Geosci. Model Dev. Discuss., 6, C776–C782, 2013
www.geosci-model-dev-discuss.net/6/C776/2013/

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Correction of approximation errors with Random Forests applied to modelling of aerosol first indirect effect" *by* A. Lipponen et al.

Anonymous Referee #2

Received and published: 13 June 2013

This paper introduces a new method for reducing errors in simplified (or reduced) models or parameterizations of physical processes used in complex geophysical models. The method estimates the approximation error in a reduced model in comparison to a more accurate model, then applies this error estimate as a correction to the reduced model. It uses the Random Forest technique to construct a regression of this error on the inputs to the reduced model. The use of the Random Forest regression technique for this type of problem is a new and potentially useful method for geophysical modeling.

General Comments

The Abdul-Razzak and Ghan (2002) aerosol activation parameterization for sectional representations (hereafter ARG-2002) does not predict cloud droplet number concentration (CDNC). It predicts number and activation fractions for each size section. The total number of activated particles gives the number of cloud droplet formed in a rising air parcel, which is equivalent to CDNC only in a non-entraining parcel model with no additional microphysics. In 3-D atmospheric models, this droplet formation, in conjunction with cloud microphysical processes and transport, together determine the CDNC. The activation fractions are important for both droplet formation and in-cloud wet removal of aerosol. The output from ARG-2002 is thus a vector quantity, and the paper does not address vector outputs (and errors) at all. Since many (if not most) atmospheric process models have multiple outputs, this is a serious limitation of the method and the paper.

ARG-2002 is itself a simplified or reduced model, and using a first-principles parcel model of aerosol activation as the accurate model would be more appropriate in this study. In addition to the reasons raised by the first reviewer, I am concerned that the simplicity of ARG-2002 (relative to a first-principles model), combined with the relatively smooth aerosol distributions (3 lognormals) used in the paper, may produce an oversimplified representation of the aerosol activation problem. As a result, the performance of the approximation error method presented here may be considerably optimistic. Using a first-principles model as the accurate model, they could also demonstrate how the AE method performs for reduction of the internal numerics of the process model, in contrast to reduction of the input parameters.

Specific Comments

Title. The title is somewhat inappropriate, as the paper contains no results that directly involve the first indirect effect. (E.g., no plots showing droplet number versus aerosol number.) "Correction of approximation errors with Random Forests applied to modeling of aerosol activation" would be better.

P 2552, L 17-19. This point about a large number of variables should be discussed at more length in the results or conclusions. (Also see later comment listing the actual input variables.)

P 2554, L 25-28. Some transition and context for the RF method is needed here. E.g., “The additive approximation error can be estimated using a number of techniques, such as . . .” Here mention other methods (e.g., neural network) in addition to RFs, and their strengths and weaknesses (if possible).

P 2556, L 12-13. Define “x” here.

Section 2.1. Some discussion is needed for when f is a vector quantity.

P 2560, Algorithm 2 definition. Notation. Use a different index variable for trees, such a j, since k is used for other things. (This will just make it easier to follow.) Steps 3 and 8. These need to be explained in considerably more detail (here or in discussion that follows). Step 7. Maybe change to “For each of the split variable candidates, construct a random set of Nsplitp split threshold values.”

P 2560-2561. I was not able to fully understand the Algorithm 2 description, and I suspect that many GMD readers will have the same problem. The description needs to be lengthened to clarify some of the details in the algorithm steps. Making it 2-3 times longer would not lengthen the paper very much. Algorithms 1 and 3 seem much simpler in comparison, so more detail is appropriate for Algorithm 2.

P 2561, Eqn. 5. It is unclear why Eqn. 5 is an equality, while Eqn. 3 is an approximate equality. If Eqn. 5 is really needed, then more explanation of why Eqn. 3 and 5 differ is needed. Also, it might be clearer to introduce an “epsilon-tilde” symbol for the approximated error, and reserve “epsilon” for the exact error as defined by Eqn. 2.

P 2562-2563, Section 3. A better description of the activation parameterization is needed. The paper says they are using ARG-2002, but the Kokkola et al. (2008) SALSA model paper describes modifications to ARG-2002. Are they using ARG-2002,

which assumes dN/dlogD is uniform within a section, or modifications to it? Do the modifications involve only the calculation of activation fractions for each section (based on the intra-sectional size distributions), or the calculation of maximum supersaturation also? Given that the 4-section model consistently underestimates maximum supersaturation (P 2565, L 15-17), this second type of modification could be important and could reduce the quite large errors.

P 2563, L 12. If the ARG-2002 output is treated as CDNC, then some discussion about assumptions involved is needed. (See general comments above.)

P 2563, L 13-14. Neglecting D < 50 nm particles is not a good assumption at high updrafts and low aerosol concentrations. Why make this assumption, as it makes the accurate model even more reduced?

Section 4.3. The input variables to the reduced models should be listed. I would expect something like aerosol number, mean size, and hygroscopicity for each section, plus T, p, and w. This gives 15 and 24 inputs for the 4 and 7 section models, so how does the Nipcands=25 work with this. Also, given the last sentence of the abstract, there should be some discussion (here or in conclusions) about this being a large number of input variables.

P 2566, L 9-10. Provide some rationale for these values of Ntrees, Nipcands, Nmaxsamples, and Nsplitp.

P 2566, L15-20. Error metrics. I suggest using RMSE rather than MSE. It provides the same information, but in units that are meaningful to atmospheric scientists. Also state the mean f from the accurate model, to put the RMSE values in perspective. (Or you could normalize the RMSE by the mean of f.) Some bias statistic should also be included.

P 2567-2568. Provide more discussion of how the accuracy and timing depend on the Ntrees, Nipcands, Nmaxsamples, and Nsplitp parameters. This would be of interest to potential users of this method.

P 2567, L 26-28. Give an example of these minor variations due to randomness. E.g, for one of the Ntrees=25 cases, state the ranges and/or standard deviations of the error metrics.

P 2568, L 15-23. The discussion here of the smallest/fastest RF models should be revised. Some things are repeated, such as the .11-.16 and .07-.11 ms times.

Section 5. Can you say more about areas for future work or improvement of the RF technique? One thing that comes to mind is giving different weights to different input variables in the random sampling of Algorithm 2 (more weight where output sensitivity to input is higher).

Minor comments

P 2552, L 2. Define reduced here, or change to "... using reduced (i.e., simplified) models."

P 2554, L 5-6. Use of "measurement model" here was not clear.

P 2552, L 13; P 2555, L 26; Section 3 heading; and other places. "Cloud droplet activation" is incorrect terminology, although frequently used. Aerosol particles activate, and this leads to formation (or nucleation) of cloud droplets. Use "aerosol activation" or "cloud droplet nucleation" or "cloud droplet formation".

P 2555, L 5. Change models to model.

P 2556, L 24-25. Suggest deleting "Note that model (2) is accurate but". Model (2) is never used (why would it be), so this point can only add confusion.

P 2557, Eqn. 3. Swapping left and right hand sides would seem more appropriate. "Epsilon" is already defined, and you are constructing "g-tilde" to approximate it.

P 2558, L 12. Maybe change "are considered" to "are used to approximate g(x)".

P 2562 L 4-5. Change to "In many atmospheric models, this process is parameterized." Some LES and cloud resolving models treat this process explicitly.

P 2563, L 2-4. State the resolution within each subrange (e.g., constant delta-logD).

P 2563, L 15. Change nucleus to nuclei

P 2563, L 20. The mathematical notation here (that f(x) is a real number) is unnecessary. Use simple text.

P 2563, L 26. Add more digits to the accurate model timing value (1 ms).

P 2565, L 4-8. Clarify the description of volume fractions. Is it that for a given mode within a given sample, they are randomly selected but fixed within a mode, and the volume fractions differ between the three modes? Also, what hygroscopicities were used for organic carbon and dust.

P 2565, L 11-13. Suggest you drop the "(x, y)" notation used to describe Fig. 4, as it is unnecessary, and these variables are used for other things. So on L 13, "... the identity line f(x) = f(x)-tilde corresponding ..."

P 2565, Eqn. 8 and 9. The notation involving log(f) could be improved. You are just transforming f and f-tilde to f-prime and f-tilde-prime, calculating epsilon-prime from the transformed f's, then applying the RF technique with these transformed quantities. Very simple, but not clear from the equations.

P 2565, L 26. The notation here (two groups of brackets) is inconsistent with the notation used for inputs to Algorithm 2 on page 2560.

P 2568. It would be helpful if somewhere in this paragraph, you restate the timing for the accurate model, as it was given several pages earlier (so perhaps forgotten by readers).

P 2568, L 15-23. Change "increment of computation time" to "increase in computation time". The increments (due to the AE calculations) are .07 and .04 for the 7 and 4 section models.

Figures. Axes labels should be larger size in many figures.

Fig. 2, 3, 5. Add labels to vertical axes.

Fig. 4, 6. I would expect to see horizontal axes used for accurate model results (the truth), and vertical axes for reduced model results (the approximations).

Interactive comment on Geosci. Model Dev. Discuss., 6, 2551, 2013.