Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC2
https://doi.org/10.5194/gmd-2020-445-AC2, 2021

Reply on RC3

Marco De Lucia and Michael Kühn

Author comment on "DecTree v1.0 – chemistry speedup in reactive transport simulations: purely data-driven and physics-based surrogates" by Marco De Lucia and Michael Kühn, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-445-AC2, 2021

Many thanks to Dr. Trinchero for his valuable comments.

Comment: *My first concern is that the limited scope of the proposed numerical exercise is not properly acknowledged and discussed. The proposed surrogate models are, in fact, tested using an over-simplified system consisting of a 1D domain and with a geochemical system defined by a few primary species and two mineral reactions only. Since the focus of the work in on the geochemical part of the reactive transport problem, it appears reasonable to consider a simple 1D system (although real media are characterized by complex flow and transport patterns that have a significant impact on the underlying reactions). However, real geochemical systems are also significantly more complex than the calcite dissolution problem considered here. They typically involve several primary species, a number of aqueous complexation reactions, different mineral assemblages and a number of additional non-linear reactions such as surface complexation, fractionation, redox reactions, etc. In the conclusion section the authors state that "The simplification... only marginally affect the validity of the benchmarks concerning the achievable speedup of geochemistry in a broad class of problems.". I think that this statement is unfactual, since the use of the proposed framework for modelling complex reactive transport problems is not obvious and clearly it has not been demonstrated here. I warmly suggest to the authors to revise the discussion section by clearly stating what are the limits of this study and what is the left challenge required to model more complex and realistic systems using this surrogate-based approach.*

Response: We agree with this comment, also raised by the second reviewer, and reformulated accordingly the discussion in the revised manuscript.

Our original statement about "a broad class of problems" has been now precised. We intended here that there are many real life examples dominated by advection and with only two minerals: for example, flow-through experiments in controlled settings. We added a more detailed discussion of these issues in the reformulated manuscript discussion.

Comment: *My second concern is somehow related to the previous one: I found that parts of the manuscript lack of factuality. It is the case, for instance, of line 280, where it is said that "extrapolating to grids with 10^5 or 10^6 elements, speedups in the order of 25-50 are achievable for this chemical problem.". Besides the fact that, according that what it is stated in the introduction ("Chemistry usually represents the bottleneck for*

coupled simulations with up to 90% of computational time"), the maximum theoretical speedup cannot be higher than 10x, I think that extrapolating from a grid of a few hundreds of cells to a model with a million element is simply not correct.

Response: The projected speedup figures are based on a naive extrapolation to larger grids and preliminary, unfinished simulation results. We accordingly removed this statement.

Our own previous work and different cited authors report runtimes of coupled reactive transport simulations where geochemistry takes up routinely around 90 % of the total CPU time, i.e., with a moderately complex chemistry, but it can be as high as over 99 %. This is now correctly mentioned in the introduction in the revised manuscript.

Comment: *Along the same line, arguing that numerical dispersion is an advantage since "it spreads the perturbation" (line 205) is at least questionable.*

Response: We agree, and reformulated this phrase. What we meant is that this "spreading" is of advantage from the machine learning and sampling of parameter space standpoint. In our data-driven approach, we are using the "true" data from three coupled reactive transport simulations to train the surrogates. In this case, the three different grid resolutions have the exact same chemistry but different hydrodynamic settings, since the different nu introduce numerical dispersion. If you think of transport as being an operator which "perturbs" the previous chemical state, the presence of numerical dispersion has the effect of spreading the aqueous concentrations around the "central values" that they would have if all the simulations were run with no numerical dispersion (assuming here dispersion-free simulations with the same time steps). Since two of our three simulations have dispersion, they introduce in the simulations and hence in the training dataset points which would have otherwise not been included.

Comment: *More generally, the authors should use a plainer language and avoid expressions such as (in capital) "subjected to the VERY same advection equation", "can be considered WITH AN ABUSE OF LANGUAGE state variables" "is NOTHING ELSE THAN the maximum value", etc.*

Response: We changed all these expressions.

Comment: *The list of references is a bit poor and is biased towards works performed by the authors. For instance, in line 25, when a short discussion on HPC is introduced, relevant references to efforts and achievements made to parallelise reactive transport problems should be included. In line 30, a single reference (of one of the authors) is used to introduce the issue of heterogeneity of subsurface formations. Since heterogeneity is not addressed here, I suggest removing this paragraph. Otherwise, just consider that, within the branch of stochastic hydrology, there are several works that have addressed the influence of heterogeneity on reactive transport modelling and thus appropriate references should be included.*

Response: We only partially agree with this critique. This work in fact does not deal at all with HPC, we just wanted to point out an argument which is in our opinion fundamental, even if "culturally difficult" to accept by many in the community. Throwing more CPUs at a reactive transport problem in order to obtain an accurate solution of a complex PDE can be regarded as overkill if one considers that there are many hidden and not so hidden uncertainties in the parametrization of that PDE. Concretely for reactive transport, there are huge uncertainties in the thermodynamics, kinetics, and the actual geological heterogeneity of subsurface. What we are trying to explore here is the possibility to obtain less accurate solutions but at a fraction of the computational cost. Of course, this manuscript only represents some first steps in this direction, and no definitive solution.

Furthermore, we cited all peer-reviewed papers we know of which deal with model reduction and machine learning applied to reactive transport, and specifically substituting the geochemical process with surrogates. If we still missed any further, we would be grateful for any hint. Nevertheless we integrated some further references about HPC, uncertainty and heterogeneity in reactive transport simulations, while removing some own citations.

Comment: *Eq. (1): any reason why not to use a more stable and less restrictive implicit scheme?*

Response: First and foremost, we implemented the same advection scheme used by the PHREEQC simulator in order to have direct comparison when replacing the geochemical subprocess with surrogates. Our focus is namely on geochemistry, not on hydrodynamics. Furthermore, at this early stage we restricted our analysis to fixed time steps in order not have to deal with time as a free variable for geochemistry. Within this context, there is in our opinion no need for a better implementation of advection. These assumptions are now more clearly stated in the paper and also better discussed in the conclusion and future work.

Comment: *"Note that equation 1 is not written in terms of porosity, so that effectively the Darcy velocity is assumed equal to the fluid flux or, alternatively, porosity is equal to unity. This assumption does not have any impact on the calculations besides the initial scaling of the system.". This is unclear or at least I am missing something here. If porosity is constant (as it is in this work) the transport equation can be written either in terms of Darcy velocity or in terms of groundwater velocity. This does not imply any scaling.*

Response: We agree, and precised accordingly this statement in the revised manuscript. The advection is written in terms of groundwater or seepage velocity and the scaling only concerns how to setup the batch geochemical simulations in terms of minerals amounts. We used the convention of moles of minerals per kg of solvent in the porous medium.

Comment: *Methods such as kriging offer error models, based, e.g., on the distance of the estimated point from the nearest training data". This is wrong. Kriging provides estimates of the variance given the distance and a chosen covariance function.*

Response: If one has already performed kriging, then the variogram (or generalized covariance) model is already known and fixed. The error model (the kriging variance) is then cheap to compute. Nevertheless, we amended the sentence to avoid any misunderstanding.

Comment: *Line 235 "We used the max value of each label divided by 10-5 as scale.". This seems to be very empirical. Please elaborate a bit more.*

Response: It was indeed an empirical finding, only noted in the paper since it contradicts the corresponding statements in the documentation of xgboost and, generally speaking, of tree-based regression methods. It was found by trial and error. We have no further explanation for this issue; it could be an issue of the specific software package (xgboost) or of the values of our labels which are very small. These facts are now reported in the manuscript.

Other comments

- *Line 80: change "computations run" with "computation runs"*
  Corrected.

- *Most of the figures lack of axes and scales.*

We added scales and axes labels in the first figure (Figure 2) where variables' profiles are displayed. The following profiles are left unchanged since they only serve as graphical comparison.

- *Last sentence at page 12. This sentence is not clear, please reformulate it.*
Corrected.