Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-229-RC1, 2018

Geoscientific Model Development Discussions

Interactive comment on "Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10" *by* Christoph A. Keller and Mat J. Evans

Anonymous Referee #1

Received and published: 11 December 2018

Review of

Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10 by Keller and Evans

Overview

The paper reports on the application of a machine learning (ML) approach (random forest regression, RFR) to replace the calculation of the chemical mechanism in the GEOS-Chem chemistry model. The RFR technique is able to reproduce the standard

application of GEOS-Chem for a 1-month simulation with a good degree of agreement. No computational optimisation has been carried, in fact the presented ML approach increases the cost by a factor 2.

General remarks

ML approaches are more and more tested to improve computational performance of Earth-System models. As the paper seems to be the first application of ML in a global CTM the therefore pioneering effort fully justifies publication in GMD. However, a revised manuscript should provide more detail on the following two main points and the more specific point raised in the specific remarks.

1) More detailed discussion of the motivation, limitations and prospect of the ML approach.

2) More detailed explanation of the RFR technique and its implementation of the model.

to 1)

ML is used to reproduce relations between input and output data sets for which there is (i) no quantitative understanding of the processes,i.e. no "model", or (ii) to reproduce complex model simulations in a cost efficient way. I think the current paper belongs more to the latter category. Therefore the reported increase in cost is not yet an indication that ML will find a place in CTM modelling. Avenues for cost saving of CTM simulations could be discussed with more concrete detail, especially approaches to optimise the application of chemical mechanisms by using reduced schemes in area away from the polluted regimes.

Apart from forecast applications, CTMs are often tools to study the response to variations in emissions on atmospheric composition. It is therefore an important question to know to what extent the accuracy of the ML scheme, i.e. good agreement with the standard CTM, can be maintained without the need to retrain the RF when emission change within a certain range. An actual evaluation of such a test (say 10% reduction

in emissions) could be a welcome extension of the paper.

to 2) A common GMD reader (as I am) may not be familiar with the RFR technique, which is why an extended explanation of the approach would be welcome. Further, the implementation in the CTM should be explained in more detail. For example, it was not always clear if the RFR is trained for all grid points, i.e. composition regimes, together or if a spatial or temporal stratification was applied and specifically trained algorithm would be called for each grid point.

Specific comments

P1 L12: please clarify that “error” is not error w.r.t observations

P1 L20: “85% slower” is not completely clear. Better say factor 2 more expensive or increase in time.

P2 L5: Provide unit of species to check consistency of formulae (density). Consider adding diffusion term if the equation is meant to refers to grid box mean.

P3 L24: As photolysis rate computations can be costly, please explain, why photolysis rates were not subject to the ML approach. They might be a good candidate too.

P3 L25: Please clarify what “each training data set” means - each model grid point ?

P3 L26/27: It is not clear what the numbers in brackets represent (space and time dimensions ?). The number given for the space and time gridpoints of a CTM simulation is not surprising. Please clarify if these numbers are untypical or a challenge to the development of RFR.

P4 L5: please explain the underlying idea of RF in a couple of sentences

P4L13: Why 30 trees ?

P4 L17: What is the computational cost / time to build the forest?

P5 L10: Please clarify what the “->” terms mean in Figure 1, the respective photolysis

rates ?

P5 L11: Why does O3 itself does not appear in the feature importance list? Why is NO3 photolysis more important than NO2 photolysis for ozone?

P5 L14: Does this refer to the troposphere as a whole or to specific regions?

P5 L17: I am not sure “misleading” is the right word here. It is just a manifestation of the fact that relative ozone tendencies are small.

P5 L31: Explain the impact of SO2 (as aerosol precursor ?)

P8 L20: Why is there no stronger increase in NOx when no chemical loss occurs in Figure 8?

P9 L17: I would argue this is exactly the same of a chemical mechanism. There is no dependency between grid point especially as column depends (ozone) of the photolysis calculation has been taken out of the equation. Please clarify, why this is a unique feature of the ML approach.

P9 L20: Please comment on memory cost of the RFR implementation. The demands for floating point operation and memory determines the type of suitable architecture.

P10 L5: I think this point has a large potential, i.e. to select the complexity of the chemical mechanism according to location. In remote areas less demanding chemistry calculation might be sufficient, which are the majority of the grid points. (There might be issues with latency in the parallelism if specific points have more complex schemes to solve)

P10 L22: Please specify to what extent the current RFR and standard implementation conserves the stochimetry.

P11 L4: Please provide more detail why this is easy and would be successful as small changes also accumulate over time.

P11 L22: Please quantify in more detail what amount of changes in emissions would be acceptable without need to retrain the RFR.

P11 L26: Despite the computational cost there is a fundamental reason why ensemble approaches do not have the same importance in AQ forecast than in weather forecasting, namely the less chaotic dependence of AQ forecast on AQ initial conditions. But I agree that significantly cheap AC simulations would be of great benefit. For example, high resolution NWP model runs could afford to have AC modules embedded, which could be used to simulate the impact of aerosols and chemistry on the weather.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-229, 2018.