Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-125-RC1, 2021

Comment on gmd-2021-125

Anonymous Referee #1

Referee comment on "Parameterization of the collision–coalescence process using series of basis functions: COLNETv1.0.0 model development using a machine learning approach" by Camilo Fernando Rodríguez Genó and Léster Alfonso, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-125-RC1, 2021

General comments

The study introduces a new parameterization of the collision-coalescence process that is based on the results from machine-learning procedures, with an aim to eventually use it in weather forecasting models. The authors utilized 100,000 size distributions of drops (including both cloud droplets and raindrops) to obtain the tendencies (time derivative) of 0^{th}-5^{th} moments, which were used for training a machine (80%) or evaluating the machine's predictions (20%). Each droplet size distribution was assumed to be a composite of two lognormal size distributions, represented by 6 parameters. The paper compares the evolutions of drop size distributions predicted by the machine-learning-based parameterization and explicitly calculated by the method in Bott et al. (1998). The authors concluded that the differences were always less than 10% and therefore it has a promising potential for the future implementation in weather forecasting models.

The overall idea of utilizing the machine-learning method is innovative and aligns with what the cloud-modeling community has started working on in recent years. The results of the study are interesting and provide promising suggestions for the future model improvements. At the same time, the paper seems to require some improvements in its structures and also in providing sufficient information. Most importantly, the conclusions would become much more solid and significant if (i) more than one test simulation is done and/or (ii) if the comparison to an existing parameterization is shown. Regarding (i): although a large number of samples were used for training the machine, the overall evaluation of the new parameterization seems to rely only on one simulation (Table 4), particularly its comparison with the explicit calculation by Bott et al. (1998) under the same condition. The prediction accuracy must be somewhat dependent on each case and it is not known if this one test case falls in the "well-" or "badly-" predicted group. Regarding (ii): the prediction would always have some errors, but the magnitude of the errors is important, particularly in comparison to errors made by other existing parameterizations. Therefore, I think (i) more test simulations to compare the predictions with Bott's calculations and/or (ii) comparison with existing two-moment parameterization is necessary to draw a solid conclusion. I would highly suggest (ii). Detailed suggestions

are listed below.

Specific comments

Lines 12-13: It seems very important to clarify what was calculated and what was predicted/estimated. Since it's supervised learning, the machine did not calculate the moments based on equations, but they must have been calculated in advance elsewhere and the results (inputs & output) were fed into the machine to train it. Afterwards, during the testing/validation phase, the total moments were predicted, not calculated by physical equations, by the trained machine. I understand the overall meaning but the readers may be misled that the machine can analytically solve the SCE and calculate the tendencies of the moments. But in reality, the machine simply gives the prediction based on what it learned before. Therefore, the word "predict/estimate" sounds more appropriate than "calculate".

Line 27: Adding a short explanation on a self-preserving form would be helpful (e.g., what it is, why it gets formed, etc.), especially if this is relevant to collision-coalescence.

Section 2: The structure of this section would become better if it's modified, so that there are 2.1 and 2.2, instead of only 2.1. In my observation, the first section in 2 (that I suggest to convert to 2.1) is dedicated to the time derivative of moments, regardless of collision-coalescence. Subsection 2.1 (that I suggest to change to 2.2) is providing the SCE. Mathematically speaking, I had hard time connecting the two, Eqs. 6 and 13, as Eq. 6 is not mentioned later in the paper, although I understood/knew them individually. Therefore, I suggest that the authors add a few sentences at the end of Section 2 to summarize the entire section.

Lines 211-212: Although mentioned later, it would be better to mention here why the third moment tendency is not calculated.

Figure 4: The figure would be more helpful if the authors instead provide a distribution (line or bar plots) of all the data rather than a scatter plot of every 100 data. Moreover, if the information (e.g., minimum, maximum, mean, median, etc.) can be provided separately for two lognormal distributions on Table 1, this figure can be omitted, as the information overlaps.

Table 3: If the authors can add a column for a prediction score, that would be helpful too, if Matlab has a function to calculate prediction scores. The actual values of MSE may be difficult for the readers to assess the accuracy of the prediction. For example, in the text, MSEs on the other of 10^{-4} are considered to be a good performance, but could you explain this assessment in more detail? For example, above what number is considered a poor performance, and why, etc.

Section 4: I think this section can be included as a subsection of 5.1 in the following Section 5, or even as 2.3 in Section 2.

Table 4: I understand that these conditions were chosen based on Clark (1976), but I think it would strengthen the argument that this case (or f_1) is a good representation of the training data on which the machine was trained, if the authors mention the mean values in Table 1.

Lines 357-358 and Figures 9 and 10: It is difficult to conclude whether the differences between what's predicted by the new parameterization and what's calculated by Bott's

code are small enough or not, only from the figures. However, if you can add predicted values from other existing two-moment parameterizations (one frequently used in weather forecasting models), that would give the readers some insight; in Figure 10, for example, if another parameterization predicts 100 cm^{-3} at t=900s, then the new machine-learning-based parameterization would be a better predicter. Furthermore, if such a comparison can be done for more than one case, the results would become much more solid and substantial.

Table 5 and Figure 12: While the authors clearly state the percentage differences between the predictions and the explicit calculations, its physical meaning also needs a clarification. For example, what does the -8% error of M2 tendency prediction physically mean, and why could it be underestimated by the machine? Even more, for instance, how does this magnitude of errors compare to the errors made by other existing parameterizations?

Section 7: The authors conclude that the overall prediction accuracy was high, but additional analyses and/or a comparison with existing parameterizations seems to be necessary to draw the conclusion. Although the errors in Figure 12 remained less than 10%, how about other existing parameterizations? Would they be within 5%, or more than 50%? I think such a comparison would provide the readers more in-depth understanding and better assessments of the presented ML-based parameterization.

Technical corrections

Please double-check the singularity/plurality of verbs throughout the paper.

Line 9: Either "drop spectra are" or "drop spectrum is". This sentence sounds a bit long,

especially the second clause. It can be shortened.

Line 11: “This basis-function parameterization”

Line 14: “following a uniform distribution” can be omitted. If not, “following” can be replaced by “that has” etc., for example.

Line 22: I think the initial sentence would need a modification. A DSD simply describes the size distribution of a droplet population, and unless it’s fitted into a predefined shape (e.g., lognormal distribution), it can be an exact representation of sizes (e.g., 1000 bins). Therefore, “well” is not necessary unless it’s a fitted distribution (e.g., lognormal). For example, "*Size distributions of droplet populations, namely droplet size distributions (DSDs), are often well represented by a lognormal distribution.*" would have a clearer message, though the authors mention this information later in the Introduction.

Line 23: The publication year is missing after “Marshall and Palmer”. Also, it should be followed by “who” instead of “whom”.

Line 24: “has shown” instead of “have shown”

Line 33: “This type”, does this mean lognormal?

Line 36: DSD was already defined earlier, so it’s not necessary to re-define it here.

Line 43: “a huge amount of equations, which number ranges” can be re-written “a large number of equations, ranging”

Line 47: “where it is introduced” can be re-written “where a simple but… is introduced”

Line 56: “substance” can be re-written “hydrometeors”, and “dependent of” to “dependent on”

Lines 70-72: As it approximates the droplet size distributions by two lognormal distributions, rather than using bins, I am not sure if “This approach simulates the explicit approach” is the accurate description. The strength of the authors’ approach seems to be the time-varying parameters for the two lognormal distributions, in contrast to the conventional bulk schemes, which can be emphasized here.

Line 75: “domain” can be re-written “size spectrum”

Lines 80-83: This sentence seems long and especially after "or being previously…" is unclear. Therefore, I suggest re-writing it more concisely.

Lines 15 & 96: "stablish" can be re-written "establish"

Line 121: An explanation of *N* is missing. Also, the current format looks like "NR^p" is the moment, so it's better to clarify in the text that R^p is the moment.

Line 129: It would help the readers greatly if you add (I=2) at the end of this sentence.

Line 145: The explanations on the equations 11a-c are missing. For example, what is *k* in the equations?

Line 172: "reasonably well"

Line 179: Although mathematically written in Eq. 17, please clarify the meaning or name

of *z* here.

Line 182: “consists of”

Line 189: “the Levenberg-Marquardt optimization” – and is there a reference for this method?

Tables 2 and 5: I think M3 can be omitted here.

Figure 6: Since only the overall decreasing tendency is discussed in the text, rather than the detailed values in this figure, I think the five panel plots can be summarized in one larger plot with 5 lines with different colors, though this is just a suggestion.

Line 256: "because" instead of “due to”

Figure 7: Since the values from the explicit calculations are the “goal/right” values, I think they should be plotted on the y axis rather than on x (i.e., suggest swapping x and y axes). Also, the plots would look better if the x- and y ranges are identical within each plot (e.g., the plots for M1 and M4 seem to have different ranges for x and y axes).

Line 344: “where is a …” to “ where a related… is seen”

Figure 11: Though this is a small point, it would be better for the two panel plots to be placed top-and-bottom instead of left-and-right, as they share the x axis.

Line 381: Either “objective to further test” or “objective of further testing”

Line 419: “loose” to “lose”

Line 432: “fact that leads to an improvement in precision” what does this mean? Precision of prediction? I think this sentence can be shortened.