Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-393-RC1, 2021

Geoscientific Model Development Discussions

Interactive comment on "Physically Regularized Machine Learning Emulators of Aerosol Activation" *by* Sam J. Silva et al.

Anonymous Referee #1

Received and published: 7 January 2021

This is an interesting paper reporting an effort of applying three machine learning regression methods, including ridge, gradient boosted regression trees, and all-connected neural networks, to derive a parameterization for aerosol activation that mimics the behavior of a physical parcel model, the Pyrcel model by Rothenberg and Wang (2015), largely in analog to the procedure in Rothenberg and Wang who used a probabilistic collocation method combining with a Latin hypercube sampling. The authors found that when certain parameterization or simplified model of aerosol activation were introduced as an additional constraint in training, the performance of trained machines would be significantly improved. Overall, the authors have demonstrated that, as orchestrated by several other efforts, machine learning algorithms can be powerful tools for deriving or rederiving parameterizations for physical processes. The paper is

well structured. The results are presented in a good order and also informative. Nevertheless, certain points need to be addressed or clarified before the paper can be accepted for publication.

One interesting result shown in Fig.5 is that a high-capacity machine (XGBoost) trained physically naively can simply match the performance of the one trained with physical constraint. From the perspective of statistics, using a physical constraint in training is simply to provide a better defined a poster scope so that the machine can be trained to easily reach a desired performance with low training cost. However, as far as the base model is regarded as the ground truth, a well-performed machine could be trained without such constraint, as demonstrated in Fig.5 credited to the authors. Generally speaking, the performance of a machine learning model could be optimized with increasing capacity, thus a point here worthy a discussion is whether the cost of coupling a simple model or alternative parameterization (likely with a considerable cost) with a low-capacity model would be better than a high-capacity model alone in application. In this sense, a better purpose of using alternative parameterization here seems just evaluating the alternative parameterization itself.

It was shown in the paper that a number of predicted points by the machines exceeded the physical bound of activation ratio of [0, 1] (more evident in Fig.4). In many cases, this type of outcomes might be a result from use of unnormalized multidimensional features. Firstly, the authors might need to mention the number or ratio of these points. Secondly, had the authors tested training with normalized features? If not, what is the specific reason for not doing so?

Rich resources for machine learning nowadays make the task to understand the sensitivity of targeted outcome to input features much easier. Besides the sensitivity study presented, had the authors used functionalities such as feature selection and feature importance to analyze the sensitivity of the performance of trained machines to the features?

Specific comments.

Line 20-25, the sentences could be rearranged to make the arguments lining up more logically, a suggestion is to move “Cloud formation... Seinfeld and Pandis, 2006)” (Ln 20-22) to ahead of “These aerosol-cloud...” (Ln 25) and modified “Cloud formation” to “It is because that cloud formation”; then change Ln 23 “Hobbs, 2006) and by changing” to “Hobbs, 2006). Aerosols can also change”.

Line 27: “quite” could be removed.

Line 47, “few observations”: did the authors mean “without observations”? If so, the sentence can stand, otherwise, change “few” to “a few”.

Line 48, change “few” to “a limited number of”.

Line 55, “are unable” to “are still unable”.

Line 56, “will longer run times” to “with longer run times”?

Table 1. The caption should include definitions of features, and please change the font and reformat subscript to make them more readable.

Line 118, should use (1) after the equation instead of Equation 1? The same is applied to later equations. Also, please change font size, and also add a space after “,” inside beta ().

Line 132, add “with” after “emulator”.

Line 257, remove one of the two “in”.

Line 286-287, “This strongly...”, as discussed in the previous general comment, the key here for training a better performing machine perhaps is to choose an algorithm adequate for the problem, i.e., nonlinear one for a nonlinear problem.

Fig. 7, Results of activation fraction versus hygroscopicity: what would the high-capacity XGBoost model behave?

Line 311-318, the discussion about training with GPU is adequate, however, the type of chip might not be a central issue for applications of trained machines (just a matrix of coefficients) in practice.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-393, 2020.