Originally published as:

European Geosciences Union General Assembly 2017, EGU
Division Energy, Resources & Environment, ERE

Integrating surrogate models into subsurface simulation framework allows computation of complex reactive transport scenarios

M. De Lucia[a,*], T. Kempka[a], J. Jatnieks[b], M. Kühn[a]

[a]*GFZ German Research for Geosciences, Section 3.4 Fluid Systems Modelling, Telegrafenberg, 14473 Potsdam, Germany*
[b]*GFZ German Research for Geosciences, Section 1.5 Geoinformatics, Telegrafenberg, 14473 Potsdam, Germany*

Abstract

We showcase a flexible, extensible yet efficient framework for reactive transport modelling, including the ability to replace "full physics" geochemical simulations with surrogate models for speedup. Surrogates are data-driven models trained on a set of pre-calculated simulations by means of machine-learning methods. We offer also an input-output-error visualization component for interactive assessment and tuning of their accuracy. Our framework, based on open source or freely available software, makes possible complex reactive transport simulations and ease further research on optimized algorithms to tackle many geoscientific problems.

1. Introduction

Numerical simulations of subsurface processes are required to assess feasibility of, e.g., CO_2 or hydrogen storage projects, recovery scenario optimization, geothermal projects, estimation of long-term environmental impacts, nuclear waste disposal as well as forecasting energy or petroleum production. The complexity and the diversity of hydrodynamical, geochemical, geomechanical and thermal processes requires highly specialised software for each subprocess in order to profit from state-of-the-art methods and numerical algorithms. The coupling of these processes - i.e., the simultaneous simulation of concurrent and interacting physical phenomena - can be achieved with "software bridges" which make the specialized subprocess simulators communicate to each other. To this specific end the authors developed a flexible simulation framework [1,2] providing an efficient workflow for integrating diverse data sources and diverse software packages to perform coupled process simulations.

---

* Corresponding author. Tel.: +49-331-288-2829 ; fax: +49-331-288-1529
  E-mail address: delucia@gfz-potsdam.de

In the case of coupled reactive transport modelling [3], which considers subsurface geochemical interactions between migrating fluids and rocks, hydrodynamic simulations are complicated by the transport of an arbitrary number of solute species and the chemical reactions occurring between those solutes and the rock-forming minerals. The phenomenological model used to simulate chemical reactions, either based on the Law of Mass Action or Gibbs Energy Minimization, relies upon a large number of thermodynamic and kinetic parameters and its solution is hampered by non-linearity and stiffness of the resulting differential-algebraic system of equations, often generating numerical instabilities and convergence failures. For this reason, massive parallelization of reactive transport simulators is necessary for speed, but still does not guarantee successful simulations. In general terms, the total computational cost of reactive transport simulations is governed by geochemistry, which can account for up to 90% of CPU time [4]. This is an important limiting factor for their wide application. As result, only relatively simple geochemical systems at significantly lower spatial resolution and timescales than pure hydrodynamics simulations are achievable, mainly on 2D or coarse 3D models, and many relevant geological features such as spatial heterogeneity, are often completely disregarded [5].

However, geochemical reactions are affected by extremely large uncertainties both in the characterization of the subsurface and in the parametrization of the phenomenological model itself. Typically, the kinetic constants and the measurements of reactive surfaces of minerals display an uncertainty which may span across several orders of magnitude [6,7]. Also the thermodynamic parameters associated to geochemical reactions are affected by large uncertainties, especially for high salinity, high temperature fluids [8,9]. Thus, expensive and unstable reactive transport simulations in presence of such large uncertainties appear unjustified.

A promising way to enhance calculation speed is represented by the substitution of the "full physics" complex geochemical simulations by statistical surrogate models trained by machine-learning algorithms on a large set of pre-calculated simulations of the specialised software, as was demonstrated by a proof-of-concept implementation [10]. The surrogates are efficient to compute and smooth, which is a decisive advantage for coupled simulations. The price to pay for such an increase in computational efficiency is a loss of accuracy, which is however justified by the large uncertainties associated with the geochemical processes as discussed above.

This contribution shows recent enhancements in our simulation framework concerning reactive transport modelling, specifically aimed at the inclusion of surrogates instead of standard geochemical models for coupled simulations. Starting point and building blocks of the framework are the MUFITS reservoir simulator [11,12]. It is a freely available research code, based on finite volumes, able to efficiently simulate multiphase flow problems. In order to employ MUFITS as hydrodynamical engine for reactive transport, we implemented a multispecies mass transport module, which is at the moment limited to pure advection, intended for file-based coupling with MUFITS. The open source "full physics" simulator PHREEQC [13] has been extensively used in our research group for geochemical modelling. In particular, we developed the Rphree[1] [8] interface to the high-level language and environment R [14]. This interface served as foundation for the development of the infrastructure for surrogate training as exposed in [10]. The R environment is particularly suited for machine-learning, given its status as leading platform in this domain and the plethora of extension packages implementing state-of-the-art data science algorithms. We therefore decided to use R as the main programming language for this reactive transport infrastructure and implemented the coupling with MUFITS, the transport module, the reactive transport algorithms and some visualization functions in form of R extension packages. For better performance, the most computationally intensive algorithms are written in C++ using the Rcpp interface [15].

Furthermore, we recognised the need for a Visual Analytics tool in order to interactively explore the accuracy of surrogate models. The next section illustrates our approach in this matter. In the following sections, we illustrate with more detail the implemented transport module. Finally, a few test cases of reactive transport simulations calculated with the newly developed framework are given.

2. An interactive visualization component to explore and tune surrogate models

The creation of data-driven surrogates for geochemical models is a challenging task and a fertile research area with many open questions [10]. Namely, geochemical models involve many input-output parameters, thus requiring

---

[1] Rphree is open source software downloadable at: `rphree.r-forge.r-project.org`

ensemble prediction methods to be used as surrogates. Moreover, the process of training a surrogate is difficult to automate since user input is required to decide which response - an output variable - of the "full physics" model has to be reproduced with maximum accuracy by the surrogate. Furthermore, geochemical variables such as concentrations may display discrepancies of several orders of magnitude among each others.

To support this task we designed a visualization tool (Fig. 1) able to interactively explore input-output data of model simulations and to display the error of the surrogates in specific regions of the parameter space. This leverages the SPR (Stacked Parameter Relation) visualization previously implemented in the our GCex program [16]. It is designed to show multiple output responses to combinations of multiple input parameters, while guiding the user towards regions of strongest response model variations or surrogate error. Furthermore, it allows to read out actual values and ranges directly from the visualization. Since the visualization itself is independent of the sampling strategy, it is actually suited for general purpose **input-output-error** visualization, not limited to geochemical models.

The possibility to explore the differences between multivariate surrogates and "full physics" models is crucial to detect anomalies or excessive inaccuracies in specific regions of the parameter space. The interactive analysis helps highlighting which specific regions of parameter space need sampling refinement or perhaps a different prediction model.

3. Advective mass transport module

An advective transport module has been implemented with a first order explicit upwind scheme leveraging the hydrodynamic calculations performed by MUFITS, to which it is sequentially coupled. For advection, explicit schemes of higher order are state-of-the-art due to their accuracy, stability and low numerical dispersion, in particular in complex, heterogeneous and transient settings [17]. However, given the simplicity of its implementation, and within the first phase of development, the accuracy granted by a first-order scheme has proven sufficient for most of our applications. In future, the module could be extended to consider also diffusion and kinematic dispersion or implement more advanced schemes. However, this further development would involve significant programming efforts, which should rather be integrated in the flow simulator for efficiency. As it is now, our toolchain makes possible to tackle many practical advection-dominated problems and, most of all, serves as basis to assess the viability of surrogate models for speeding up reactive transport simulations, which is the main focus of our current and future research.
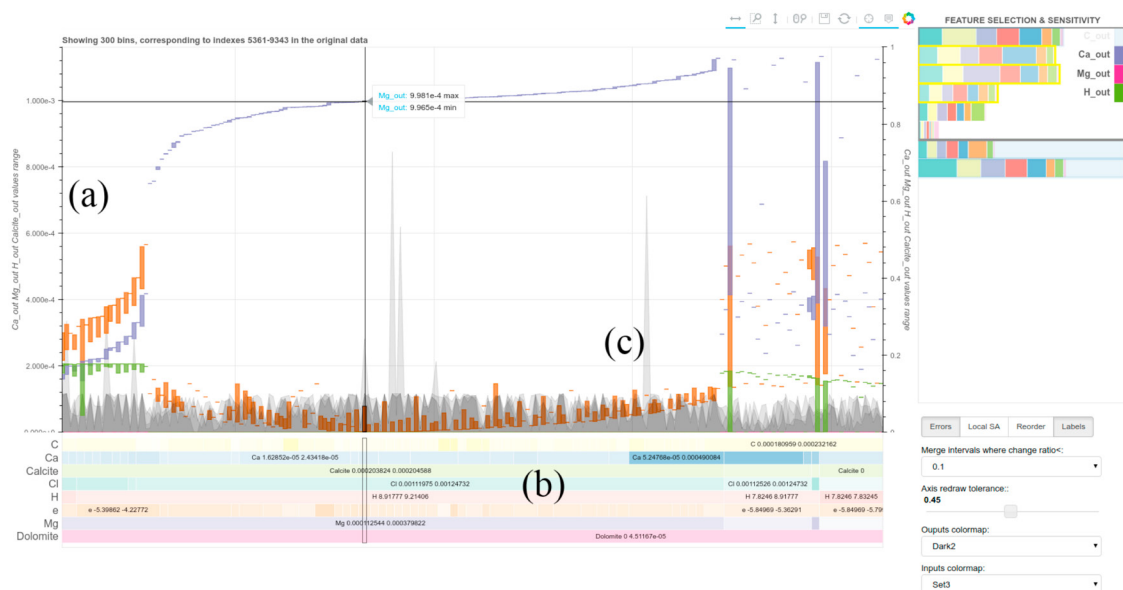

Fig. 1: interactive Visual Analytics component displaying (a) the response of many variables to (b) multiple input parameters (the stacked x-axes) in the "full physics" simulations and (c) the corresponding error of the selected surrogate (shaded bars).

The advantages of the chosen scheme are: simplicity of implementation, calculation speed, as well as guaranteed mass balance (conservativity). The mass conservation equation written in terms of molar concentration C for a grid element indexed n at the time step $t + 1$, neglecting changes in saturation and porosity, reads:

$$C_n^{t+1} = C_n^t + \frac{\Delta t}{\varphi_n S_n V_n} \cdot \left[\sum_{n \to m} \frac{F_{n \to m}}{\rho_n} C_n^t + \sum_{m \to n} \frac{F_{m \to n}}{\rho_m} C_m^t \right] \tag{1}$$

where m is an index covering neighbouring cells, F is mass flux of the transporting (water) phase $[kg \cdot s^{-1}]$, positive if entering the element n, negative otherwise; S is the saturation of transporting phase and ρ is its density; V is the volume of the element and φ its porosity; the subscript $n \to m$ indicates the outbound fluxes from the element n from the neighbouring element m and $m \to n$ the inbound fluxes.

This scheme relies, at each time step, on the fluxes along each connection between each element of the simulation grid; together with element volumes, this represents the only grid-dependent information needed. As result, the same code can be used independently of the dimensionality of the simulation grid and the element type (1D to 3D, tetrahedra or hexaedra, etc.). For the implemented scheme, the Courant-Levy-Frederich condition on allowable time step writes:

$$\Delta t \leq min \left\{ \left| \frac{V \cdot \varphi \cdot S \cdot \rho}{F} \right| \right\} \tag{2}$$

This condition poses quite strict limits on time stepping as compared to implicit or higher order explicit schemes. The scheme however is so efficient that, in the framework of sequential non-iterative coupling with hydrodynamics, it is advantageous to perform inexpensive "inner transport iterations" to reach the hydrodynamic time-step. In view of coupling with chemistry, at least in presence of moderate transient and moderate feedbacks between chemistry and flow, it is also therefore possible to let hydrodynamics or chemistry control the overall coupling time-step, and to run several inner iterations of transport before actually running chemistry.

The transport scheme is both accurate and computationally efficient for moderately transient hydrodynamic problems. It has been tested from simple homogeneous 2D (Fig. 2), 3D (Fig. 3, left) or large 3D multiphase heterogeneous models (Fig. 3, right). With large models such as that in Fig. 3 (right), which depicts the Ketzin pilot site for underground CO_2 storage [18], a transport iteration is performed in few tenths of a second.

In the next section we provide an estimation of the efficiency of transport calculations relative to CPU time spent in chemical simulations.

4. Reactive transport simulations

Our framework is conceived and implemented as a toolbox adapted to design coupling strategies between hydrodynamical and geochemical processes. However, it is also built for computational efficiency, therefore allowing for complex simulations. Having access to virtually all functionalities of the MUFITS and the PHREEQC simulators, the user can setup reactive transport models of multiphase flow profiting from the accurate and numerically efficient thermophysical description of the first program and all the features of the latter one, including mixed kinetic-equilibrium calculations, state-of-the-art thermodynamical databases and the Pitzer activity model for concentrated solutions. Among the strengths of our simulation framework we can name:

- equilibrium and kinetic geochemical calculations in PHREEQC
- surrogates can be plugged-in in the coupled simulations instead of PHREEQC models
- built-in parallelization of chemistry and multispecies mass transport on shared memory computers
- control of the feedback between chemistry and hydrodynamic flow (e.g., permeability change) with straightforward programming
- data visualization in the R environment; data export in standard formats through the VTK libraries for use with external tools, e.g. ParaView

An exemplary 2D geochemical benchmark involving the transport of eight chemical elements and two minerals (the `calcite_pqc` benchmark from [19]) is depicted in Fig. 4. A magnesium chloride ($MgCl_2$) solution is injected in a porous medium homogeneously saturated with water at thermodynamic equilibrium with calcite. With the movement
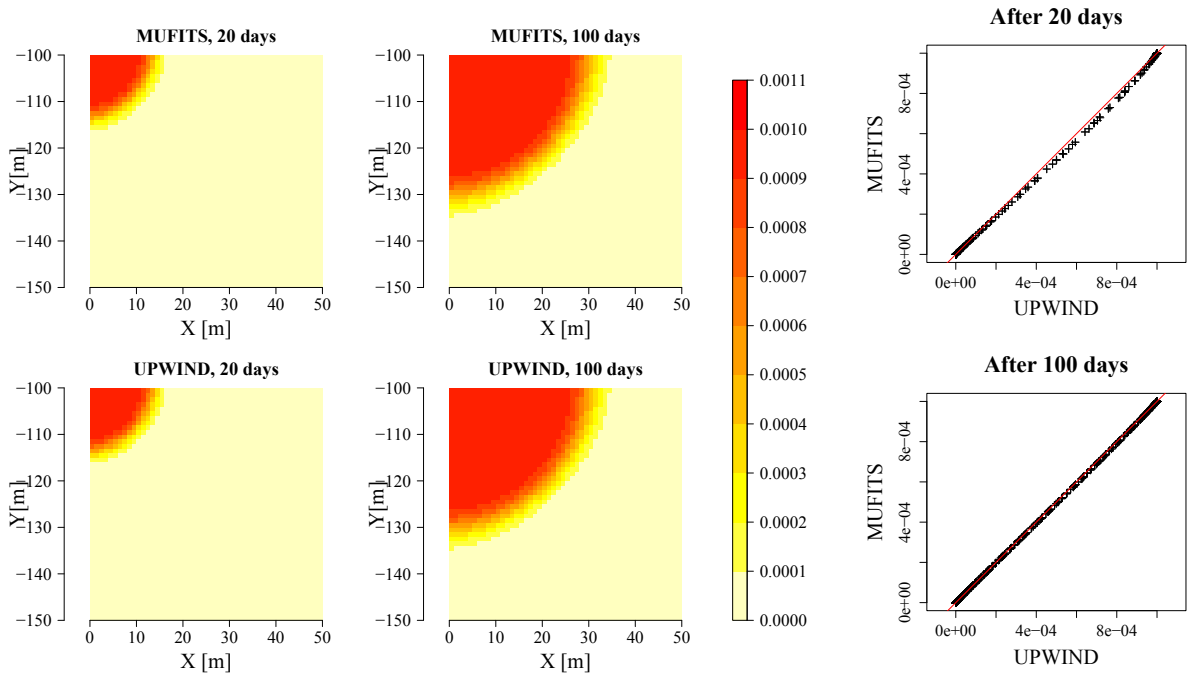
Fig. 2: Comparison of solute advection computed by MUFITS and with the UPWIND scheme implemented in the present work on a two-dimensional model comprising 50 x 50 elements. The model depicts the injection of a tracer in the top-left corner of the model with fixed pressure at the injection point and at the opposite model boundaries. Left: salt concentration in brine calculated through the GASSTORE option of the MUFITS simulator (upper row) and with the transport module, denoted UPWIND (lower row). Right: Scatter plot of the resulting concentrations after 20 and 100 days.

of the injected solution front, calcite starts to dissolve and dolomite is temporarily formed, to redissolve little after the passage of the reaction front. This simulation is performed on the same 50 x 50 rectilinear grid shown in Fig. 2, with injection on the topleft corner of the model and steady state saturated flow. Table 1 charts the total time for the simulations with increasing number of CPUs and in particular the acceleration achieved by ensuring that only the
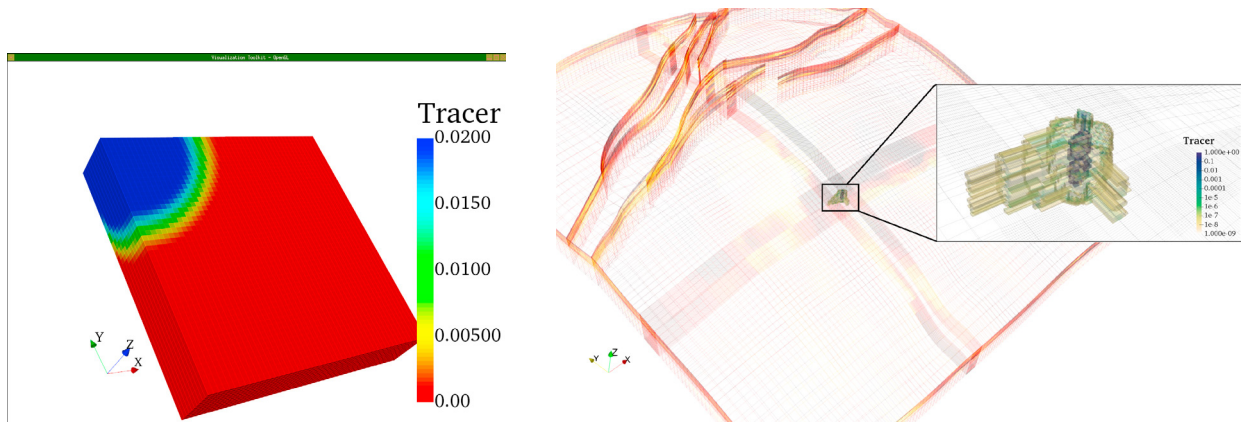

Fig. 3: Examples of transport problems calculated with the presented tools. Left: 3D transport benchmark, interactive visualization based on VTK implemented as a framework component. Right: migration of an aqueous tracer in a CO_2 injection scenario. The heterogeneous grid comprises over 640,000 elements.
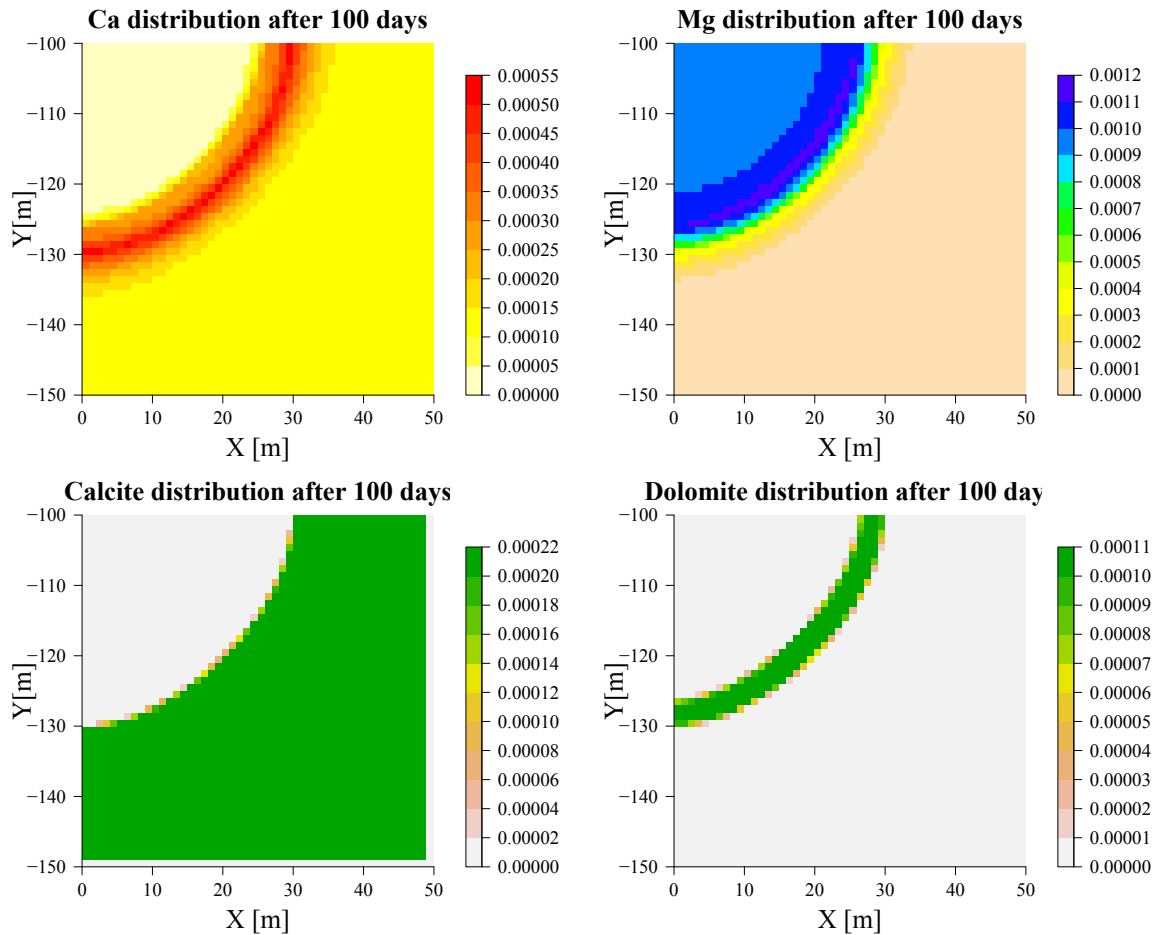
Fig. 4: 2D reactive transport benchmark. A magnesium chloride ($MgCl_2$) solution is injected in a porous medium saturated with water at thermodynamic equilibrium with calcite. With the movement of the injected solution front, calcite starts to dissolve and dolomite is temporarily formed. The depicted simulation time of 100 days has been reached in 79 coupling iterations with constant time stepping.

minimum number of geochemical calculations are performed at each time step. The principle of this algorithm, here referred to as "Reduce", is explained in [8].

It is apparent how the implemented advective transport has a small cost compared to chemical calculations, even if they are extremely simple as in this benchmark. The toolset scales almost linearly with increasing number of employed CPUs, however the overhead due to parallelization is only justified up to 4 CPUs in this simple case. An important

Table 1: Timings of the benchmark reactive transport simulations of Fig. 4 with multiple CPUs and impact of the "Reduce" compression algorithm (see text and references therein). Left: total wall clock time for 79 iterations. Right: CPU time spent in each task for the simulation with "Reduce" algorithm.

Total time	1 CPU	4 CPU	8 CPU
Normal	208 s	57 s	53 s
Reduce	119 s	57 s	56 s

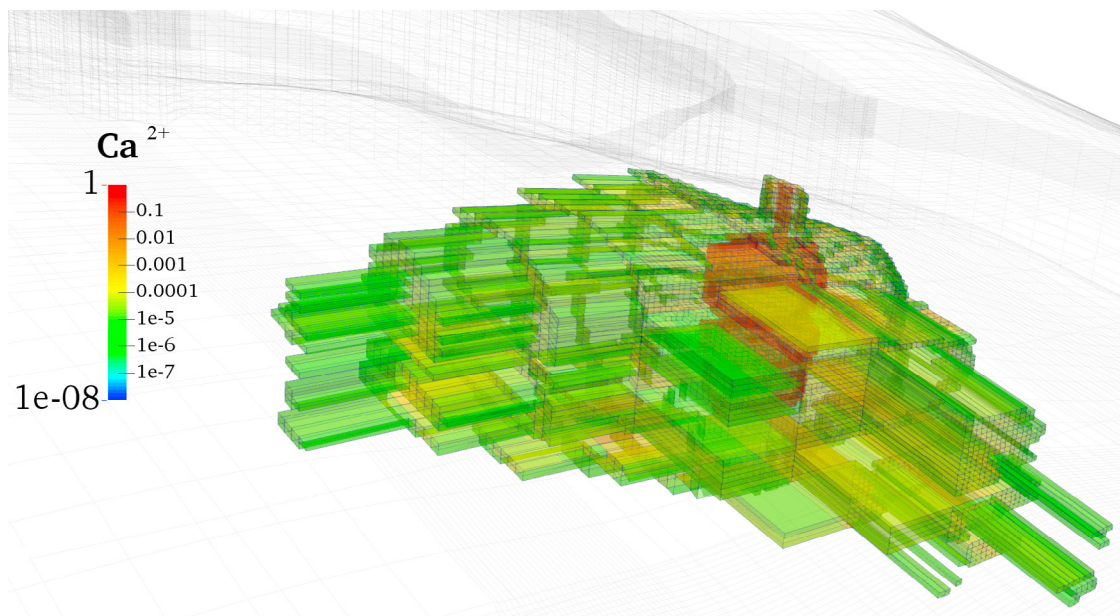|	1 CPU	4 CPU	8 CPU
Chemistry	91.0 s	23.0 s	21.0 s
Advection	0.8 s	1.4 s	0.8 s
Reduction	1.6 s	1.6 s	2.2 s

Fig. 5: Exemplary reactive transport simulation calculated on the same large, heterogeneous Ketzin model of Fig. 3 (right).

factor controlling the achievable complexity of simulations in terms of resolution of the simulation grid (number of elements) is, again, the overhead due to data passing between the hydrodynamical engine and the subprocesses transport, which, in a sequential file-based coupling, is comparable to the CPU cost of chemistry.

However, despite the inefficient file-based coupling, our framework is perfectly able to tackle quite large, heterogeneous simulation grids. An exemplary run of reactive transport model in the context of multiphase flow (CO_2 injection scenario) is depicted in Fig. 5. Here is shown the increase of dissolved calcium following the dissolution of calcite due to contact with injected CO_2. On such a large simulation grid, few dozens time steps were calculated in less than half an hour on a common multicore workstation.

5. Conclusions

We presented a newly developed toolchain for performing complex reactive transport simulations using the high-level language R to implement process coupling. Building blocks of the framework are the scientific reservoir simulator MUFITS and the PHREEQC geochemical simulator. The main novelty here lies in the ability to program tailored algorithms for reactive transport simulations leveraging the capabilities of specialised subprocess simulators. In particular, the framework offers a facility to directly plug-in reduced models for chemistry and to implement and test innovative coupling strategies. Furthermore, it makes much more comfortable to perform tasks such as sensitivity and uncertainty analysis, visualization and postprocessing, since all data are interactively available in the R environment.

Several open questions related to the application of surrogate models still remain unanswered, i.e. how to ensure mass balance between the different chemical species, which is indispensable for reactive transport applications. Specifically to tackle such challenges, which constitute the major future research axes, there is need for a truly extensible simulation toolchain. Only within such a framework it is in facts possible to rigorously assess the errors introduced in coupled simulations by numerical approximations of the submodels.

Furthermore, we presented some benchmarks and complex simulations in order to highlight the usability and the capabilities of the framework. The framework is currently under heavy development; we envisage to release these tools in form of R extension packages upon reaching sufficient maturity.

References

[1] Kempka, T., Nakaten, B., De Lucia, M., Nakaten, N., Otto, C., Pohl, M., et al. (2016) "Flexible simulation framework to couple processes in complex 3D models for subsurface utilization assessment." *Energy Procedia* 97: 494–501. doi:10.1016/j.egypro.2016.10.058.

[2] De Lucia, M., Kempka, T., Afanasyev, A., Melnik, O. and Kühn, M. (2016) "Coupling of Geochemical and Multiphase Flow Processes for Validation of the MUFITS Reservoir Simulator Against TOUGH." *Energy Procedia* 97: 502–508. doi:10.1016/j.egypro.2016.10.060.

[3] Steefel, C., Appelo, CAJ., Arora, B., Jacques, D., et al. (2014) "Reactive transport codes for subsurface environmental simulation." *Computational Geosciences* 19(3): 445-–478. doi:10.1007/s10596-014-9443-x.

[4] De Lucia, M., Kempka, .T and Kühn, M. (2015) "A coupling alternative to reactive transport simulations for long-term prediction of chemical reactions in heterogeneous CO_2 storage systems." *Geoscientific Model Development* 8(2): 279–294. doi:10.5194/gmd-8-279-2015.

[5] De Lucia, M., Lagneau, V., de Fouquet, C. and Bruno, R. (2011) "The influence of spatial variability on 2D reactive transport simulations." *Comptes Rendus Geoscience* 343(6): 406–416. doi:10.1016/j.crte.2011.04.003.

[6] Marty, N.C., Claret, F., Lassin, A., Tremosa, J., Blanc, P., Madé, B. et al. (2015) "A database of dissolution and precipitation rates for clay-rocks minerals." *Applied Geochemistry* 55: 108-–118. doi:10.1016/j.apgeochem.2014.10.012.

[7] Black, J., Carroll, S., and Haese, R. (2015) "Rates of mineral dissolution under CO_2 storage conditions." *Chemical Geology* 399: 134—144.doi:10.1016/j.chemgeo.2014.09.020.

[8] De Lucia, M. and Kühn, M. (2013) "Coupling R and PHREEQC: Efficient programming of geochemical models." *Energy Procedia* 40: 464–471. doi:10.1016/j.egypro.2013.08.053.

[9] Dethlefsen, F., Haase, C., Ebert, M. and Dahmke, A. (2012) "Uncertainties of geochemical modeling during CO_2 sequestration applying batch equilibrium calculations." *Environ Earth Sci* 65(4): 1105–1117. doi:10.1007/s12665-011-1360-x.

[10] Jatnieks, J., De Lucia, M., Dransch, D. and Sips, M. (2016) "Data-driven surrogate model approach for improving the performance of reactive transport simulations." *Energy Procedia* 97: 447-453. doi:10.1016/j.egypro.2016.10.047.

[11] Afanasyev, A. (2013) "Application of the reservoir simulator MUFITS for 3D modelling of CO_2 storage in geological formations." *Energy Procedia* 40: 365–374. doi:10.1016/j.egypro.2013.08.042.

[12] Afanasyev, A. (2015) "Hydrodynamic modelling of petroleum reservoirs using simulator MUFITS." *Energy Procedia* 76: 427–435. doi:10.1016/j.egypro.2015.07.861.

[13] Parkhurst, D. and Wissmeier, L. (2015) "PhreeqcRM: A reaction module for transport simulators based on the geochemical model PHREEQC." *Advances in Water Resources* 83: 176–189. doi:10.1016/j.advwatres.2015.06.001.

[14] R Core Team. (2017) "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

[15] Eddelbuettel, D. and Francois, R. (2011) "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40(8):1–18.

[16] De Lucia, M., Jatnieks, J. and Sips, M. (2015) "GCex: a Visual Analytics approach for interactive exploration of geochemical models." *Energy Procedia* 76: 616-–622. doi:10.1016/j.egypro.2015.07.881.

[17] LeVeque, R.J. (2002) "Finite Volume Methods for Hyperbolic Problems." Cambridge University Press.

[18] Kempka, T., De Lucia, M. and Kühn, M. (2014) "Geomechanical integrity verification and mineral trapping quantification for the Ketzin CO_2 storage pilot site by coupled numerical simulations." *Energy Procedia* 63: 3330–3338. doi:10.1016/j.egypro.2014.11.361.

[19] Kolditz, O., Görke, U-J., Shao, H. and Wang W. (2012) "Thermo-Hydro-Mechanical-Chemical Processes in Porous Media." Springer.