Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2020-445-RC2, 2021

Comment on gmd-2020-445

Glenn Hammond (Referee)

Referee 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-RC2, 2021

"DecTree v1.0 - Chemistry speedup in reactive transport…" demonstrates the performance of two machine learning algorithms relative to PHREEQC (accessed through R-PHREEQC) for reactive transport motivated by Engesgaard and Kipp's Test Case B from 1992. The algorithms include: (1) a fully-data driven backbox approach that trains multivariate regressors with PHREEQC and (2) a physics-based surrogate model that predicts future states based on nonlinear extrapolation of the current state through a trained decision tree. The latter is termed "feature engineering". Both algorithms reproduce the PHREEQC result rather well. Algorithm #1 provides speedups up to 4x on the finest grid with error <~1% while #2 provides up to ~7x speedup with 0.3% error.

This model paper has technical merit and details new science, the feature engineering in algorithm #2. The authors acknowledge the simplifying assumptions in the conceptual model. For instance,

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.

Specific Comments:

In the context of parallel computing, how is training implemented across processes (e.g. in the case of distributed-memory computing)?

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.

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).

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?

Line 56 – beforehands -> beforehand

Line 107 – Please clarify what is meant by "Charge imbalance and redox potential (pe) can be safely disregarded for this redox-insensitive model…"

Line 123 – How is the term "state variables" an abuse of language?

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?

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.

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.

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"?

Line 312 – "…, but also not relying on." Reword?

Line 339 – in facts to -> in fact to

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.

Line 430 – How would the authors handle variable time step size?

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

Line 441 – with the own -> with their own

Figure 13 (Line ~442) – What causes the oscillatory behavior for pH at late times for Grid 200?

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.