Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2020-445-RC3, 2021

Comment on gmd-2020-445

P. Trinchero (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-RC3, 2021

The manuscript presents two different approaches to solve reactive transport problems using surrogate models. The first approach relies on purely data-driven regressors whereas in the second approach the system is informed by physical-based rules. The first approach should be considered as a black-box, although geochemical calculations are performed on demand when the mass balance error exceeds the selected tolerance, and as such its predictive capability degrades rapidly in the out of sample region. The physical-based approach provides a compartmentalization of the parameter space and thus allows simpler univariate regressions to be employed. More importantly, by tightening the system to the underlying physical/chemical processes, the physical-based surrogate model offers more reliable predictions, even in the out of sample region. The drawback of the physical-based approach is that the definition of the underlying rules might be tedious, which makes it less attractive for practitioners.

I think that the work is interesting for the geoscientific community and the manuscript is generally well written. However, I do have some major concerns that should be carefully considered and possibly addressed before the ms is accepted in its final form.

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.

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

Other minor comments are stated below.

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.

Line 80: change "computations run" with "computation runs"

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

"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.

"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.

Line 235 "We used the max value of each label divided by 1·10-5 as scale.". This seems to be very empirical. Please elaborate a bit more.

Most of the figures lack of axes and scales.

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