Model evaluation by a cloud classification based on multi-sensor observations by Hansen et al. submitted to GMD.

General comments and recommendation:

This manuscript describes a hydrometeor classification algorithm that turns modeled thermodynamic profiles and profiles of hydrometeor concentrations into a hydrometeor classification, which can then be used to evaluate the models' ability to produce the right hydrometeor type at the right time and location. It is inspired by the CloudNet algorithm, in the sense that it uses the same hydrometeor categories, but otherwise is based solely on modeled output. In this case, it uses a given constant concentration threshold for all hydrometeors to decide which hydrometeor type dominates. In doing so, the authors make a useful, though simple, first step towards comparing hydrometeor types that are observed and modeled. This allows first glances into whether a model produces the right hydrometeors/clouds and can be used in time series, point to point comparisons or larger area comparisons, and the authors show a few of these for the COSMO operational weather model and the ICON LES model.

Despite this nice first effort, the authors do not convince (me) that this modeled classification provides more insight into the model's behavior than the available model output already can do. The authors state that modeled biases they find (for instance, the overestimation of ice in COSMO) can not be addressed using ice cloud fraction profiles, but do not show that their method works better. Because the classification is solely based on model output, and does not - unlike the title suggest - take into account hydrometeor aspects that are detected with multi-sensor information such as particle sizes and radiative effects - it cannot be used to make fair comparisons with observations. Instrument simulator techniques that have long been developed to address such issues are not discussed. The simple threshold for hydrometeor concentrations is not further tested or discussed, for instance, what this implies for a comparison with the instrument-based classification (by the CloudNet algorithm) where much more information about hydrometeors is available based on lidars and radars.

What surely does not help is that the manuscript suffers from poor writing, including poor argumentation, structuring, repetition, spelling/grammar mistakes and spoken language. Especially at the outset of the manuscript (the introduction and methods) the authors fail at clearly delineating what their new classification entails and implies.

My recommendation is to reject this manuscript, and ask the authors to work on their method in light of what has been happening in much more detail by the community, including the use of simulators, and present more convincingly and in detail how this provides detailed insight in how the model performs, more so than other methods or model diagnostics can do. I also ask the authors to considerably improve on the writing. Some specific comments to think about when preparing a new manuscript are included below.

Specific comments:

Title: This is a model evaluation by a cloud classification inspired by the CloudNet algorithm but based on modeled hydrometeor concentrations. The current title suggests that multi-sensor information is used in a detailed assessment of hydrometeor concentration type, size, radiative effects, which is not the case.

Abstract:

L3: an important step, not necessarily the first.

From the second paragraph: not immediately clear whether the classification is only applied to models, or if a revised version is applied to models and to observations.

L10/11: regarding different cloud types : unclear. You mean, the accuracy of forecasted cloud types?

Third paragraph: can the authors include results about how the classification works for COSMO/LES, and what such an evaluation has taught you about the vertical structure of cloud and cloud type in either model.

L13: as the observations. Which show detailed cloud structures, the obs or LES?

Section 1. Introduction:

The first two paragraphs are oddly written. You first make it sound like only a model variable is derived, but LWC is not specifically a model variable, it is a variable on its own, that can be derived from obs and models. Why is this bad? You mention evaluation of mixed-phase clouds, and that a modeled classification offers potential to investigate mixed-phase clouds in models. A classification can be used to evaluate ALL types of clouds - but the CloudNet classification may so far be only operational for supersites that are located in (midlatitude) regions where mixed-phase clouds are common. This is an important difference.
P2L10: atmospheric models like ..... now you should mention which models, and then cite literature.

P2L16: I disagree with your saying that the CloudNet products have not been used to evaluate the representation of clouds in models, in fact, you mention a whole list of them on L10. What you probably mean to say is that you can make qualitative comparisons between Cloudnet-derived products and models, and more precise quantitative (what you call direct) comparisons. The latter might have happened, but is not entirely fair, if the model does not have a comparable classification.

L17: a surrogate "offers"...

L19: Similar cloud classifications - what does similar here refer to? Unclear.

L 23: rain radars do not see cloud droplets, only rain drops, so how can they provide detailed cloud information?

L26: most atmospheric models are numerical models. Are you making a classification that works on a certain range of model resolutions? Which ones? Is this applicable to everything from DNS to a GCM?

L26-33: this again is very oddly written, and the content of sentences is not logically connected. First you write that you make a new classification, so now we're expecting to hear more details on that. Then you write that standard metrics are not part of (a, your, which?) classification, ok fine, then what is? Then you talk about fuzzy methods ... is this what you use? Details on that this is inspired on methods

often used in the evaluation of precipitation is useful, and can be added, but really as a reader we are expecting to hear what is specific about your classification.

Section 2. Data & methods

P3L12/13: Remove "due to the remote sensing measurement characteristics"

A number of poor grammar/spelling here. The abundant use of e.g. throughout the manuscript is one example, including spoken language like: Isn't. Don't. Till. Spelling mistake: see occurence in the caption of Fig 4. This list is not comprehensive.

P4 second paragraph: Are we assumed to now understand how the classification works? Will you explain Fig 2? How are the thermodynamic profiles used? How does this connect to next paragraphs? Can you write it in a way that allows us to understand that what follows will be an explanation of Fig 2?

L25-35: which thresholds on QC/QI? You mention it later - include here?

P5L4-5: "would be thus difficult to interpret" poor grammar

P5L1: "have to be chosen" - but are in the end not chosen based on the instrumentation, but purely from a model point of view. For the lidars, the signals are certainly very sensitive to the concentration of particles, but for the radars, the signal will additionally be very sensitive to the particle sizes: a few large particles may already give you a return. Especially for cases where both ice crystals and cloud droplets are present, or where drizzle/rain drops and cloud droplets are present, the concentrations of the larger particles (crystals and rain drops) can give returns to the radar that would make the original CloudNet classification categorize this as a mixture of both, whereas the concentration of the larger particles might be small enough to let the model classification categorize this as just cloud droplets. COSMO only has a 1 moment scheme, and therefore you could not retrieve any information about the particle size, and perhaps this is why you have chosen just a simple threshold. But other models might have a 2 moment scheme, including the ICON-LES I believe, and because the lidar/radars are sensitive to both, this paper should at least have a discussion about how sensitive the choice of this threshold is.

What does a *significant* concentration mean for the model? How low/high is this concentration with respect to what we usually find in clouds.

P6L8-9: "by averaging the categories to the most frequent one": how does one average to a most frequent one? Don't you just select the most frequent one, no averaging involved? Is the most frequent one also the one that has the highest concentration? What if there is a warm cloud with one strong rain event, then the most frequent cloud category might be cloud droplets, but the highest concentration or strongest returns are at times of rainfall. My point is that such aspects are not discussed.

Section 3:

First sentence, first paragraph. It is an overstatement to say that your classification is consistent with the original CloudNet algorithm applied to observations. If anything, it is inspired by the CloudNet algorithm by selecting the same categories. The subsequent selection is purely based on modeled concentrations of different hydrometeors, which are in no way compared or comparable to lidar and radar retrieved signals that are functions of concentrations, particle sizes and more. The

authors also do not show how the original CloudNet algorithm does the selection, so there is no way for the reader to assess how much more detail the original classification entails.

PL26 : the temperature and humidity profile may be correct, but the microphysical scheme wrong. There is no way to separate the origin of these errors from a qualitative comparison of cloud categories. L29: detailed insights: give me an example.

P13L15/16: "Nevertheless, differences for example between the categories of "Drizzle or rain" and "Drizzle/rain & cloud droplets" can be identified, which provide detailed insights into the models' microphysics." Please provide the reader with what detailed insight you obtained regarding ICON's microphysical scheme. If you would give this to Axel Seifert, what does he learn from this what he could/did not know from other diagnostics.

P8L20: I am not at all convinced by your statement that identifying a modeled overestimation of ice clouds at the expense of clear skies is not feasible when looking at the mean cloud fraction profiles. Please show the ice fraction profiles of COSMO and the observations and explain why it does not reveal COSMO's bias.

P8L23-24: "Nevertheless, especially at high altitudes, the lower sensitivity of the remote-sensing instruments has to be considered which could reduce the observed frequency of occurrence of "Ice clouds". Isn't this exactly why you wish to have a modeled classification that more closely resembles what the instruments are seeing. Implementing instrument simulators in the model, and then doing the classification there, would be a fairer comparison wouldn't it. Now we are left wondering which one is right, and we cannot - unlike the authors state - identify the origin of errors and prove a detailed insight of why the model is wrong.

The first paragraphs of sections (point-to-point comparisons/fuzzy verification) repeat much of what has been written before, although they are written more clearly than before. The summary is certainly much clearer and better written that earlier parts of the text.