Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC1
https://doi.org/10.5194/gmd-2020-445-AC1, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

# Reply on RC2

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-AC1, 2021

Many thanks to Dr. Hammond for his valuable and constructive comments.

**Comment:** *Line 445 - "The simplifications concerning the transport and the coupling itself (stationary flow; pure advection with dispersive full explicit forward Euler scheme; no feedback of chemistry on porosity and permeability; initially homogeneous medium; kinetic rate not depending on reactive surfaces) are obviously severe, but most of them should only marginally affect the validity..." I believe that increasing complexity will more than "marginally affect" the results. The addition of high-ionic strength solution, aqueous speciation with a large pH swing, complex rate expressions for mineral precipitation-dissolution (transition state theory with temperature-dependent rate constants, rate limiters, prefactors, etc.) significantly increases nonlinearity and may degrade ML algorithm performance. It is unclear how these processes can be adequately simplified (i.e. for the physics-based surrogate approach). Perhaps this test problem would be marginally impacted, but for real-world reactive transport in heterogeneous porous media, the robustness of presented ML algorithms is not clear to me. But that is future research and does not diminish the results presented here.*
**Response:** We agree on the poor formulation of this thesis. The corresponding discussion has been thoroughly rephrased.

**Comment:** *In the context of parallel computing, how is training implemented across processes (e.g. in the case of distributed-memory computing)?*
**Response:** We did not address nor face this problem for this paper. We assumed that all the results from the "training simulations" are available from the beginning to the ML algorithm for training and test. This is implicit when we write "pre-calculated training dataset.". Once the whole dataset is available, it depends on the regressor and its implementation whether or not the training can be run in parallel or not, and on which hardware. In the provided examples and datasets we always used only one CPU except for training of xgboost which used 4 threads in a shared memory machine, as reported in the manuscript.

This question is of course appropriate when considering a parallel reactive transport architecture in which the training of ML model begins at runtime together with the coupled reactive transport simulation. However it's not really possible to comprehensively answer this question, because it is largely dependent on the actual method used for regression, and specifically if it supports "incremental training" or not, and on which architecture it can be parallelized (GPU, shared memory, HPC cluster-MPI parallelization).

Specifically for our physics-informed surrogate approach, at the moment we did not implement any "forests of trees" based on random subsets of the original training data set. This would be an obvious way to achieve parallelization for both the learning and the prediction phase.

**Comment:** *Line 21 "Chemistry usually represents the bottleneck for coupled simulations with up to 90% of computational time". Bear in mind that if only 90% of the run time if chemistry, infinite speedup of chemistry by ML will only provide 10x speedup overall.*
**Response:** We were imprecise in this sentence. 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 well over 99 %. We rephrased accordingly.

**Comment:** *Line 26 -- "parallelization alone still is not sufficient to ensure numerical convergence of the simulations". What is meant by the term "numerical convergence"? Is this convergence to an accurate solution? Grid or conceptual model refinement can improve accuracy, but HPC enables the solution of more unknowns in less time. If the convergence is nonlinear convergence, HPC seldom improves solver convergence and usually degrades it (hopefully slightly).*
**Response:** Agreed, we were again imprecise in this formulation. We meant - and now also write:

"However, the problem of difficult numerical convergence for the geochemical sub-process, routinely encountered by many practitioners, is not solved by parallelisation."

**Comment:** *Line 49 -- "Any regression algorithm can be employed to replace the "full physics" equation-based geochemical model." If this is true, does it need to be stated?*
**Response:** No, we removed that phrase.

**Comment:** *Line 107 -- Please clarify what is meant by "Charge imbalance and redox potential (pe) can be safely disregarded for this redox-insensitive model..."*
**Response:** We clarified by reformulating the paragraph, which now reads:

"The implemented advection relies on transport of total elemental concentrations instead of the actual dissolved species, an allowable simplification since all solutes are subjected to the same advection equation (Parkhurst et al., 2015). Total dissolved O, H and solution charge should be included among the state variables and thus transported, but since this problem is redox-insensitive, we can disregard charge imbalance and only transport pH instead of H and O,
disregarding changes in water mass. pH is defined in terms of activity of protons:

$$pH = -\log_{10}[H^+]$$

and is hence not additive. If we further assume that the activity coefficient of protons stays constant throughout the simulation, the activity [H+] can be actually transported. The resulting simplified advective model shows absolutely insignificant errors when compared to the same problem simulated, e.g., with PHREEQC's ADVECTION keyword (not shown).

**Comment:** *Line 203 -- There is much discussion early in the manuscript regarding the choice of forward Euler and and CFL=1 to avoid numerical dispersion, but here the authors are attributing the difference in solution between the various grid resolutions to "numerical dispersion". Would it be better to attribute the error to time truncation error? No mass should be diffusing between cells.*
**Response:** To generate the reference simulations and thus the training data for this data-driven approach we used different grid refinements but fixed time steps across all of

them. This means that in the coarser grids we generate numerical dispersion, which in turn makes the simulations actually different hydrodynamical problems rather than different refinements of the same hydrodynamical problem. Since however chemistry is exactly the same across the three grids, this choice allows us to:

1. eliminate the influence of time step;
2. increase the information content in the training dataset;
3. spread the "perturbations" induced by transport (which is now not pure advection but advection + numerical dispersion) to a given partial geochemical result.

If we did run all three reference simulations with their own nu=1 time-step, we would have obtained completely equal trajectories for the chemistry of each grid element, just sampled at different times. Introducing numerical dispersion in two of them also introduces non-linear changes in aqueous concentrations, thus "spreading" the actual sampling of chemistry. Hence, using the three reference simulations as training data "covers more ground" in the parameter space
expected for chemistry, and adds information content to the dataset.

**Comment:** *Line 280 -- Please provide the basis for projecting speedups of 25-50. I believe that it would be better to demonstrate the speedups for 10^5-10^6 elements or with more complex chemistry, rather than to state it. This is especially the case with increasingly complex chemistry as the ML will be more complicated.*
**Response:** These figures were obtained by naive extrapolation of the trend lines of figure 4 and from unfinished work. We removed any speculation about achievable speedups.

**Comment:** *Line 459 -- Please explain how the simplifications only marginally affect the validity of the benchmarks concerning achievable speedup. From previous discussion, it is not clear that the algorithms employed in the benchmarks can handle increased complexity (non-uniform flow velocity, non-uniform grids, nun-uniform time stepping, heterogeneity, ...). If that is the case, the simplifications invalidate the algorithm for complex problems.*
**Response:** We addressed some of these issues now in the discussion. They open up such a large amount of topics that it is not possible to cover thoroughly; however, for example with variable time stepping and non-uniform grids, we devised some further improvements which will be object of future work. More details in the next specific answer.

**Comment:** *Line 430 -- How would the authors handle variable time step size?*
**Response:** One first possibility is represented by training the surrogates for a small, fixed time step and just repeat it within one chemistry iteration to match the new required time step, which we did for the physics-informed decision tree approach. Of course this would only allow simulation of multiples of the "training time step". Another immediate idea, based on the previous one but extending it to non-multiples of the training time step $Delta t_t$, is represented by performing interpolation of the predicted outputs between the $(n-1)cdotDelta t_t$ and $ncdotDelta t_t$ inner iterations (the samples used for interpolation may be more than two, of course) such that the required time step falls within that interval. There would be need to account for non-linearity of the interpolated variables and mass and charge balance would need to be reassessed.

A more general way would be to treat time step as completely independent variable. This would mean that "time step size" is now an "input", a separate column in the training data required for the surrogate to operate. This would of course require a quite large dataset for training, which can be mitigated, in the reference reactive transport
simulations with variable time stepping, by simply outputting partial results of chemistry within a required large time step.

We outlined these ideas in the discussion section, clearly labelling them as "future work".

**Minor comments**

*- Line 56 -- beforehands -> beforehand*
Corrected

*- Line 123 -- How is the term "state variables" an abuse of language?*
Rephrased

*- Line 162 -- A figure showing the inhomogeneous distribution of parameter space for a simple, real problem scenario would be a nice addition to the paper. Would Figure 5 work?*
While this is a valuable suggestion for future work, we do not believe it belongs in the manuscript.

*- Line 310 -- "...is the first, natural feature engineering tentative." Is a bit confusing to me. "tentative" is an adjective. Could it state, "is the first feature to be engineered"?*
Rephrased

*- Line 312 -- "..., but also not relying on." Reword?*
Amended

*- Line 339 -- in facts to -> in fact to*
Corrected

*- Line 379 -- If a different "learning" is required for each grid, please elaborate regarding what "learning" would be required for an unstructured or non-uniform grid? Would each grid cell have its own learning? Such a discussion would make the paper more informative for real-world applications.*
Added a paragraph to the discussion.

*- Figure 12 (Line 439) -- Moving the upper (Reference/Surrogate) legend to the middle of the right side would improve the figure.*
Corrected

*- Line 441 -- with the own -> with their own*
Corrected

*- Figure 13 (Line ~442) -- What causes the oscillatory behavior for pH at late times for Grid 200?*
These are inaccuracies introduced by the surrogate.