Review of “Evaluating Arctic clouds modelled with the Unified Model and Integrated Forecasting System“ by Young et al.

The authors analyse the capabilities of the Unified Model and the IFS to accurately capture clouds and surface radiative fluxes measured during the Arctic Ocean 2018 expedition. They present simulations covering roughly a month-long period and they discuss the different performance of the models over this time period. Model simulations differ in various settings and physical parameterisations. While finding systematic biases in cloud cover and to a lesser extend in cloud microphysical structure, thermodynamic and moisture biases inherited from the driving model are found to play a major role.

Overall, the paper presents a very interesting and comprehensive analysis and the paper is generally well-written. The results are mostly presented in a clear manner. However, some clarifications / more extensive presentation are needed in some places, while the paper is repetitive in other sections (details and suggestions in the comments below). However, these are mostly minor issues, that should be easy to fix. If these corrections have been, I recommend the paper to be published.

Comments

- l. 96: What hypotheses? The bullet point list given above does not list hypotheses but rather model components to be investigated.
- l. 148-151: Please briefly discuss the spatial heterogeneity or homogeneity of the modelled/ observed cloud fields. Are there any issues to be expected from using the closest grid point to the ship location for verification?
- l. 173 ff: How representative are the climatological values for sea-ice cover and thickness for the month considered here? This could be particular an issue given the rapid shift in conditions due to climate change in recent years.
- l. 178 ff: How are sea-ice cover and sea surface temperature modelled in the UM?
- l. 220 ff: What are the assumptions / settings in the turbulence parameterisation in UM_CASIM-100? Are they similar to UM_RA2M or UM_RA2T or altogether different?
- l. 237 ff and earlier: The claim to test different settings with a simulation where multiple parameters / parameterisations are changed (here: albedo parameterisation and cloud microphysics) is a bit troublesome. Assigning changes or non-changes is difficult if there is more than one scheme changed. The authors should better acknowledge this in the text here and in the presentation of the results later on (btw. this holds also for the comparison of CASIM simulation to UM_RA2M and UM_RA2T).
- l. 255: This is a vital point in the later analysis, therefore a short summary on how periods of “consistent meteorology“ are identified and what characterises them is needed here. Only referencing earlier work is not sufficient here.
- l. 304: The differences in the values given here are actually larger than the -1.9 Wm^{-2} maximum bias given above for the simulation claimed to be a worse match to observations.
- l. 347: The failure of the model to reproduce persistent clouds at higher altitudes is actually also notable in the TWC metric.
- l. 403: Give some brief insight in why the ice-phase should be more active in UM_RA2M than UM_RA2T despite them using the same microphysics. Is this due to mixing processes or the differing cloud schemes?
- l. 405 ff: Please clarify if the LWP is calculated for z>150 m in observations and model?
- l. 429 - 432: Please rephrase this sentences and make clearer references to the Figures.
- l. 445: It is fine to have the details of this simulation in the SI, but at least provide the aerosol number concentration in the BL in the main text for comparison with the 100 cm^{-3} simulation.
- l. 490 ff: Would it not make sense to look at relative rather than absolute biases in moisture content. The tampering off above 4 km could be due to the general decrease of q with altitude.
- l. 547 ff: How much of the moisture biases actually translates into an relative humidity bias, which is more important for cloud cover than pure moisture biases.

- l. 606 ff: Can you estimate how much of these biases are actually inherited from the global model? This is discussed later in section 4 and also in the abstract, but initial / lateral boundary condition biases are not mentioned here.
- Section 4: This section essentially is a somewhat more concise presentation of the key results already presented in section 3. This is repeated again in section 5. Hence there is a lot of repetition in these sections, which makes the paper more lengthy and tedious to read than necessary. Please consider either shortening the summary of results in section 4 or moving the few instances of proper discussion / comparison with other studies to section 3 or 5.
- l. 637: Can you at least briefly provide some examples of the “other meteorological factors and incorrect model biases“ you mention here.
- l. 679: Despite the present study covering a month of simulations, it is not possible to draw a link between model biases in the present study and climate change signals.

Technical corrections

- l. 30: “cloud cover likely contributes to th~~e ever-present~~ near-surface temperature bias“
- l. 153: “ from Jun**e** 2019 to Jun**e** 2020“
- l. 338: I do not think surface properties are extensively discussed rather thermodynamic conditions are analysed in the following.
- l. 394: Not sure what you mean with “consistent with altitude“ here.
- l. 401/402: What are “shaded standard deviations“?
- l. 416: precipitation is increased compare to what?
- l. 471: replace “noise“ with “variability“
- l. 537: “a more prominent than a secondary layer“ ? Please rephrase.
- l. 566: Not sure which “effect“ you are refering to here.
- l. 728: “less dominant ice **microphysics**“