We thank the reviewers for their thoughtful comments on the manuscript and have addressed them in detail below. Reviewer comments are in *blue italics*, and our responses are in black normal text.

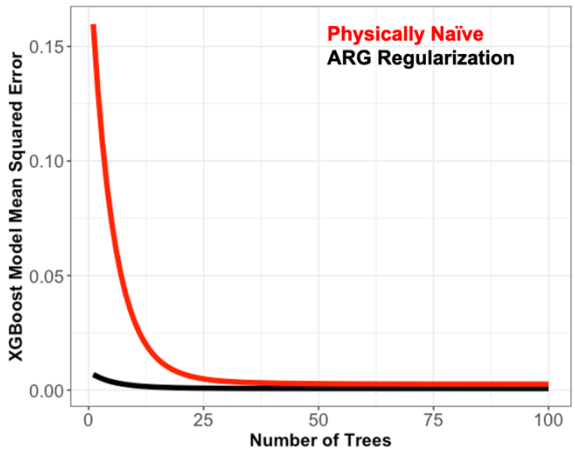
Response to Reviewer #1

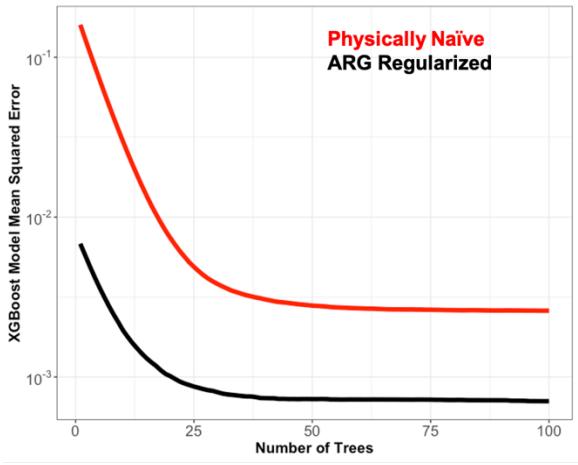
One interesting result shown in Fig.5 is that a high-capacity machine (XGBoost) trained physically naively can simply match the performance of the one trained with physical constraint. From the perspective of statistics, using a physical constraint in training is simply to provide a better defined a 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 in- creasing 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.

The reviewer makes a good point that there is an inherent tradeoff between the capacity and skill of these methods, which does have implications for computational cost. We have updated the manuscript to include this in the discussion of Figure 5 on lines 283-285:

"For a given machine learning technique, increased capacity typically comes with increased computational cost. Including physical information through physical regularization can thus be a computationally efficient strategy for achieving a given model accuracy with lower capacity."

The reviewer has pointed out an issue in our communication of the results summarized Figure 5. While the additional model capacity does certainly improve the performance of these machine learning techniques, the physically regularized model always has better skill. We have updated the figure to now be in log-scale to better illustrate this result.
The difference between the two figures is shown below:





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?

We agree with the reviewer that normalization can be an important step in the machine learning model development. We do normalize the features in this work, and update the text to be explicit in that sense on line 174:

"All features were standardized through a Z-score normalization where the mean was subtracted from each feature, followed by dividing each feature by its standard deviation."

The ratio of points outside of [0,1] is now stated in the text as well, on lines 224-226:

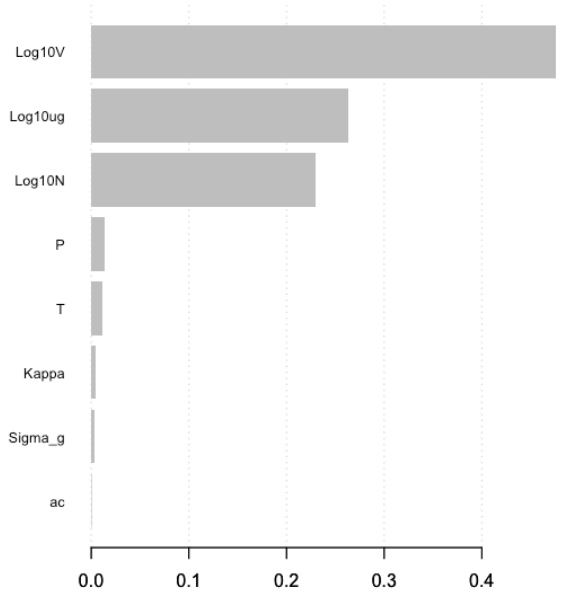
"For cases very near the mass-conserving bounds of 0 to 1 (~10% of the test data), the emulators all predict activation fraction values that extend beyond those bounds. Other than for the linear ridge regression, these deviations outside of the mass-conserving bounds are all very small (less than 0.01)."

While these unphysical predictions are certainly an important caveat regarding model skill, the bounds of [0,1] are reasonably easy to enforce in a large-scale modeling framework (e.g. through clipping).

Rich resources for machine learning nowadays make the task to understand the sen- sitivity 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?

The reviewer raises an interesting point about the potential value of interpretability methods in the development of machine learning emulators. We did not apply them in detail in this work for several reasons. Primarily, we begin this analysis with specific and concrete knowledge of which features contribute to the prediction (i.e., the parent data generating model is known). Additionally, we have a relatively small number of features to begin with, so feature selection techniques for subset selection and computational efficiency were not a design requirement. Lastly, the majority of these selection and importance algorithms (e.g., Layer-wise Relevance Propagation, XGBoost Importance/Gain, etc.) are specific to a given machine learning technique and cannot be easily compared across the DNN, XGBoost, and Ridge methods used in this work.

We show an example of the XGBoost importance metrics below, specifically the gain metric. This demonstrates that while all features are important for prediction (values > 0), certain features that the parent model is most sensitive to (e.g. vertical velocity and aerosol population parameters) are more prominent in the XGBoost model as well.



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".

Done. Thank you.

Line 27: "quite" could be removed.

Removed.

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

Updated to "a few".

Line 48, change "few" to "a limited number of".

Done.

Line 55, "are unable" to "are still unable".

Corrected.

Line 56, "will longer run times" to "with longer run times"?

Updated.

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

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 ().
Updated, thank you.

Line 132, add "with" after "emulator".
Added.

Line 257, remove one of the two "in".
Done.

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.
Agreed.

Fig. 7, Results of activation fraction versus hygroscopicity: what would the high- capacity XGBoost model behave?
The high-capacity XGBoost model behaves similarly to the physically regularized XGBoost model, with larger errors.

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.
We agree that the training gains are larger on GPUs than the application of already trained models. We keep the statement general to account for other complexities in the potential machine learning pipeline (e.g. online learning).