Response to Referee 1 on revised manuscript 'Machine learning of cloud types in satellite observations and climate models'

Peter Kuma, Frida A.-M. Bender, Alex Schuddeboom, Adrian J. McDonald, and Øyvind Seland

November 2, 2022

Dear Dr. Steven Sherwood,

Thank you very much for the second round of comments. Please find our response below. In the following text, the original comments are in **bold**, followed by our response.

Kind regards,

Dr. Peter Kuma on behalf of the authors

1. As raised by both reviewers of the original paper, it was very odd that the performance on the test data was slightly higher than on the training data. Now that the authors have more clearly described their approach, it is clear that they did not use independent test and training data. The test data come from a subset of the same years as used in training. The authors choose instances at random from within those years, but due to spatio-temporal autocorrelation of the data (and the fact that nothing is said about any effort to ensure that the training and test data tiles don't overlap), this will not result in an independent test. The authors should either redo their analysis with non-overlapping training and test periods, or else they should just use all the available data for training and explain that they did not do an independent test at all (this would be unsatisfactory and against usual practice, but would at least be an accurate description).

The sentence in our revised manuscript describing the validation was wrong (L260, 'We trained the ANN on CERES and IDD data in years 2004, 2005, 2007 and 2009–2017, with years 2007, 2012 and 2017 used as a validation dataset, representing 25% of the total number of years.'). The training and validation years were strictly separate, with the training dataset consisting of years 2004, 2005, 2009–2011, 2013–2016, and 2018–2020, and the validation dataset consisting of years 2007, 2012 and 2017. We changed the sentence to: 'We trained the ANN on CERES and IDD data in years 2004, 2005, 2009–2011, 2013–2016 and 2018–2020, with years 2007, 2012 and 2017 used as a validation dataset, representing 20% of the total number of years.' We apologise for this error.

The validation dataset was used in the training only as a stopping criterion – the training process was stopped when the loss function on the validation dataset was not improved for three consecutive training rounds. Therefore, the input of the validation dataset on the training process was likely negligible.

In the second revised manuscript, we also perform validation by training the ANN on the training data excluding station data from a number of geographical locations – quarters of the globe (north-east, north-west, south-east, south-west), as well as four smaller regions over North Atlantic, East Asia, Oceania and South America. In this way, we test the ability of the ANN to predict cloud occurrence over stations for which no information was supplied during the training, we well as temporally to years in included in the training (2007, 2012 and 2017). The new validation is described in Sect. 3.3 and 4.1. The new validation includes comparison of between the ANN and the IDD stations in the regions excluded from the training in terms of RMSE of daily means and all-time means, and a receiver operating characteristic (ROC).

2. Related to (1): the authors do not present any quantification of the skill of their ANN on the IDD data. The only skill scores reported compare the ANN classifications based on GCM or reanalysis radiances to those based on CERES observed radiances. This is legitimate as a way of comparing models to obs, but does not assure us that the basis of comparison is physically meaningful or that the station data have actually added anything. We can roughly assess the skill by comparing the first two rows of Fig. 3, but this is hard to see, and only addresses the skill of time-averaged fields whereas the method should also capture time variations (does it?). They do not assess skill on the micro-classifications

at all, yet present some results on those later. If they are going to present such results they need to show whether the ANN can actually distinguish any of those cloud classes.

As mentioned in the point above, we added a new validation which compares the ANN with IDD station data not included in the training. We also determine the accuracy on daily scales as RMSE and ROC. We do not relay on accuracy of the ANN on daily scales for the main results of the study. All presented figures are long-term means. To some extent Fig. 5 (cloud optical depth–cloud top pressure; in the second revised manuscript Fig. 8) and Fig. B1 (comparison with SOM) depend on daily scale accuracy, even though they present the information as long-term means. The other main figures (Fig. 3, 4, 6–8 and 9; in the second revised manuscript Fig. 6, 7, 9–12) are based entirely on long-term means, and no daily scale information is compared between the ANN applied on models and the ANN applied on CERES.

3. The authors note that one prior study (Shuddeboum and MacDonald 2021) that was similar did not find any relationship to ECS. In Appendix B they seem to explore in detail the alternate classification system, but then do not say anything about why this discrepancy occurred. Can they work out based on this analysis why the answers were different?

The strongest factor might be the small set of models analysed in Schuddeboom and McDonald (2021) – only 8 models. The two cloud classifications are not the same by definition, and therefore there is no reason to think that they have to show the same relationship to ECS. We performed the comparison between the two classifications (1) as a sanity check for our classification scheme, and (2) to provide context for the new classification, alongside a comparison of the cloud top pressure – cloud optical depth histograms with the ISCCP classification (Sect. 4.3). The classification schemes represent particular decomposition of the radiation fields from which they are derived. Different statistical associations can be identified depending on the type of decomposition, but of course such associations might not always imply a causal physical relationship. In their Fig. 8 they analyse association between global compensating error in terms of SW CRE of the cloud regimes and ECS. This is conceptually different from our analysis of association of the RMSE of cloud type occurrence probability and ECS.

We added a sentence in the discussion and conclusions: 'The reason why their result is different from ours might be due to a number of factors, such as a small number of models analysed by Schuddeboom and McDonald (2021), a different set of models, their focus on SW CRE errors vs. our focus on the RMSE of cloud type occurrence probability, and a very different cloud classification method.'

4. The fact that the authors chose to assign a 50/50 prior to a null model vs a linear effect model is very important and should be stated in the main text, not just Appendix A. Arguably since the linear-effect model is more complex and carries an additional parameter, it should be assigned a much lower prior, which would change the results (especially using only one model per modelling centre, in which case the most probable hypothesis would switch from the linear-effect model to the null model). Just by eye, the relationships in Fig. 9b-d do not look strong to me so I think the probabilities being estimated by the authors are being skewed by the statistical assumptions. I do agree that the stratiform correlation in Fig. 9a looks impressive (though hard to see due to the colour) but even in this case I would surely not give it a 99.99% chance of being a real relationship, as the authors do!! 80-90% maybe. Due to the sensitivity of their probabilities to the arbitrary choices listed in Appendix A, I think it would benefit readers to also quote a standard correlation coefficient for the panels in Fig. 9 at least.

We added a sentence 'We note that for the statistical analysis we used priors for the null and alternative models equal to 0.5 (Appendix A).' to the first paragraph where we mention the statistical analysis (Sect. 4.6). Using equal priors is a common practice. In this case, it does not favour the alternative model because it has more free parameters (three parameters: slope, intersect and error spread) than the null model (two parameters: mean and error spread), which implicitly penalises the alternative model. To supplement the Bayesian statistical analysis, in the second revised manuscript in Fig. 11 and 12 we also include the coefficient of determination, correlation coefficient, 95% confidence interval of the slope, Bayesian probability that the slope is greater or lower than zero, and the p-value of a z-test for the difference of means of two groups of models in the bottom 50% and top 50% of ECS, TCR and cloud feedback. The result of the z-test is 3×10^{-6}, 3×10^{-6} and 4×10^{-4} for ECS, TCR and the cloud feedback, respectively.

A z-test for the stratiform cloud type in Fig. 9a in the first revised manuscript (Fig. 11d in the second revised manuscript) for the difference of mean between groups of models with ECS below 4 K and above is 3×10^{-5}.

What is not included in the statistical estimates is the impact of model interdependence, which is hard to quantify objectively and would have the largest impact on reducing confidence in the identified relationships. As mentioned earlier, this is not an uncommon practice and also applies to other similar studies such as Schlund et al. (2020). We repeated the statistical tests with

a limited set of independent models (AWI-ESM-1-1-LR, CNRM-CM6-1-HR, CanESM5, EC-Earth3P, INM-CM5-0, IPSL-CM6A-LR, MPI-ESM1-2-HR, MRI-ESM2-0, NorESM2-LM and UKESM1-0-LL). The p-value of the z-test is 7×10^{-3}, 7×10^{-3} and 6×10^{-2} for ECS, TCR and the cloud feedback (Fig. S9) for the 27 cloud types, and 5×10^{-3}, 5×10^{-3} and 5×10^{-2} for the four cloud types.

5. Given that Figs. 6-8 don't really show interesting trends with ECS or indeed anything that really stands out, I'd strongly consider eliminating or simplifying them to show whatever it is the authors want to show. Fig. 8 is especially hard if not impossible to see anything in, especially given my comments under (2) above.

In Fig. 6a (Fig. 9a in the second revised manuscript) the error in the cumuliform cloud type stands out quite starkly, together with its dependence on ECS. We performed a z-test for the difference between the cumuliform cloud type error between a group of models with ECS < 4 K and a group of models with ECS >= 4, and the p-value is 3×10^{-4}. We included results of the z-test in the second revised manuscript (Fig. 9a). Fig. 6b (Fig. 9b in the second revised manuscript) for example shows that the stratiform cloud type tends to be increasing with GMST in low ECS models and constant/decreasing in models with high ECS. This relationship is tested statistically in Fig. 9a (Fig. 11 in the second revised manuscript). We think this justifies keeping the figure.

We moved Fig. 7 and 8 to the supplement and made visual improvements in the figures to improve readability. Some useful information that is shown in these figures is that the error in the cumuliform cloud type is attributed to the Cu and Cu+Sc type (and not, for example, Cb).

Minor issues:

The text at line 121 seems to say that only two radiation values were used for each training sample, when in fact it is two vectors of 256 normalised radiation fluxes.

We changed the sentence to 'In detail, the ANN input were samples consisting of two channels of SW and LW radiation (256 values for each channel in 16 by 16 pixel samples), ...'

The text at line 434-445 seems to repeat things that should have been covered in the introduction. All text in this section should do is refer back and say how the presented findings relate to past work. Actually I'm not sure most of the work described here is even relevant to the paper.

We removed the first sentence of the paragraph and integrated the rest into the first two paragraphs of the introduction.

Response to Referee 2 on revised manuscript 'Machine learning of cloud types in satellite observations and climate models'

Peter Kuma, Frida A.-M. Bender, Alex Schuddeboom, Adrian J. McDonald, and Øyvind Seland

November 2, 2022

Dear anonymous referee,

Thank you very much for the second round of comments. Please find our response below. In the following text, the original comments are in **bold**, followed by our response.

Kind regards,

Dr. Peter Kuma on behalf of the authors

Summary

The authors have revised the manuscript to address the reviewers' previous comments, which has generally improved the paper. This includes dialing back the conclusions regarding implications for climate sensitivity, adding much more detail to help readers understand the methodology, and performing the analysis at a pixel-by-pixel level. However, there remain several things that need to be addressed before the paper is suitable for publication.

Major Comments

*** I am still quite unclear on how the ANN works. The input TOA flux data are used to predict the probability of the 4 cloud types at each pixel in each 16x16 pixel domain, but the ground truth labels in most cases will only occupy a small portion of the domain where the IDD stations are. How can the pixels with no ground truth learn anything from the CERES TOA fluxes?**

The reference IDD station data are only available on some pixels of a sample. The ANN uses a U-Net type of network to quantify the cloud type occurrence probability at every pixel of the sample. The loss function is only calculated from those pixels where an IDD station is available. The other pixels do not have any reference during the training process. However, the samples are large enough to contain synoptic-scale cloud patterns in the normalised CRE fields. Therefore, the ANN can learn to recognise these patterns and associate them with certain cloud types. In addition, the value of the normalised SW and LW CRE at any given pixel also provides information about the possible cloud type. Because the samples are centred at random geographical points during the training process, an IDD station will end up located in different random pixels of the sample (on the 16×16 grid). This provides training for all pixels of the sample on the 16×16 grid. The U-Net design of ANNs trains coefficients at a number of size scales of the sample in a number of downsampling and upsampling steps. The coefficients are therefore affecting the whole sample and not only one pixel. In this way, the ANN is able to quantify all pixels of the sample, despite only having training data for a limited number of pixels. Overfitting is prevented by the use of a validation dataset and a stopping criterion, which stops the training process once the independent validation dataset stops improving for three consecutive training steps.

In the second revised manuscript, in Fig. 3 we show that the ANN is able to extrapolate synoptic-scale and global-scale patterns to regions for which it had no training information. In Sect. 4.1 we provide more validation which also includes evaluation over four regions excluded from training: North Atlantic, East Asia, Oceania and South America.

*** I don't understand what is meant by the "application phase." Does this refer to the phase when you deploy the trained ANN to make predictions on unseen data? If so, then why does it only use 20 random samples per day rather than all of the TOA radiation data?**

Yes, the application phase refers to making predictions from unseen CERES or model normalised CRE data. The input in the application has to have the same format as during the training phase, i.e. must be a 16×16 grid of SW and LW normalised CRE. This means we cannot supply the whole normalised CRE fields to ANN covering the entire globe at once. The reasons why we use samples instead of for example the fields on a regular longitude–latitude grid is to avoid issues due to spatial transformation in a geographical projection. Every geographical projection causes deformation which can cause biases in the prediction. We limit this by centring the local geographical projection in the centre of the samples. We also want to prevent the ANN training from recognising geographical patterns such as coastlines instead of training only on the cloud fields. This cannot be fully prevented, but it can at least partially mitigated by training it on randomly located samples.

The reason why we only use 20 samples and not more is due to performance reasons. 20 samples already cover a substantial part of the globe on a single day.

*** Why are there no figures demonstrating the skill of the ANN in predicting unseen data? I see that the ANN is trained on CERES and IDD data in years 2004, 2005, 2007 and 2009–2017, with years 2007, 2012 and 2017 used as a validation dataset. My understanding of how ML studies are typically done is that the data is split into three categories: training, validation, and testing. It appears as though here you have used all of the data for training and validation, but did not reserve some data for doing out-of-sample testing. How can we be sure that the ANN works well on unseen data and has not over-fit to the training data?**

The sentence in our revised manuscript describing the validation was wrong (L260, 'We trained the ANN on CERES and IDD data in years 2004, 2005, 2007 and 2009–2017, with years 2007, 2012 and 2017 used as a validation dataset, representing 25% of the total number of years.'). The training and validation years were strictly separate, with the training dataset consisting of years 2004, 2005, 2009–2011, 2013–2016, and 2018–2020, and the validation dataset consisting of years 2007, 2012 and 2017. We changed the sentence to: 'We trained the ANN on CERES and IDD data in years 2004, 2005, 2009–2011, 2013–2016 and 2018–2020, with years 2007, 2012 and 2017 used as a validation dataset, representing 20% of the total number of years.' We apologise for this error.

The validation dataset was used in the training only as a stopping criterion to prevent overfitting – the training process was stopped when the loss function on the validation dataset was not improved for three consecutive training rounds. Therefore, the input of the validation dataset on the training process was likely negligible.

As mentioned in the point above, we included more extensive validation in Sect. 4.1 which covers temporal and geographical extrapolation.

*** All of the analysis is basically in frequency of occurrence space rather than in within-regime cloud property space. But surely the latter should be a large part of the story. A model could for example get the frequency of occurrence of each regime perfectly right but the cloud properties (cloud fraction, albedo, altitude) within the 4 regimes could be biased. Is there a reason that within-regime cloud properties are not evaluated as well as frequencies of occurrence?**

To some extent the cloud type occurrence probability estimated by the ANN already contains information about cloud properties such as cloud fraction, cloud optical depth and altitude. Therefore, an evaluation of the cloud type occurrence probability is also an evaluation of the said properties, even though they are not treated separately.

We added Sect. 4.5 which analyses cloud properties (cloud fraction, cloud top pressure and cloud optical depth) in models relative to CERES as a function of cloud type. The reason why we did not analyse these were both time constraints and the fact that these properties are relatively hard to compare across models and observations. Some of the difficulty arises from potentially different definitions of cloud fraction in models, and if an instrument simulator is used, it should be consistent for all of the models/observations. This is hard to achieve when it is not possible to run the models and instead we have to rely on what is available in the archives. Even if some of these properties may be consistent in CMIP models, MERRA-2 and ERA5 use other conventions. Normalised CRE which was used in the rest of our analysis was chosen as an input because it is consistently defined in models and observations and known with relatively high accuracy from satellite observations.

Minor Comments

L8: delete "a" before "top"

We corrected this in the revised manuscript.

L55: there are issues with subject-verb agreement ("they...is...has...")

We reformulated this sentence to: 'Therefore, they can be used as a metric for model evaluation which, unlike metrics based on more synthetically-derived cloud classes, is easy to understand and has a very long observational record.' The singular here refers to the metric.

L99: I don't understand this statement about grouping together multiple cloud genera, since throughout the paper the results for 10- and 27-type classifications are also shown.

This was not meant to be a statement referring to all of the classifications, only the classifications of 4 and 10 cloud types. We clarified this in the revised manuscript: "For practical purposes, in our analysis we grouped together multiple cloud genera to a smaller number of 'cloud types' (four and ten), in addition to using the full set of 27 WMO cloud genera."

Section 2: I dislike the organization here. It goes from Methods description (Section 2.1) to Data used (Section 2.2) back to Methods description (Section 2.3). Why not put the Data section first?

We re-ordered the sections so that Data (Sect. 2 in the second revised manuscript) comes before Methods (Sect. 3 in the second revised manuscript). The reason for the original order was that the outline section might be useful to the reader before reading the data section for understanding the general aim. The outline section was added in response to a previous comment asking for a general overview of the method before describing the details.

Figure 2: Suggest calling the radiation fields what they actually are (normalized CRE) rather than "reflected TOA radiation" (colourbar) or "shortwave and longwave radiation (caption). Do the 4 cloud type maps have to sum to 100%? It doesn't look like this is the case. Is there a clear-sky probability?

We changed the colourbar label to 'Normalised CRE (%)' and the caption to 'Shown is normalised shortwave (SW) and longwave (LW) cloud radiative effect (CRE), and the probability of cloud type occurrence calculated by the ANN for the classification into four cloud types.'

The cloud type occurrence probability does not generally sum up to 100% because the cloud types are not mutually exclusive in the station measurements. Each WMO station report can identify up to three cloud genus/species in the three SYNOP/BUOY fields C_L, C_M and C_H. Clear-sky probability was not predicted by the ANN. In the Fig. 2 (Fig. 1 in the second revised manuscript) caption, we added: 'Note that the cloud types are not mutually exclusive, and therefore do not have to sum to 100%.'

L291: "histograms" should be singular

The plural refers to the four histogram in Fig. 5b–e (Fig. 8b–e in the second revised manuscript).

L333-334: Why are you reporting the years like this? It doesn't make any sense, as I noted in my previous review.

The sentence in question: 'It was calculated from the first 100 years for CMIP abrupt-4xCO2, with the exception of two models for which the first 100 years were not available: for MPI-ESM-LR years 1850–1869 and 1970–1989 were used, and for MRI-CGCM3 years 1851–1870 and 1971–1990 were used (1850 is the start of the abrupt-4xCO2 experiment, and the time period is not supposed to correspond to reality).'

The years here correspond to the years as in the product files. This is to allow anyone to reproduce the results. We understand that these are not real years, but rather arbitrarily chosen years which depend on the chosen start of the abrupt-4xCO2 experiment in the model run. Some models do not provide all years in the time period between the first and 100th year of the abrupt-4xCO2 experiment, but they still provide limited time periods in the beginning and end of the experiment. For these models, we explicitly listed those time periods (1850–1869 and 1970–1989 in MPI-ESM-LR, and 1851–1870 and 1971–1990 in MRI-CGCM3). Again those are not real years, but rather what is in the *time* variable in the product files.

We reformulated the sentence to make it more clear: 'It was calculated from year 1 to 100 of the CMIP abrupt-4xCO2 experiment. Some models do not provide all years in this time period. These models are MPI-ESM-LR, for which we used years 1850–1869 and 1970–1989, as in the *time* variable of the product files, and MRI-CGCM3, for which we used years 1851–1870 and 1971–1990. These years do not correspond to real years, but rather an arbitrary time period starting with 1850 used for the abrupt-4xCO2 experiment in these models.'

Figures 18-19: These are completely ineffective and uninformative figures that should be removed. Are responses of the individual 27 cloud types really trustworthy? Even if they are, is examining responses with this level of granularity bringing any new any insights? I doubt it.

The figures were added in response to a comment that an analysis of only four cloud types is not sufficiently different from other cloud classifications. Some readers might still be interested in the large sets of cloud types. The supplementary plots 'geo_cto_historical_10_x.png' and 'geo_cto_historical_27_x.png' show that the larger sets of cloud types have geographically distinct distribution, and therefore it can be expected that different physical processes are involved. They also have clearly different cloud optical depth – cloud top pressure distribution. For atmospheric modelling, it is an important question whether certain physical processes leading to the formation of a particular cloud type is well represented. One of the main points of comparison between models and observations is to provide information for modellers, which they could use for model improvement. More granular results can help to identify particular processes and geographical locations which are responsible for biases. A new insight from the classification for 10 cloud types is that the cumuliform cloud type bias is dominated by the Cu cloud genus (and not, for example Cb, which is also in the same group in the classification of four cloud types). We recognise that Fig. 9 and 10 may be too detailed for most readers. In the revised manuscript, we moved them to the supplementary information. We also made the plots colour blindness friendly in line with the journal guidelines.

L345: "the the"

We corrected this in the revised manuscript (as well as other instances of the same mistake).

L358: What is the P value for high clouds? Which value of P marks the transition from "statistically identifiable" to "not statistically identifiable"? In the previous paper, the Bayes factor (ratio of the two probabilities) was reported, but now just the probability of the null hypothesis is reported. Is there a reason for this?

We just want to clarify first that 'P(...)' in the manuscript refers to posterior probability and not a p-value. We used a value of 5% as a cutoff for statistically identifiable, which is commonly used with p-values and confidence intervals, i.e. 95% probability that the null hypothesis can be rejected. Because of the deficiencies of p-values (Colquhoun, 2014), p-values are usually weaker than the actual probability of the null hypothesis at the same threshold.

$P(M_0)$ for the high cloud type was 0.7. This value is included in Fig. 11 in the second revised manuscript.

Because we use prior probability of 0.5 for the null and alternative hypotheses, the Bayes factor (BF) is simply related to the probability of M_0 as $P(M_0) = BF/(1 + BF)$. We did not include BF in the first revised manuscript in response to referee comments noting that not many readers are going to be familiar with BF. In the second revised manuscript, we include multiple other measures of statistical association and significance in Fig. 11 and Fig. 12 in the second revised manuscript (the coefficient of determination, correlation coefficient, confidence interval of the slope, probability that the slope is greater/smaller than 0, and the p-value of a z-test for the difference of group means).

We clarified the sentence: 'Relation with the high cloud type was not statistically identifiable (probability below 5%).'. In Appendix A, we added: 'For statistical significance we assumed $P(M_0)$ below 0.05.'

L397-414: I don't see the value in this discussion. All of the options discussed would still not allow for an unambiguous estimate of the response of clouds to global warming relevant for ECS. This is because of spurious trends in the datasets, the influence of factors other than just global warming (aerosols, most notably) on the trends, the fact that the observed warming pattern over the satellite period is very different from that expected in response to CO2, and other things. I am not aware of any TOA flux measurements on MISR, MODIS, CloudSat, or Calipso, so I'm not sure why those would be used instead of a radiometer like CERES. The idea of running a COSP simulator in a model to generate fields to be run through an ANN to tell you about cloud types seems really bizarre since COSP is already providing detailed information about cloud types. I suggest deleting this paragraph.

The point of the discussion is not necessarily related to ECS. It would be useful to know the trend in the cloud types in the historical record regardless of ECS or aerosols, and if models can reproduce this trend in the *historical* CMIP experiment, which is based on real CO_2 concentration. Aerosol effects do not preclude such a comparison because they are commonly included in the *historical* experiment. Passive satellite instruments generally provide radiance measurements which can be potentially utilised with the ANN instead of normalised CRE – the ANN does not have to operate on flux which is corrected for angular and diurnal variation, but any kind of radiation field which carries information about clouds. We used normalised CRE because it is a very clean field to work with in terms of being corrected for angular and diurnal variations and accurately comparable between models and satellite observations. In the case of active instruments (CloudSat and CALIPSO), 2-dimensional fields (in time and height) of backscattered radiation could be used in the training process. For comparison with models, an equivalent output from the model would be needed, and such output can be provided by an instrument simulator. In the case of normalised CRE as used in our analysis, a simulator is not needed because models provide equivalent fields without the need for a simulator. The reason why radiation information other than TOA flux from CERES would be

used is because other satellites provide a more long-term records, and because active satellite like CloudSat and CALIPSO provide a qualitatively very different view of clouds – passive instruments can only see the top of clouds, whereas the active instruments can see the vertical structure of clouds. Our cloud types are different from other cloud classifications because they correspond directly to the WMO cloud genera, and COSP provides a different classification. Even though spurious trends in long-term satellite datasets such as the NOAA satellite series might preclude the evaluation of trends now, it is quite possible that in the future these datasets will be improved enough to determine trends (options for future research are what we are trying to lay out in this section). There is no reason why our method cannot be improved in the future by applying it on higher resolution radiation fields either from a satellite or future high-resolution climate models. Higher resolution could carry enough information to identify cloud genera much more precisely, and this is something which might not be possible with more traditional cloud classification methods.

We reformulated the paragraph to clarify these points and split it into two paragraphs.

L428: I'm confused. Zelinka et al call what an open question? The previous sentence just looks like a statement of fact – that better present-day cloud properties is associated with larger cloud feedback, similar to what is found in this study. What is the question?

From Zelinka et al. (2022), Sect. 3.4: 'While caution is necessary given the relatively small sample size, an important question is why better simulating present-day cloud properties is associated with larger cloud feedbacks. We leave this as an open question for future research.'

We reformulated the sentence to: 'They concluded that the explanation for this association is an open question for future research.'

In the next line, is it really necessary to directly quote that paper (Zelinka et al) for a fairly mundane statement summarizing the results from another paper (Tsushima et al)? Usually quotes would be reserved for something where the exact phrasing is vital or compelling.

We removed this part.

L438: I believe the correct citation is "Jiménez-de-la-Cuesta" and Mauritsen (2019)

We corrected this in the revised manuscript.

L343-445: This paragraph seems to be all over the place and it is not clear what point you are trying to make.

We removed this paragraph in the second revised manuscript because it referred to Fig. 7 and 8, which are in the supplementary information of the second revised manuscript.

Appendix A: suggest telling the reader what values of P or Bayes factor represent statistical significance (e.g., something analogous to p values being less than 0.05 for a statistically significant result at 95% confidence)

We added: 'For statistical significance we assumed $P(M_0)$ below 0.05.'

Figure B1: I still don't really know what I am looking at here. Is there a way of showing the reader what "perfect" validation looks like? I have no idea what "right" or "wrong" looks like.

The two classifications are different and a perfect match cannot be expected. Something close to the best result would be if the cloud types in CERES/ANN corresponded strongly to specific cloud regimes in ISCCP/SOM. A wrong result would be if the the CERES/ANN cloud types corresponded to many different ISCCP/SOM cloud regimes with about equal probability, or if they corresponded to regimes which are not physically consistent (in terms of cloud optical depth and cloud top pressure) with the cloud type. The comparison is included for providing a link to another cloud classification method rather than to strictly validate our results, because any two cloud classifications are going to be different depending on their definition.

References

Colquhoun David. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. open sci.* **1:** 140216140216. http://doi.org/10.1098/rsos.140216.