallow layers) and the second peak few hundred meters below LCL ~~. The distribution of the upper edge of~~ in the precipitation layer. Like in the observations, there is a similar mode of thin radar-detectable clouds (Fig. 6b, along the main diagonal), however, at generally lower heights, and precipitating clouds with deeper cloud tops compared to the ~~upper mode is relatively more gentle than in the~~ observations. The clear difference of observed ~~outflow~~ stratiform and precipitating cloud layer heights between the eastern and western part of the data is not pronounced in the SRM data, even though the coverage of the ~~model fits better to the observations~~ SRM matches the observations better than the LEM. This indicates ~~,~~that the shallow convection and outflow process ~~is not modeled as seen during the field experiment~~as well as the formation of a stratiform layer are modeled differently than observed.

~~Bimodal distributions of~~ There are a couple of observed cases (right of the diagonal in Figures 6 d-f) with the lidar-detected cloud top heights ~~were also observed from space~~ below the radar base height. These signals relate to lateral raindrop transport out of the precipitating core with a patch of cloud beneath. Such cases also occur in the LEM dataset and less frequently in the

SRM dataset. The smaller grid spacing in the LEM could be favorable for a more likely lateral transport of raindrops into a neighbor grid cell in the LEM compared to the SRM.

We can compare the observed and modeled bimodal cloud top distribution to the those found by Genkova et al. (2007) and Leahy et al. (2012). The ~~former identified cloud top height maxima~~ modes identified by the former at 650 and 1500 m above sea level ~~in an area similar to this study from about 150 scenes between September 2004 and March 2005. Both modes~~ seem to be lower than the ones observed in the present study ~~, considering that the heights of the LCL is in the dropsonde, SRM, and LEM datasets in this study have means and standard deviations of , , and respectively~~with the LCL height at about 750 ± 150 m, even though they studied a similar domain in a similar season. However, Genkova et al. (2007) denote a vertical uncertainty of 250 to 500 m. ~~Leahy et al. (2012) observed the upper layer around , the lower at about above the sea surface in tropical Pacific trade wind cumulus (S, W). These values~~ The modes observed in similar cloud regimes by Leahy et al. (2012) at around 800 and 2000 m are closer to the values in the present study, even though the ~~similar cloud regimes are investigated in different areas~~observations took place in a different domain over the tropical South Pacific.

To conclude: Bimodal lidar cloud top height distributions were observed and their clear separation is well reproduced by the LEM but not by the SRM. The lower mode of ~~thermal driven~~ very shallow thermal-driven clouds is closely above the LCL, while the upper mode is closely below the trade inversion (Stevens et al., 2017), i.e., about 1.3 km higher up. The SRM, however, shows one prominent mode of cloud top heights with its maximum at rather lower heights. However, the SRM also produces deeper clouds with their frequency decreasing with height. Neither model reproduces the often observed radar echoes embedded in the non-precipitating upper stratiform ~~outflow~~ mode. To shed light on the conditions under which these clouds are ~~simulated infrequently~~ underrepresented compared to observations, comprehensive LWP observations refine the statistics in the next section.

5.3 LWP classes

The stratification of the observations and model ~~data~~ output into different LWP classes can give more detailed insight into the regimes under which the models perform better or worse. LWP classes are chosen to represent barely detectable clouds ($< 10\,\mathrm{g\,m^{-2}}$), clouds which are not completely optical thick ($< 50\,\mathrm{g\,m^{-2}}$), classical cumulus clouds ($50\,\mathrm{g\,m^{-2}} <$ LWP $< 100\,\mathrm{g\,m^{-2}}$), thicker clouds which are still considered in satellite retrievals ($100\,\mathrm{g\,m^{-2}} <$ LWP $< 300\,\mathrm{g\,m^{-2}}$) (Wentz and Meissner, 2000), and even more water bearing clouds ($300\,\mathrm{g\,m^{-2}} <$ LWP $< 1000\,\mathrm{g\,m^{-2}}$). An overview of cloud top heights and radar base in the different datasets and LWP ranges is presented in Fig. 7 and discussed in the following. ~~To ease this discussion, we define three layers in which the lidar and radar signals occur. Every signal below LCL is in the layer. Typically, only the radar base is in this layer. Signals within above LCL are in the so called cloud layer. Signals above LCL + are called which also includes the tops of raining clouds as depicted in Fig. 4.~~

It is remarkable ~~,~~ that high cloud ~~top heights in the outflow~~ tops in the stratiform layer were often observed by the lidar under low LWP conditions (below $10\,\mathrm{g\,m^{-2}}$). ~~However, the lidar top heights in the outflow layer are relatively more frequent, when extending the class from LWP <~~ Such clouds likely correspond to thin "veil" clouds frequently observed near the upper boundary layer, i.e., below the trade inversion, in the stratocumulus to cumulus transition by Wood et al. (2018) and

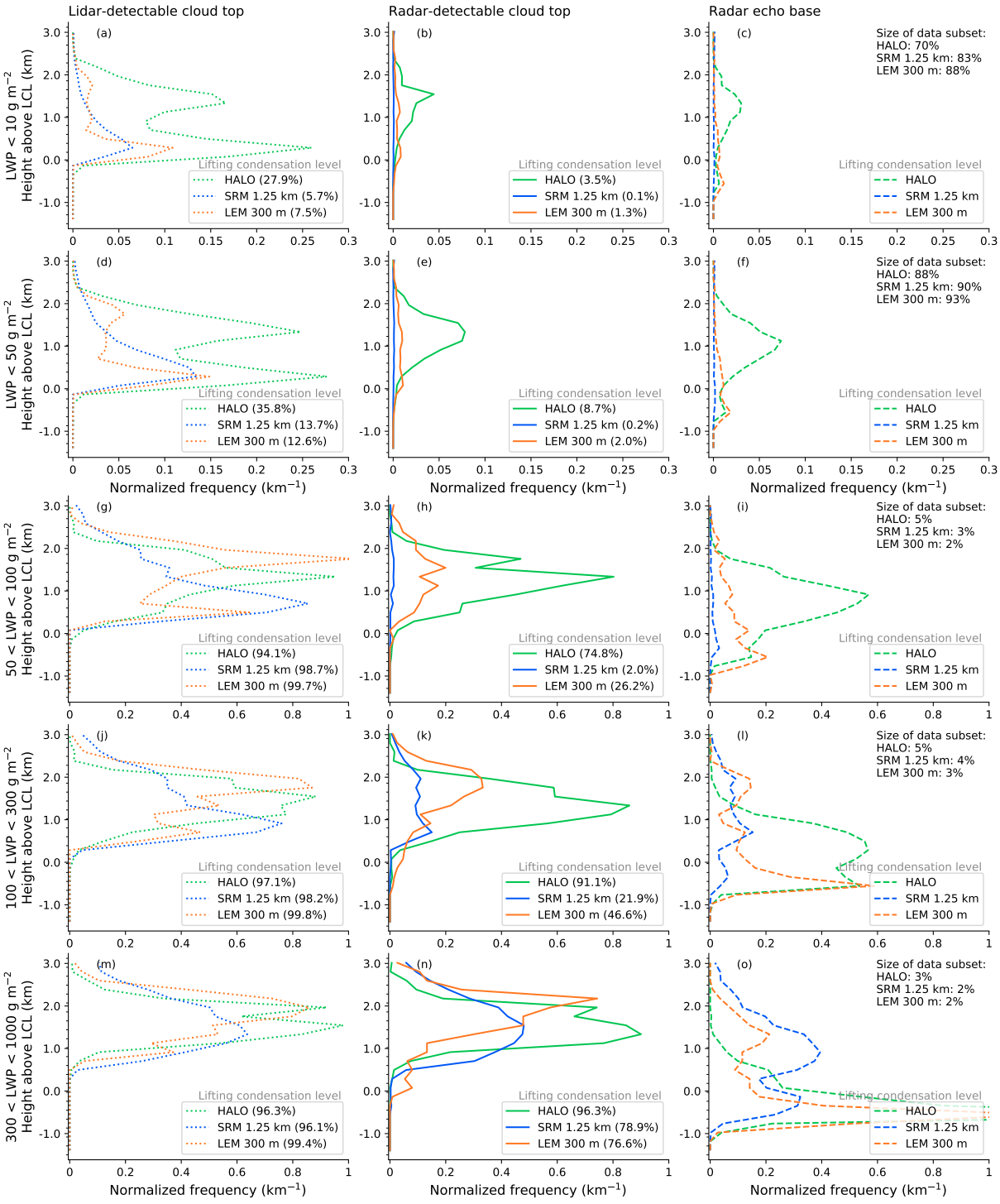

Figure 7. ~~Like~~ Similar to Fig. 5 but classified by liquid water path (LWP). Columns represent lidar cloud top, radar cloud top, and radar base of observed and ~~forward simulated~~ forward-simulated lidar and radar signals. Rows represent different LWP ranges. Note the different x-scale used in the upper two rows.

O et al. (2018). They report on geometrically and optically thin clouds with low droplet number concentration (about 5 cm^{-3}) but relatively large droplets with radii ranging from 15 to 30 µm. Droplets of such sizes are large enough to provide a radar reflectivity above the detection threshold.

Extending the LWP class from 10 g m^{-2} to ~~LWP <~ 50 g m^{-2} .~~ includes more additional lidar-detectable stratiform cloud coverage to the statistics than very shallow cloud coverage. This means even more veil clouds are included, which were estimated to have a typical LWP of about 25 g m^{-2} (Wood et al., 2018). In all cases with LWP < 50 g m^{-2}, the ~~outflow~~ stratiform layer was observed about 1.5 times more often by the lidar than the ~~thermal layer clouds.~~ layer of very shallow clouds, which is a bit more often than in the LEM and SRM (see also Fig. 8a).

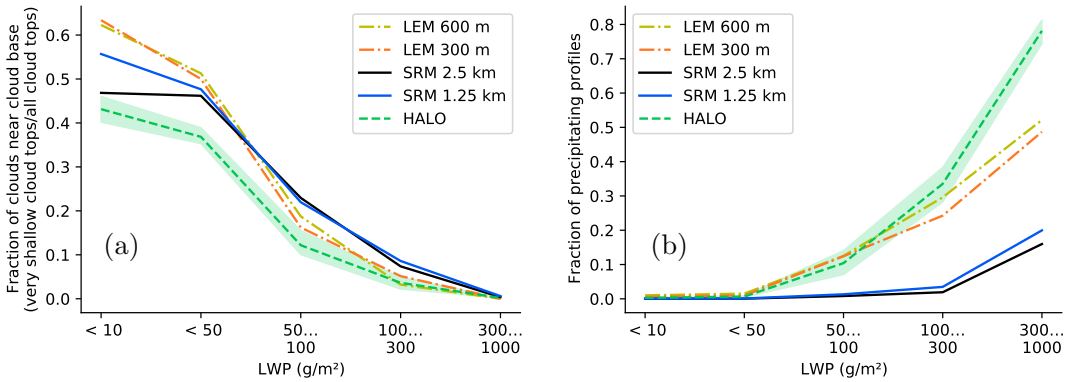

Figure 8. Fraction of very shallow clouds and precipitating clouds as a function of LWP. (a) Fraction of clouds near cloud base defined as ratio of very shallow clouds (lidar-detectable cloud top within LCL and LCL + 600 m) to all shallow clouds (lidar-detectable cloud top height < 4000 m). (b) Fraction of precipitating profiles defined as ration of radar echo below LCL to all profiles. Shading of the HALO dataset represents the uncertainty of the LWP retrieval.

In general, it is no surprise that the ~~distribution~~ distributions of lidar cloud tops in low LWP conditions (< 50 g m^{-2}, Fig. 7a and d) ~~is~~ are similar to those of the whole dataset (Fig. 5a), as most of the scenes have a low LWP ~~. For example, the statistics of the lidar-detectable top-height of scenes with LWP < in the SRM with only one mode and in the LEM with two modes is in general the same as discussed in the previous Sect. 5.2. However, the classification by LWP shows the trend in both the observations and the LEM data that outflow or precipitating clouds are more likely with higher LWP. Likewise, the thermal mode disappears in the observation and LEM datasets for higher LWP (>).~~

(compare Fig. 11b in Jacob et al. (2019) for the HALO dataset). The statistics of radar-detectable cloud top and base heights in scenes with LWP < 10 and 50 g m^{-2} in Figs. 7b, c, e, and f are different ~~to~~ from the overall statistics (Fig. 45b), as the radar is often not sensitive enough to detect clouds with such little LWP. The lidar-detected clouds are about seven (three) times more frequent than those detected by the radar on HALO in scenes with LWP < 10 g m^{-2} (< 50 g m^{-2}). In the LEM simulations, this ratio is about five for both LWP limits. The relative smaller increase ~~of~~ in radar-detectable clouds means that clouds in the LEM with 10 < LWP < 50 g m^{-2} probably ~~consist out of too~~ have only small droplets and thus miss a radar-detectable drizzle component. About a twelfth of the observed radar clouds with LWP < 50 g m^{-2} are categorized as precipitating, while the LEM

depicts half of them as precipitating. No statement on the SRM precipitation fraction can be made as only 0.2 % (i.e., less than 200 profiles) of the SRM scenes with LWP < 50 g m^{-2} show radar-visible cloud tops below 4 km at all.

The lidar detected a cloud in 96 % of the observed scenes with LWP > 50 g m^{-2}. In the remaining cases, the lidar ~~probably~~ either missed clouds with only ~~partially~~ partial coverage in the microwave radiometer footprint (~~≈ 1 km). Further~~≈ 1 km) or few but large raindrops were horizontally transported out of the cloud core, such that they are only visible to the microwave radiometer. Likewise, not all clouds in scenes with LWP > 50 g m^{-2} contained radar-detectable hydrometeors. This difference between lidar- and radar-detectable clouds with LWP > 50 g m^{-2} is in principle also reproduced by both models. In the observations, about four of five clouds detected by the lidar were also seen by the radar in the 50 to 100 g m^{-2} LWP class. However, only a quarter of the lidar-detectable LEM clouds are also detectable by the simulated radar. The ratio in the SRM simulations is even smaller. The radar base on the other hand shows ~~,~~ that the LEM models about half of the radar-detectable clouds as precipitating, while precipitation was only observed for a quarter of the observed radar clouds with 50 < LWP < 100 g m^{-2} ~~.~~ (compare Fig. 6).

In scenes with LWP between 100 and 1000 g m^{-2}, the radar-detectable clouds in both models form two groups. They either precipitate or ~~form an outflow like structure~~ belong to the stratiform layer with a base clearly above the LCL. Such a separation was not observed from HALO. ~~In the observations, about a third of the 100 to clouds precipitate, while most others have base heights within above LCL. In the observed dataset with LWP > , about four fifths precipitate. The single mode of lidar-detectable cloud top height in the SRM increases with LWP. Finally, radar-detectable clouds appear more frequently in the SRM when LWP > . However, in these cases only a quarter of the radar-visible SRM clouds actually show a precipitating signal below the LCL.~~

~~The mode of~~ The mode of non-precipitating radar-visible clouds under high LWP (> 300 g m^{-2}) conditions in both models~~can also~~ , however, can be explained by heavy clouds in the model consisting purely of cloud droplets~~only~~. A model cloud with LWP > 300 g m^{-2}, for example, which is 300 m deep must on average contain at least ~~about~~ 10^{-4} kg kg^{-1} liquid. This means that such ~~cloud doesn't need any contribution from raindrops to be radar-detectable~~ a cloud is radar-visible without containing any raindrops (compare Fig. 2). ~~However, such heavy non-precipitating clouds are observedrather infrequently.~~

Figure 8 summarizes the dependence of the very shallow cloud mode and precipitation on the LWP classes. It includes the output from a coarser nest of both ICON model versions to shed light on the influence of grid-resolution versus model configuration. In general, all models and resolutions simulate the observed and expected reduction of very shallow clouds with increasing LWP (Fig. 8a). However, relatively more very shallow clouds were modeled than observed. The likelihood of precipitation with higher LWP (Fig. 8b) and hence also in deeper clouds, separates the datasets. The increase in precipitation up to LWP ≤ 300 g m^{-2} in both LEM resolutions corresponds better to the observations than the SRM outputs do. For very high LWP, too little high-LWP clouds precipitate, which could be caused by a too weak autoconversion process in the LEM. The additional outputs from coarser resolved LEM and SRM nests suggest the different resolutions being less important than the choice of parametrizations, vertical levels and microphysical scheme. Differences in sampling of the LEM and SRM outputs, however, also influence the results, but sensitivity testing the sampling area and considered dates showed no major influence as analyzed in appendix C.

The stratification of the data by LWP shows that both models cannot represent non-precipitating but radar-visible drops that were observed under all and especially low LWP conditions. These drops are probably larger than those represented by the Gamma distributions of the cloud hydrometeor class in both models. Radar-visible model clouds precipitate more often than observed, which means they ~~consist of~~ contain already very large droplets, but the fraction of radar-visible clouds is in general too small. Non-precipitating clouds, consisting ~~presumably~~ of cloud-type hydrometeors only ~~,~~ were produced by both models under high LWP conditions ($> 300\,\mathrm{g\,m^{-2}}$), but such cases were not observed.

6 Summary and conclusions

Observed statistics of hydrometeor profiles and liquid water path (LWP) of oceanic shallow cumulus clouds are compared against those produced by two high resolution models of the ICON family. The observations and model runs were part of the NARVAL experiment over the tropical Atlantic east of Barbados in the dry winter season 2013. The instruments were operated from the research aircraft HALO at an altitude between 13 and 14.4 km in a ~~nadir pointing~~ nadir-pointing orientation. The two models ~~from the ICON family are the so called~~ are the so-called storm resolving model (SRM) and the ~~large eddy~~ large-eddy model (LEM)~~with~~ . Primarily, outputs with grid-spacings of 1.25 km (SRM) and 300 m ~~grid-spacing, respectively~~(LEM) are analyzed, but additional outputs are coarser grid-spacings (2.5 km and 600 m) are considered as well. The SRM resolves the shallow cumulus layer ~~with 10 to 13 layers~~(700 – 3000 m above sea level, (asl.)) with 14 levels, while the LEM has ~~23 to 28~~ 24 levels in that layer.

The upper part of the hydrometeor profile is characterized by radar and lidar observations, while the lower part of the hydrometeor profile is characterized by the radar only~~due to lidar extinction~~. The LWP is retrieved from microwave radiometer measurements. When looking at the high occurrence of low-LWP scenes in the models (83 and 88 % below $10\,\mathrm{g\,m^{-2}}$ ~~,~~ in SRM and LEM, respectively, Fig. 7), it becomes evident that common sensitivity thresholds for the instruments and models are urgently needed to assess clouds in this regime. Thus, forward simulations of the radar and ~~liar~~ lidar observations using instrument specific sensitivity thresholds and relationships between the observables and the model output are used to allow an apples-to-apples comparison between the HALO observations and the ICON model output (Lamer et al., 2018). A lidar backscatter ratio threshold of 20 suggested by Gutleben et al. (2019) is applied to clearly distinguish between backscatter from dust aerosols and cloud droplets. A comparison of the airborne measurements to ground-based radar records reveals a reliable radar reflectivity detection threshold of −20 dBZ for the airborne radar over the full column. The forward simulations show that most clouds with $q_c > 10^{-7}\,\mathrm{kg\,kg^{-1}}$ in the model are detectable with the respective backscatter lidar threshold. The radar, in contrast, is primarily sensitive to the "rain"-category ~~hydrometeor~~ hydrometeors in ICON. Only the highest amounts of liquid q_c in a cloud-water-only cloud in the model are detectable by the radar.

The observations reveal ~~two prominent modes~~ a bimodal distribution of cumulus cloud top heights separating the clouds into ~~two layers.The lower~~ a mode of very shallow cumulus, defined from the lifting condensation level (LCL) to LCL + 600 m, and a stratiform layer, defined from LCL + 600 m to 4 km asl.. The very shallow mode of cloud tops relates to shallow, non-precipitating boundary layer clouds reaching up to a few hundred meters above the ~~lifted condensation level (LCL). The~~

~~upper~~ LCL. The stratiform mode is mostly driven by shallow moist convection and also contains ~~stratiform shallow outflow anvils~~ detrainment clouds, which are often formed from the outflow of active shallow cloud anvils at around 1.3 km above LCL. The ~~lower~~ very shallow mode consists of mostly thin water clouds that are best seen by the backscatter lidar and are frequently missed by the radar. In contrast, ~~the upper mode clouds~~ clouds in the stratiform mode contain more and larger droplets that scatter sufficient microwave radiation to be detected by the radar in addition to the lidar. Overall, the ~~upper~~ stratiform mode was observed ~~more frequently, but both modes are similarly~~ 2.5 times as often as the very shallow mode and was in general more frequent in scenes with ~~little condensate (LWP <~~ higher LWP (> 50 g m^{-2}) than in those with lower LWP. In the ~~outflow~~ stratiform layer, the lidar detected the cloud tops slightly higher than the radar. This indicates that small particles with low radar reflectivity are present at the upper part of the ~~outflow~~ stratiform layer. Higher LWP values are associated with more precipitation echoes below the LCL and with deeper ~~outflow layers~~clouds, even though already 10 % of the scenes with medium LWP (50 – 100 g m^{-2}) showed precipitation. Also, a clear trend with higher cloud tops in the ~~upper~~ stratiform mode in the western part of flight tracks is observed that is probably related to higher sea surface temperatures in that area enforcing convection.

The bimodal cloud top height distribution is reproduced by the LEM, although the total cloud fraction is lower than observed. The radar forward simulations ~~suggest that the LEM produces less large particles in the outflow regimes~~show that too few large particles or in general too small cloud droplets in the stratiform regime are produced in LEM. The observed increase of radar-detectable clouds between LWP of 10 and 50 g m^{-2} is not reproduced by the LEM. This is consistent with the overall trend of the models that produce ~~smaller than observed particle sizes~~to few large particles. However, the LEM describes more of the radar-detectable clouds as precipitating than observed. This indicates that large radar-visible drops probably ~~can not~~ cannot be kept long enough in the model cloud layer before falling out. An observed cloud layer deepening with LWP can be also found in the LEM.

Different than the LEM, the SRM produces no clear separation between the two cloud layers. Cloud tops are typically at 500 to 700 m above LCL. Small differences in the warm autoconversion (AU) parametrizations might be a reason for the reduced frequency of deeper shallow clouds. The AU formulation is similar in the LEM and the SRM, but as the SRM cloud droplet number concentration ~~q_{nc}~~ N_c is constant (Table~~.~~ 1) but smaller than the average in-cloud ~~q_{nc}~~ N_c in the LEM~~(not shown)~~ , and as the AU rate increases with decreasing ~~q_{nc}~~ N_c (Seifert and Beheng, 2001, eq. 16), the AU in the SRM is expected to be stronger on average. Therefore rain could form quicker in the SRM and thereby reduce the average cloud life time, cloudiness, and also cloud top height. Indeed, especially the radar-visible cloud top heights of the LWP heavy clouds in the SRM are in general lower ~~then~~ than in the LEM (Figs. 7k and 7n). One could hypothesize further that a faster ~~warm precipitation cycle~~ warm-precipitation-cycle reduces the strength of the shallow convection, ~~so than in consequence, less clouds would reach the tropical inversionlayer, which could~~ such that fewer clouds reach the trade inversion, which would force the cloud to create the shallow outflow in the stratiform layer, that is produced too rarely by the SRM~~too seldom~~. However, there are other differences between the LEM and SRM that could contribute to differences in cloudiness and rain production. For example, the lack of a clear gap in the cloud top frequency distribution might be also due to the lower vertical resolution of the SRM ~~with 10 to 13 layers in the shallow cumulus layer (compared to 23 to 28 in the LEM) as the gap would require that~~ which would require

always the same few model layers ~~contain no cloud top.~~ to be cloud free. Furthermore, higher horizontal resolution influences the cloud formation and resolves buoyancy production better. However, the models and model outputs used here indicate a stronger dependence of cloud production on the model setup than on the resolution as shown in appendix C. The clearly observed east-west difference in the height of the ~~upper~~ stratiform cloud layer is only weak in the SRM. This indicates that processes of the precipitating ~~shallow convection~~ shallow-convection cumulus clouds are not fully represented in the SRM. The SRM cloud distribution is rather insensitive for different LWP classes except for a cloud deepening and precipitation increase with increasing LWP. This study primarily considers the grid-resolved clouds in the SRM. This might be an unfair comparison as the SRM also contains a diagnostic scheme for sub-grid-scale cloudiness used in the radiation calculations. Thus, the additional sub-grid-scale cloudiness is briefly assessed in the appendix B. In summary, clouds modeled from diagnostic equations would moderately increase the SRM cloudiness, but would not alter the vertical structure significantly, i.e., the diagnosis does not solve the missing cloudiness in the ~~outflow~~ stratiform layer.

Both models show clearly non-precipitating radar-visible clouds with LWP > 300 g m^{-2} which were not observed in that way and probably come from very high amounts of pure cloud water. In other cases, both models tend to produce precipitation that is also detectable below LCL once the cloud is visible to the radar and it seems that large radar-visible but ~~just slowly sedimenting~~ only slowly-sedimenting non-precipitating drops like in drizzle are missing. This is probably due to the size constraint in the ICON microphysics (Seifert and Beheng, 2001), that implies a threshold between cloud PSD and rain DSP at 40 µm, i.e., cloud PSD is assumed to not contain a significant number of droplets with diameter larger than this threshold. Our observation of larger but non-precipitating particles is in line with findings by Siebert et al. (2013) and Wolf et al. (2019) who observed cloud droplet effective radii on the order of this threshold in the same region but in generally moister months, i.e., they also note the principle presence of large cloud droplets.

Finally, it has to be noted that the available datasets have a great spatiotemporal overlap but do not match perfectly. The consequences of this are probably less severe than they would be for example in the mid-latitudes, a region that is heavily influenced by synoptic systems, because the study area and period is characterized as mostly undisturbed (Vial et al., 2019) and the variation from flight to flight in the winter season is limited (Jacob et al., 2019). ~~Nevertheless, the~~ The exact choice of domain and dates used in this study are analyzed in the appendix C by taking subsamples from the 1.25 km SRM dataset, but no significant impact on the cloud statistics can be found. The methods presented in this study show high potential to benchmark realistically driven ~~large eddy~~ large-eddy simulations. Even ~~with slightly different underlying meteorological statistics the~~ if the matching between model and observations could be improved in future studies the analysis provides insight into processes that are well represented by the models and which phenomena are difficult to model with the respective setup. ~~However, absolute numbers of cloud frequencies should be interpreted carefully.~~

Enhanced observations with several research aircraft, vessels, and autonomous platforms and coordinated model applications during the ~~upcoming~~ recent EUREC^{4}A field study in early 2020 (ElUcidating the Role of Cloud-Circulation Coupling in Climate Bony et al., 2017) will provide an even more comprehensive view on the trade wind shallow cumulus clouds. For that, the methods presented here are ready to by applied to future EUREC^{4}A studies. Also, cloud-chasing ship-based observations can observe individual cloud cycles including the transition from pure cloud to drizzle onset and probably rain production,

while airborne observations survey the cloud field to report on the representativeness of the in-detail studied cloud. As shallow cumulus clouds also ~~will~~ were be probed in-situ in addition to the remote sensing setup used in this study, a closer look into the drop size distributions in the ~~outflow~~ stratiform layers will be enabled in upcoming studies.

Code and data availability. The source code of CR-SIM was made available by Oue et al. (2020) at https://www.bnl.gov/CMAS/cr-sim.php (last accessed online: Nov. 6, 2019). The PAMTRA source code was made available by Mech et al. (2020) at https://github.com/igmk/pamtra/ (last accessed online: Nov. 6, 2019). The airborne radar and dropsonde can be found under https://doi.org/10.1594/WDCC/HALO_measurements_2 (Konow et al., 2019). The LWP retrieval data from the HAMP microwave radiometer can be found under https://doi.org/10.26050/WDCC/HALO_measurer (Jacob et al., 2019). The BCO data are accessible to the broader community through Stevens et al. (2015). The ICON SRM and LEM outputs were produced by Klocke et al. (2017) and made further public by Vial et al. (2019). The ICON SRM was run using revision "28436M" of the "icon-nwp/icon-nwp-dev" branch (Klocke et al., 2017). The ICON LEM was run using the ICON release 2.3.00 (Stevens et al., 2019).

Appendix A: Radar sensitivity

To estimate the practical sensitivity limit of the HALO radar observations, HALO radar statistics are compared to ground-based measurements taken at the BCO. The BCO radar operates at the same Ka-band frequency as the airborne radar, but has a better sensitivity due to a larger antenna and longer integration time (Lamer et al., 2015). Therefore, the lower MDS of the BCO radar offers the opportunity to assess the practical sensitivity limit of the HALO radar.

A comparison can only be made on a statistical basis as the BCO and HALO radars do not sample the same volume. To avoid statistical effects of the diurnal cycle identified by Vial et al. (2019), BCO data are only considered roughly during the time when HALO was flying, i.e., between 12:00 and 21:00 UTC (8:00 and 17:00 local time) on the 8 flight days.

The higher BCO radar sensitivity compared to the HALO radar is notable in the height-resolved reflectivity histograms in Fig. A1. The BCO radar frequently measures reflectivity signals down to −70 dBZ at around 500 m with a clear frequency maximum below 1 km for Z up to −20 dBZ. Klingebiel et al. (2019) identify such weak signals at BCO below −50 dBZ as originating from sea salt aerosols and only signals above −50 dBZ are attributed to clouds. Clouds with reflectivity between the HALO radar MDS (−32 dBZ) and −20 dBZ and within 4 km above sea level are observed in 8.5 % of the time at BCO but only rarely (< 1.2 %) by HALO. Only clouds with a reflectivity higher than about −20 dBZ are similarly or more often observed by HALO than at BCO. Thus, we use −20 dBZ as the practical cloud detection threshold of HALO and use this value to define "radar-detectable clouds" in the observations and forward simulations.

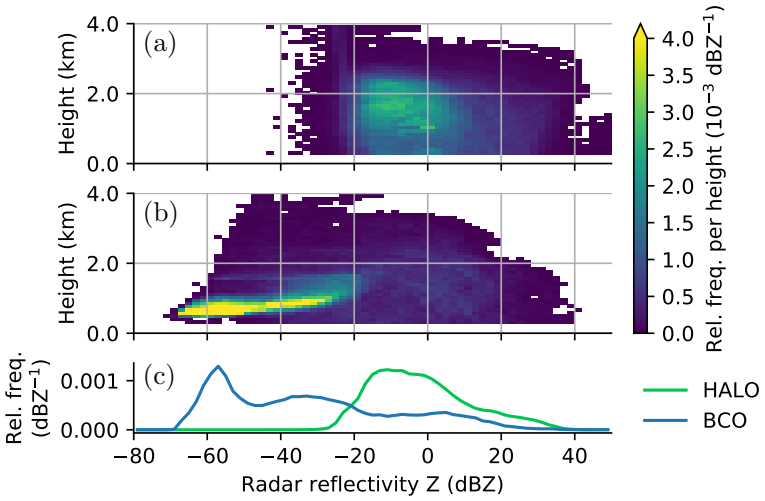

Figure A1. Height-resolved radar reflectivity distribution of shallow cumulus from (a) HALO radar and (b) BCO radar during flight days of NARVAL-South. Marginal distributions (c) show the probability density of reflectivity from HALO and BCO below 4 km. BCO data are limited to hours between 12:00 and 21:00 UTC (8:00 and 17:00 local time) on every flight day to match aircraft operation time. The probability density function of each height is normalized to the maximal possible number of data points.

Appendix B: Sub-grid clouds in the SRM

The ~~SRM~~ vertical cloud structure of the SRM deviates stronger from the observations than that of the LEM, as discussed in ~~section~~Sect. 5. This might be because the forward simulations of the SRM clouds and precipitation are analyzed based on the prognostic model ~~equations~~ outputs under the assumption that these clouds are resolved by the model grid. However, in addition to the prognostic cloud scheme, the SRM uses a diagnostic cloud scheme to model the sub-grid-scale cloud distribution used in the SRM radiation scheme. This appendix presents a rough estimation, whether the diagnostic cloud scheme ~~provides the missed outflow clouds~~ could provide the missing clouds in the stratiform mode.

The diagnostic cloud scheme uses a simple box probability density function of total water content and provides the diagnostic cloud ~~cover (CLC~~fraction (CF) and liquid cloud water content ($q_{\mathrm{c,dia}}$) (Martin Köhler, personal communication). In that scheme, the total amount of water is conserved but redistributed between the vapor, and liquid and solid cloud phases. It is assumed, that turbulent perturbations distribute the total water content in a probability density function (PDF) of rectangular shape centered around its prognostic grid-box mean. The supersaturated part of this PDF is then interpreted as diagnostic cloud cover and $q_{\mathrm{c,dia}}$. In principle, the diagnostic clouds should be analyzed as filling only their specific ~~cloud fraction~~ CF of each grid box. This means, that the diagnostic in-cloud cloud water $\frac{q_{\mathrm{c,dia}}}{\mathrm{CLC}}$ ~~covers the CLC~~ $\frac{q_{\mathrm{c,dia}}}{\mathrm{CF}}$ covers the CF fraction of a grid box. An appropriate analysis of the cloud top height distribution of the sub-grid cloudiness in the SRM would require an assumption on the horizontal overlap of sub-grid clouds within a model column. To circumvent an assumption on this, the influence of the sub-grid clouds is analyzed in terms of the CF profile. The lidar-detectable ~~cloud fraction *cf*~~ $\mathrm{CF}_{\mathrm{lidar}}$ in each height can then be

calculated as

$$\text{~~cf~~CF}_{\text{lidar}} = \frac{1}{N}\sum_{i=1}^{N} c_i \tag{B1}$$

$$c_i = \begin{cases} \text{CF}, & \text{if } \frac{q_{\text{c,dia}}}{\text{CF}} > t \\ 0, & \text{else,} \end{cases} \tag{B2}$$

with N being the number of model columns, i the column index, and t the detection threshold. ~~cf~~ CF_{lidar} describes the spatial cover in each height that contains enough cloud water to be detectable by the lidar. Analogous to the analysis in the previous section, the prognostic ~~cloud fraction~~ CF is calculated as fraction of cells in one height level, where $q_c > t$. This is a binary assumption that implies full cloud cover, if the cloud simulated from the prognostic equations is lidar detectable.

The additional ~~cloud fraction~~ CF due to the diagnostic scheme is largest (about 3.5 %) near the LCL (Fig. B1) using the sensitivity threshold $t = 10^{-7}\,\text{kg}\,\text{kg}^{-1}$ estimated from Fig. 2a. However, sensitivity tests (not shown) indicated, that the diagnostic and prognostic ~~cloud fraction~~ CF profiles derived from sensitivity thresholds between 10^{-5} and $10^{-8}\,\text{kg}\,\text{kg}^{-1}$ are not significantly different. The highest diagnostic ~~cloud fraction~~ CF is at the same height as the prognostic ~~cloud fraction~~ CF at about 500 m above LCL but about a third higher. Above its maximum, the additional ~~cloud fraction~~ CF decreases until it approaches the prognostic ~~cloud fraction~~CF. The diagnostic lidar-detectable ~~cloud cover~~ CF_{lidar} profile follows the profile of diagnostic ~~cloud cover (clc)~~ CF from the model very closely. This means, the lidar is so sensitive, that it detects all (diagnostic) model clouds with meaningful spatial extent.

As the profile shape of diagnostic clouds is very similar to the profile of prognostic clouds, we do not expect the statistics of ~~forward simulated~~ forward-simulated diagnostic clouds to differ much from what is discussed in ~~sections~~Sections 5.2 and 5.3 except for a somewhat higher frequency of lidar-detectable cloud tops. However, a proper forward simulation would have to take the sub-grid cloud overlap problem into account. The radar cloud top and base statistics are almost unaffected by the diagnostic cloud water content, as the maximum additionally diagnosed cloud water content in the SRM is only $2.2 \times 10^{-4}\,\text{kg}\,\text{kg}^{-1}$. Such contribution is insignificant for the radar-detectable cloudiness in relation to the radar detection threshold (compare Fig. 2c).

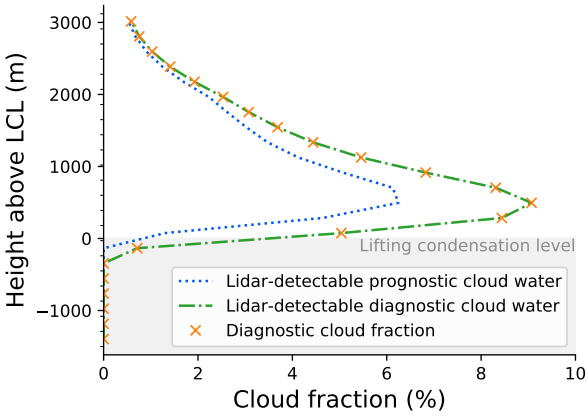

Figure B1. Mean cloud fraction profile ~~with~~ for resolved and diagnostic lidar-detectable clouds in the full SRM ~~over all cases~~output. Additionally the cloud ~~cover~~ fraction profile given by the diagnostic equations is shown.

Appendix C: Cumulative cloud cover and influence of model resolution and domain

The distribution of cloud top height detected by lidar and radar as well as the radar echo base in the observations and simulated from the model output is shown in Fig. 5. Figure C1 shows a cumulative version of that Fig. 5 to enable the comparison to analogous presentations in the literature (e.g., Medeiros et al., 2010; van Zanten et al., 2011).

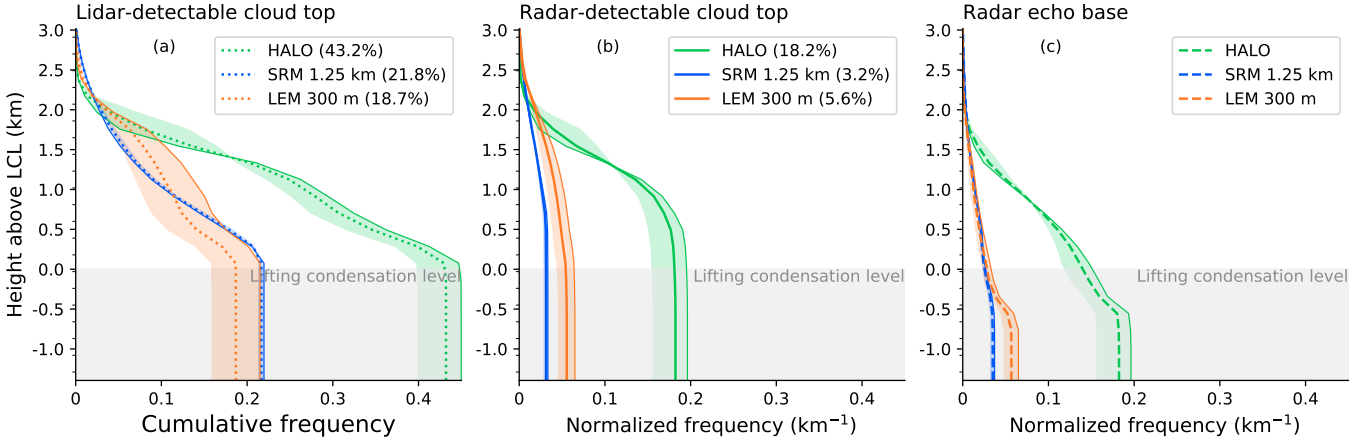

Figure C1. Cumulative cloud cover statistics on all observed and forward-simulated lidar and radar signals: (a) lidar cloud top, (b) radar cloud top, and (c) radar echo base. Same thresholds for cloud detection are used for the observed and simulated lidar and radar signals. Height is in relation to the lifting condensation level (LCL). Shadings depict western (bright edge) and eastern (dark edge) half of each dataset. The histogram bin edges are depicted as ticks on y-axis. Total shallow 2D cloud coverage (i.e., the x-axis intercept) detected by lidar and radar in each dataset are given in the legends. This figure is the cumulative version of Fig. 5.

Figure C2 investigates the influence of model resolution by including outputs from two further ICON domains with coarser grid spacings. The figure shows higher similarity among the outputs when refining the horizontal model grid from 2.5 to 1.25 km (SRM) or 600 to 300 m (LEM) than from 1.25 km to 600 m. This indicates a potentially stronger influence of the cloud representation on the model microphysical configuration (Sect. 3) compared to the horizontal resolution. However, the different spatiotemporal sampling of the model data might have to be considered here as well.

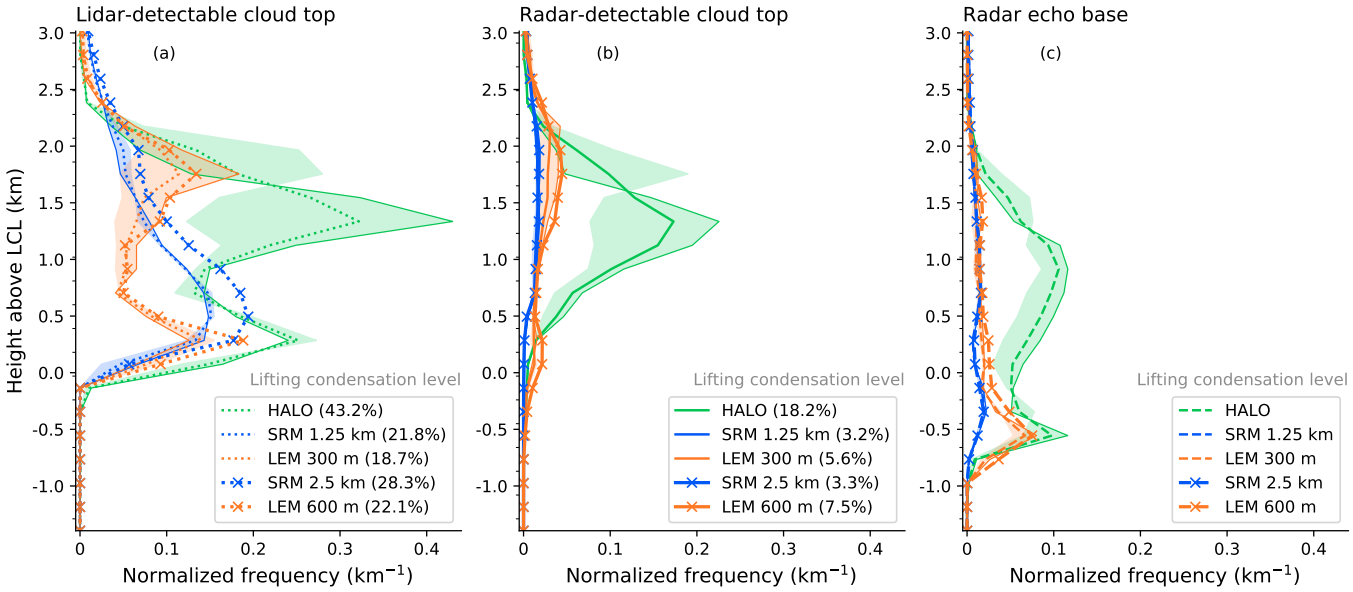

Figure C2. This figure extends Fig. 5 by including data from two additional domains from the SRM and LEM outputs at 2.5 km and 600 m grid spacings, respectively. For description see Fig. 5.

Figure C3 investigates the exact choice of domain and dates used in the analysis by taking subsamples from the 1.25 km SRM dataset. First the SRM 1.25 km dataset is restricted to those points that are near the LEM meteogram locations. The statistics of different cloud tops and bases in the different spatial subsets (Fig. C3) seems quite robust. Thus, we conclude, that the meteogram locations are in principle able to represent the cloud behavior of the full domain. Further, we restrict the SRM dataset to the four days for which also LEM output is available. This SRM subsample (also in Fig. C3) indicates a limited development of a deeper stratiform of lidar-detectable cloud, which is, however, not as prominent as in the observational or LEM datasets (Fig. 5).

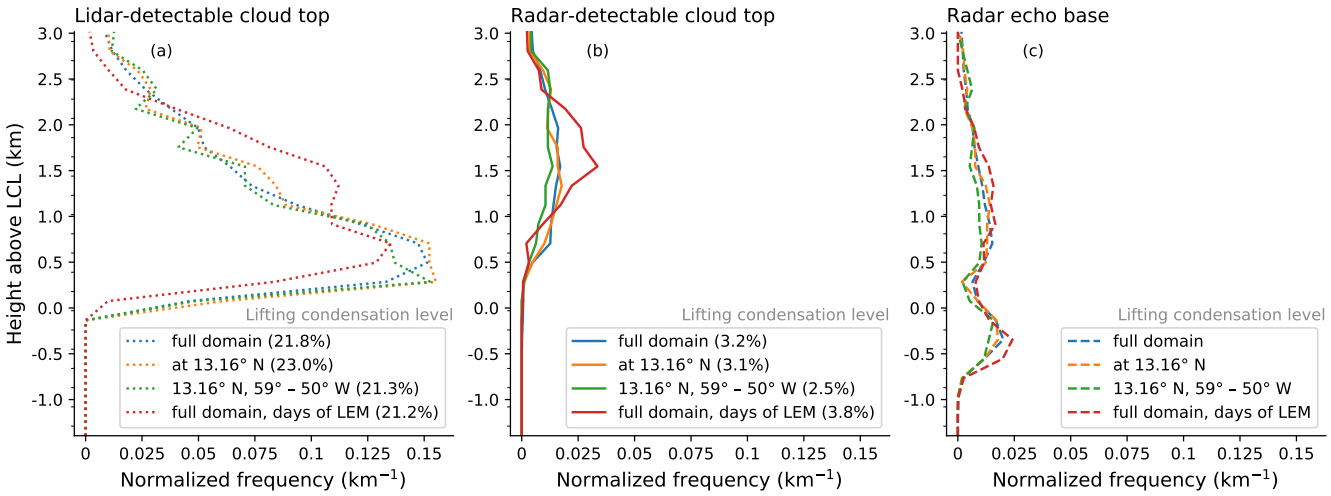

Figure C3. Comparison of cloud top and base heights in different spatial and temporal subsamples of the 1.25 km SRM dataset. Shown are the full domain (as in Fig. 5), data near the same latitude (13.16° N) as LEM meteogram locations, data near the same latitude and longitude range (59° – 50° W) as LEM meteogram locations, and full spatial domain but only the days with available LEM output. (a) Lidar-detectable cloud top height. (b) Radar-detectable cloud top height. (c) Radar echo base height. Total shallow 2D cloud coverage detected by lidar and radar in each dataset are given in the legends.

Author contributions. PK and MJ conceptualized this study. FA and SC designed the observational experiment setup and supported the interpretation of the measurements. VS supported the model interpretation. MJ performed the analysis, prepared all plots, and wrote the paper with support and input from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The work has been supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) within the DFG Priority Program (SPP 1294) "Atmospheric and Earth System Research with the Research Aircraft HALO (High Altitude and Long Range Research Aircraft)" under grant CR 111/12-1. We would like to thank Daniel Klocke and Matthias Brück for running the ICON simulations and the German Climate Computing Centre (DKRZ) for storing and supplying the data. Manuel Gutleben is thanked for making the WALES cloud top height data available. Furthermore, we would like to acknowledge the discussion on the ICON microphysics with Axel Seifert and diagnostics with Martin Köhler and Harald Rybka from the German Weather Service (DWD). The authors gratefully acknowledge the constructive comments by two anonymous referees and by Executive Editor Astrid Kerkweg who helped improving the manuscript.

References

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities, Monthly Weather Review, 139, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1, 2011.

Bodas-Salcedo, A., Webb, M., Bony, S., Chepfer, H., Dufresne, J.-L., Klein, S., Zhang, Y., Marchand, R., Haynes, J., Pincus, R., et al.: COSP: Satellite simulation software for model assessment, Bulletin of the American Meteorological Society, 92, 1023–1043, 2011.

Bony, S. and Dufresne, J.-L.: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, Geophysical Research Letters, 32, https://doi.org/10.1029/2005GL023851, 2005.

Bony, S., Stevens, B., Ament, F., Bigorre, S., Chazette, P., Crewell, S., Delanoë, J., Emanuel, K., Farrell, D., Flamant, C., Gross, S., Hirsch, L., Karstensen, J., Mayer, B., Nuijens, L., Ruppert, J. H., Sandu, I., Siebesma, P., Speich, S., Szczap, F., Totems, J., Vogel, R., Wendisch, M., and Wirth, M.: EUREC4A: A Field Campaign to Elucidate the Couplings Between Clouds, Convection and Circulation, Surveys in Geophysics, 38, 1529–1568, https://doi.org/10.1007/s10712-017-9428-0, 2017.

Bretherton, C. S. and Blossey, P. N.: Understanding Mesoscale Aggregation of Shallow Cumulus Convection Using Large-Eddy Simulation, Journal of Advances in Modeling Earth Systems, 9, 2798–2821, https://doi.org/10.1002/2017MS000981, 2017.

Crewell, S., Ebell, K., Löhnert, U., and Turner, D. D.: Can liquid water profiles be retrieved from passive microwave zenith observations?, Geophysical Research Letters, 36, https://doi.org/10.1029/2008GL036934, 2009.

Dipankar, A., Stevens, B., Heinze, R., Moseley, C., Zängl, G., Giorgetta, M., and Brdar, S.: Large eddy simulation using the general circulation model ICON, Journal of Advances in Modeling Earth Systems, 7, 963–986, https://doi.org/10.1002/2015MS000431, 2015.

Ewald, F., Groß, S., Hagen, M., Hirsch, L., Delanoë, J., and Bauer-Pfundstein, M.: Calibration of a 35 GHz airborne cloud radar: lessons learned and intercomparisons with 94 GHz cloud radars, Atmospheric Measurement Techniques, 12, 1815–1839, https://doi.org/10.5194/amt-12-1815-2019, 2019.

Frisch, A. S., Feingold, G., Fairall, C. W., Uttal, T., and Snider, J. B.: On cloud radar and microwave radiometer measurements of stratus cloud liquid water profiles, Journal of Geophysical Research: Atmospheres, 103, 23 195–23 197, https://doi.org/10.1029/98JD01827, 1998.

Genkova, I., Seiz, G., Zuidema, P., Zhao, G., and Di Girolamo, L.: Cloud top height comparisons from ASTER, MISR, and MODIS for trade wind cumuli, Remote Sensing of Environment, 107, 211–222, https://doi.org/10.1016/j.rse.2006.07.021, 2007.

Grabowski, W. W., Morrison, H., Shima, S.-I., Abade, G. C., Dziekan, P., and Pawlowska, H.: Modeling of Cloud Microphysics: Can We Do Better?, Bulletin of the American Meteorological Society, 100, 655–672, https://doi.org/10.1175/BAMS-D-18-0005.1, 2019.

Gutleben, M., Groß, S., and Wirth, M.: Cloud macro-physical properties in Saharan-dust-laden and dust-free North Atlantic trade wind regimes: a lidar case study, Atmospheric Chemistry and Physics, 19, 10 659–10 673, https://doi.org/10.5194/acp-19-10659-2019, 2019.

Heinze, R., Dipankar, A., Henken, C. C., Moseley, C., Sourdeval, O., Trömel, S., Xie, X., Adamidis, P., Ament, F., Baars, H., Barthlott, C., Behrendt, A., Blahak, U., Bley, S., Brdar, S., Brueck, M., Crewell, S., Deneke, H., Di Girolamo, P., Evaristo, R., Fischer, J., Frank, C., Friederichs, P., Göcke, T., Gorges, K., Hande, L., Hanke, M., Hansen, A., Hege, H.-C., Hoose, C., Jahns, T., Kalthoff, N., Klocke, D., Kneifel, S., Knippertz, P., Kuhn, A., van Laar, T., Macke, A., Maurer, V., Mayer, B., Meyer, C. I., Muppa, S. K., Neggers, R. A. J., Orlandi, E., Pantillon, F., Pospichal, B., Röber, N., Scheck, L., Seifert, A., Seifert, P., Senf, F., Siligam, P., Simmer, C., Steinke, S., Stevens, B., Wapler, K., Weniger, M., Wulfmeyer, V., Zängl, G., Zhang, D., and Quaas, J.: Large-eddy simulations over Germany using ICON: a comprehensive evaluation, Quarterly Journal of the Royal Meteorological Society, 143, 69–100, https://doi.org/10.1002/qj.2947, 2017.

Jacob, M., Ament, F., Gutleben, M., Konow, H., Mech, M., Wirth, M., and Crewell, S.: Investigating the liquid water path over the tropical Atlantic with synergistic airborne measurements, Atmospheric Measurement Techniques, 12, 3237–3254, https://doi.org/10.5194/amt-12-3237-2019, 2019.

Jeevanjee, N. and Romps, D. M.: Convective self-aggregation, cold pools, and domain size, Geophysical Research Letters, 40, 994–998, https://doi.org/10.1002/grl.50204, 2013.

Küchler, N., Kneifel, S., Kollias, P., and Löhnert, U.: Revisiting Liquid Water Content Retrievals in Warm Stratified Clouds: The Modified Frisch, Geophysical Research Letters, 45, 9323–9330, https://doi.org/10.1029/2018GL079845, 2018.

Khain, A. P., Beheng, K. D., Heymsfield, A., Korolev, A., Krichak, S. O., Levin, Z., Pinsky, M., Phillips, V., Prabhakaran, T., Teller, A., Heever, S. C. v. d., and Yano, J.-I.: Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization, Reviews of Geophysics, 53, 247–322, https://doi.org/10.1002/2014RG000468, 2015.

Klepp, C., Ament, F., Bakan, S., Hirsch, L., and Stevens, B.: NARVAL Campaign Report, Reports on Earth System Science 164, Max-Planck-Institut für Meteorologie, http://www.mpimet.mpg.de/fileadmin/publikationen/Reports/WEB_BzE_164_last.pdf, 2014.

Klingebiel, M., Ghate, V. P., Naumann, A. K., Ditas, F., Pöhlker, M. L., Pöhlker, C., Kandler, K., Konow, H., and Stevens, B.: Remote Sensing of Sea Salt Aerosol below Trade Wind Clouds, Journal of the Atmospheric Sciences, 76, 1189–1202, https://doi.org/10.1175/JAS-D-18-0139.1, 2019.

Klocke, D., Brück, M., Hohenegger, C., and Stevens, B.: Rediscovery of the doldrums in storm-resolving simulations over the tropical Atlantic, Nature Geoscience, 10, 891–896, https://doi.org/10.1038/s41561-017-0005-4, 2017.

Konow, H., Jacob, M., Ament, F., Crewell, S., Ewald, F., Hagen, M., Hirsch, L., Jansen, F., Mech, M., and Stevens, B.: A unified data set of airborne cloud remote sensing using the HALO Microwave Package (HAMP), Earth System Science Data, 11, 921–934, https://doi.org/10.5194/essd-11-921-2019, 2019.

Krautstrunk, M. and Giez, A.: The Transition From FALCON to HALO Era Airborne Atmospheric Research, in: Atmospheric Physics, pp. 609–624, Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-642-30183-4_37, 2012.

Lamer, K., Kollias, P., and Nuijens, L.: Observations of the variability of shallow trade wind cumulus cloudiness and mass flux, Journal of Geophysical Research: Atmospheres, 120, 6161–6178, https://doi.org/10.1002/2014JD022950, 2015.

Lamer, K., Fridlind, A. M., Ackerman, A. S., Kollias, P., Clothiaux, E. E., and Kelley, M.: $(GO)^2$-SIM: a GCM-oriented ground-observation forward-simulator framework for objective evaluation of cloud and precipitation phase, Geoscientific Model Development, 11, 4195–4214, https://doi.org/10.5194/gmd-11-4195-2018, 2018.

Leahy, L. V., Wood, R., Charlson, R. J., Hostetler, C. A., Rogers, R. R., Vaughan, M. A., and Winker, D. M.: On the nature and extent of optically thin marine low clouds, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2012JD017929, 2012.

Mech, M., Orlandi, E., Crewell, S., Ament, F., Hirsch, L., Hagen, M., Peters, G., and Stevens, B.: HAMP – the microwave package on the High Altitude and LOng range research aircraft (HALO), Atmospheric Measurement Techniques, 7, 4539–4553, https://doi.org/10.5194/amt-7-4539-2014, 2014.

Mech, M., Maahn, M., Kneifel, S., Ori, D., Orlandi, E., Kollias, P., Schemann, V., and Crewell, S.: PAMTRA 1.0: A Passive and Active Microwave radiative TRAnsfer tool for simulating radiometer and radar measurements of the cloudy atmosphere, Geoscientific Model Development Discussions, 2020, 1–34, https://doi.org/10.5194/gmd-2019-356, 2020.

Medeiros, B., Nuijens, L., Antoniazzi, C., and Stevens, B.: Low-latitude boundary layer clouds as seen by CALIPSO, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2010JD014437, 2010.

Naumann, A. K. and Kiemle, C.: The vertical Structure and spatial Variability of lower tropospheric Water Vapor and Clouds in the Trades, Atmospheric Chemistry and Physics Discussions, pp. 1–24, https://doi.org/10.5194/acp-2019-1015, 2019.

Nuijens, L., Serikov, I., Hirsch, L., Lonitz, K., and Stevens, B.: The distribution and variability of low-level cloud in the North Atlantic trades, Quarterly Journal of the Royal Meteorological Society, 140, 2364–2374, https://doi.org/10.1002/qj.2307, 2014.

O, K.-T., Wood, R., and Tseng, H.-H.: Deeper, Precipitating PBLs Associated With Optically Thin Veil Clouds in the Sc-Cu Transition, Geophysical Research Letters, 45, 5177–5184, https://doi.org/10.1029/2018GL077084, 2018.

O'Connor, E. J., Hogan, R. J., and Illingworth, A. J.: Retrieving Stratocumulus Drizzle Parameters Using Doppler Radar and Lidar, Journal of Applied Meteorology, 44, 14–27, https://doi.org/10.1175/JAM-2181.1, 2005.

Oue, M., Tatarevic, A., Kollias, P., Wang, D., Yu, K., and Vogelmann, A. M.: The Cloud Resolving Model Radar Simulator (CR-SIM) Version 3.2: Description and Applications of a Virtual Observatory, Geoscientific Model Development, 13, 1975–1998, https://doi.org/10.5194/gmd-13-1975-2020, 2020.

Romps, D. M.: Exact Expression for the Lifting Condensation Level, Journal of the Atmospheric Sciences, 74, 3891–3900, https://doi.org/10.1175/JAS-D-17-0102.1, 2017.

Rose, T., Crewell, S., Löhnert, U., and Simmer, C.: A network suitable microwave radiometer for operational monitoring of the cloudy atmosphere, Atmospheric Research, 75, 183–200, https://doi.org/10.1016/j.atmosres.2004.12.005, 2005.

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., and Düben, P.: Global Cloud-Resolving Models, Current Climate Change Reports, 5, 172–184, https://doi.org/10.1007/s40641-019-00131-0, 2019.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., and Siebesma, A. P.: Climate goals and computing the future of clouds, Nature Climate Change, 7, 3–5, https://doi.org/10.1038/nclimate3190, 2017.

Schulz, H. and Stevens, B.: Observing the Tropical Atmosphere in Moisture Space, Journal of the Atmospheric Sciences, 75, 3313–3330, https://doi.org/10.1175/JAS-D-17-0375.1, 2018.

Seifert, A. and Beheng, K. D.: A double-moment parameterization for simulating autoconversion, accretion and selfcollection, Atmospheric Research, 59-60, 265–281, https://doi.org/10.1016/S0169-8095(01)00126-0, 2001.

Siebert, H., Beals, M., Bethke, J., Bierwirth, E., Conrath, T., Dieckmann, K., Ditas, F., Ehrlich, A., Farrell, D., Hartmann, S., Izaguirre, M. A., Katzwinkel, J., Nuijens, L., Roberts, G., Schäfer, M., Shaw, R. A., Schmeissner, T., Serikov, I., Stevens, B., Stratmann, F., Wehner, B., Wendisch, M., Werner, F., and Wex, H.: The fine-scale structure of the trade wind cumuli over Barbados – an introduction to the CARRIBA project, Atmospheric Chemistry and Physics, 13, 10 061–10 077, https://doi.org/10.5194/acp-13-10061-2013, 2013.

Siebesma, A. P., Bretherton, C. S., Brown, A., Chlond, A., Cuxart, J., Duynkerke, P. G., Jiang, H., Khairoutdinov, M., Lewellen, D., Moeng, C.-H., Sanchez, E., Stevens, B., and Stevens, D. E.: A Large Eddy Simulation Intercomparison Study of Shallow Cumulus Convection, Journal Of The Atmospheric Sciences, 60, 19, 2003.

Stevens, B., Farrell, D., Hirsch, L., Jansen, F., Nuijens, L., Serikov, I., Brügmann, B., Forde, M., Linne, H., Lonitz, K., and Prospero, J. M.: The Barbados Cloud Observatory: Anchoring Investigations of Clouds and Circulation on the Edge of the ITCZ, Bulletin of the American Meteorological Society, 97, 787–801, https://doi.org/10.1175/BAMS-D-14-00247.1, 2015.

Stevens, B., Brogniez, H., Kiemle, C., Lacour, J.-L., Crevoisier, C., and Kiliani, J.: Structure and Dynamical Influence of Water Vapor in the Lower Tropical Troposphere, Surveys in Geophysics, 38, 1371–1397, https://doi.org/10.1007/s10712-017-9420-8, 2017.

Stevens, B., Ament, F., Bony, S., Crewell, S., Ewald, F., Gross, S., Hansen, A., Hirsch, L., Jacob, M., Kölling, T., Konow, H., Mayer, B., Wendisch, M., Wirth, M., Wolf, K., Bakan, S., Bauer-Pfundstein, M., Brueck, M., Delanoë, J., Ehrlich, A., Farrell, D., Forde, M., Gödde, F., Grob, H., Hagen, M., Jäkel, E., Jansen, F., Klepp, C., Klingebiel, M., Mech, M., Peters, G., Rapp, M., Wing, A. A., and Zinner,

T.: A High-Altitude Long-Range Aircraft Configured as a Cloud Observatory: The NARVAL Expeditions, Bulletin of the American Meteorological Society, 100, 1061–1077, https://doi.org/10.1175/BAMS-D-18-0198.1, 2019.

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., Rybka, H., Schubotz, W., Windmiller, J., Adamidis, P., Arka, I., Barlakas, V., Biercamp, J., Brueck, M., Brune, S., Buehler, S. A., Burkhardt, U., Cioni, G., Costa-surós, M., Crewell, S., Crüger, T., Deneke, H., Friederichs, P., Henken, C. C., Hohenegger, C., Jacob, M., Jakub, F., Kalthoff, N., Köhler, M., Laar, T. W. v., Li, P., Löhnert, U., Macke, A., Madenach, N., Mayer, B., Nam, C., Naumann, A. K., Peters, K., Poll, S., Quaas, J., Röber, N., Rochetin, N., Scheck, L., Schemann, V., Schnitt, S., Seifert, A., Senf, F., Shapkalijevski, M., Simmer, C., Singh, S., Sourdeval, O., Spickermann, D., Strandgren, J., Tessiot, O., Vercauteren, N., Vial, J., Voigt, A., and Zängl, G. a.: The Added Value of Large-eddy and Storm-resolving Models for Simulating Clouds and Precipitation, Journal of the Meteorological Society of Japan. Ser. II, 98, 395–435, https://doi.org/10.2151/jmsj.2020-021, 2020.

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G.: Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel, Bulletin of the American Meteorological Society, 94, 1031–1049, https://doi.org/10.1175/BAMS-D-12-00117.1, 2013.

Vaisala: Vaisala RD94 Dropsonde Datasheet, Technical data, Vaisala, https://www.vaisala.com/sites/default/files/documents/RD94-Datasheet-B210936EN-B.pdf, 2017.

van Zanten, M. C., Stevens, B., Nuijens, L., Siebesma, A. P., Ackerman, A. S., Burnet, F., Cheng, A., Couvreux, F., Jiang, H., Khairoutdinov, M., Kogan, Y., Lewellen, D. C., Mechem, D., Nakamura, K., Noda, A., Shipway, B. J., Slawinska, J., Wang, S., and Wyszogrodzki, A.: Controls on precipitation and cloudiness in simulations of trade-wind cumulus as observed during RICO, Journal of Advances in Modeling Earth Systems, 3, https://doi.org/10.1029/2011MS000056, 2011.

Vial, J., Vogel, R., Bony, S., Stevens, B., Winker, D. M., Cai, X., Hohenegger, C., Naumann, A. K., and Brogniez, H.: A new look at the daily cycle of trade wind cumuli, Journal of Advances in Modeling Earth Systems, 11, 3148–3166, https://doi.org/10.1029/2019MS001746, 2019.

Vogel, R., Nuijens, L., and Stevens, B.: Influence of deepening and mesoscale organization of shallow convection on stratiform cloudiness in the downstream trades, Quarterly Journal of the Royal Meteorological Society, 146, 174–185, https://doi.org/10.1002/qj.3664, 2020.

Wentz, F. J. and Meissner, T.: Algorithm Theoretical Basis Document (ATBD), Version 2, AMSR Ocean Algorithm, Tech. rep., Remote Sensing Systems, http://images.remss.com/papers/rsstech/2000_121599A-1_Wentz_AMSR_Ocean_Algorithm_ATBD_Version2.pdf, 2000.

Wirth, M., Fix, A., Mahnke, P., Schwarzer, H., Schrandt, F., and Ehret, G.: The airborne multi-wavelength water vapor differential absorption lidar WALES: system design and performance, Applied Physics B, 96, 201–213, https://doi.org/10.1007/s00340-009-3365-7, 2009.

Wolf, K., Ehrlich, A., Jacob, M., Crewell, S., Wirth, M., and Wendisch, M.: Improvement of airborne retrievals of cloud droplet number concentration of trade wind cumulus using a synergetic approach, Atmospheric Measurement Techniques, 12, 1635–1658, https://doi.org/10.5194/amt-12-1635-2019, 2019.

Wood, R., O, K.-T., Bretherton, C. S., Mohrmann, J., Albrecht, B. A., Zuidema, P., Ghate, V., Schwartz, C., Eloranta, E., Glienke, S., Shaw, R. A., Fugal, J., and Minnis, P.: Ultraclean Layers and Optically Thin Clouds in the Stratocumulus-to-Cumulus Transition. Part I: Observations, Journal of the Atmospheric Sciences, 75, 1631–1652, https://doi.org/10.1175/JAS-D-17-0213.1, 2018.

Wyant, M. C., Bretherton, C. S., Rand, H. A., and Stevens, D. E.: Numerical Simulations and a Conceptual Model of the Stratocumulus to Trade Cumulus Transition, Journal of the Atmospheric Sciences, 54, 168–192, https://doi.org/10.1175/1520-0469(1997)054<0168:NSAACM>2.0.CO;2, 1997.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M.: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core, Quarterly Journal of the Royal Meteorological Society, 141, 563–579, https://doi.org/10.1002/qj.2378, 2015.