This paper compares NWP model output from the UM (using several different cloud schemes) and the IFS against ground-based observations collected in the central Arctic during a summer time cruise that sampled both sea ice melting conditions and refreezing conditions. The results demonstrated that both model frameworks overestimated cloud occurrence, but that using a total water content method, which takes into account the insensitivity of the ground-based remote sensors to very small amounts of cloud water content (or very small particle size) provided better agreement in cloud occurrence. Even still, the models tended to overestimate the LWC relative to the observations, which was hypothesized to result in too much cloud top radiative cooling which had deleterious impacts on the simulated temperature profile. However, they also showed that the UM models, which were all limited area models, sensitive to the forcing conditions from the global model used to drive these simulations. They also showed that a more accurate treatment of aerosols in the UM-LAM with the most complex cloud microphysics did change the profile of LWC in the lowest levels, but had virtually no other effect on the cloud lifetime, precipitation amount, or the biases in the thermodynamic profiles.

I found this paper very interesting, well-motivated, and well written.

My main comment is associated with the 4th point raised in the conclusions:

I think that the sensitivity of the results to the forcing dataset used to drive the UM models casts a lot of questions on this analysis. In section 3.4 (and later sections), the authors work hard to connect errors in clouds to errors in thermodynamic profiles (it seemed like a cause-effect implication). However, I don’t think the authors have done enough to convince me that the errors in the clouds are causing the rest of the issues. I think this could be addressed reasonably simply by showing the biases in the
thermodynamic profiles over the region from the forcing dataset itself (e.g., in Fig 13 and subsequent figures). I realize that the UM models are providing 12-36 h forecasts that start from the forcing dataset, but I think adding these bias profiles would still be useful in making their case.

In a very similar and related comment: it looks like the biases in the LWC profiles from the three UM models are quite different, but the biases in the temperature and moisture profiles are essentially identical. If errors in the cloud properties are truly the driver (via radiation) of the biases in the temperature profiles, then I would have hypothesized we would see differences in the biases in the temperature profiles from the three models. Why don’t we?

Minor comments:

- The differences between the model diagnosed cloud cover and the cloud cover derived from the TWC “cloudNet simulator” was striking. I feel that there was too much emphasis on the Cv estimate; I believe that we need to use instrument simulators much more routinely in model – observation comparisons. I would like to see this emphasized more in the conclusions.
- Line 389: that all model simulations overestimate LWC in the 1-3 km range relative to the observations is interesting, especially since the obs are using an adiabatic assumption to distribute the liquid water. Thus, the true bias in LWC in the models is likely even larger than what was shown. I think this should be pointed out somewhere in the paper.
- Fig 6: the units of LWP are incorrect; I suspect they should be g/m2
- Lines 472-474: the water vapor units are g/kg, not g/m3
- The yellow color used to denote the IFS results is too faint to see well; please increase its contrast
- Line 522: Satellites provide good coverage of the arctic, and the infrared sounders do provide thermodynamic profiles (of some quality, depending on your metric). I think that some mention to the challenge of using these satellite data for DA is needed here.
- LW radiative cooling is strongly dependent on (a) the integrated water content of the cloud and (b) if there is another cloud above the radiating layer or not. Turner et al. JAMC 2018 provides a good illustration of this for arctic clouds. Line 562 is hypothesizing too much LW radiative cooling, but we have seen that the different microphysics parameterizations yield different LWCs. Is this because there are clouds above this BL that is muting this radiative impact somehow?
- Lin 591: are the correlation coefs on Fig 14 for the “orig inv” or the “adj inv” dataset? It is not clear
- Line 607-609: if there is too much radiative cooling at cloud top (because the LWC is too high), then this would result in greater LW radiative warming in the lower part of the BL, which could lead to this warm bias in the surface (a possible explanation for the warm bias near the surface).