An emulation-based approach for interrogating reactive transport models

Authors: Angus Fotherby[1*], Harold J. Bradbury[1], Jennifer L. Druhan[2], Alexandra V. Turchyn[1]

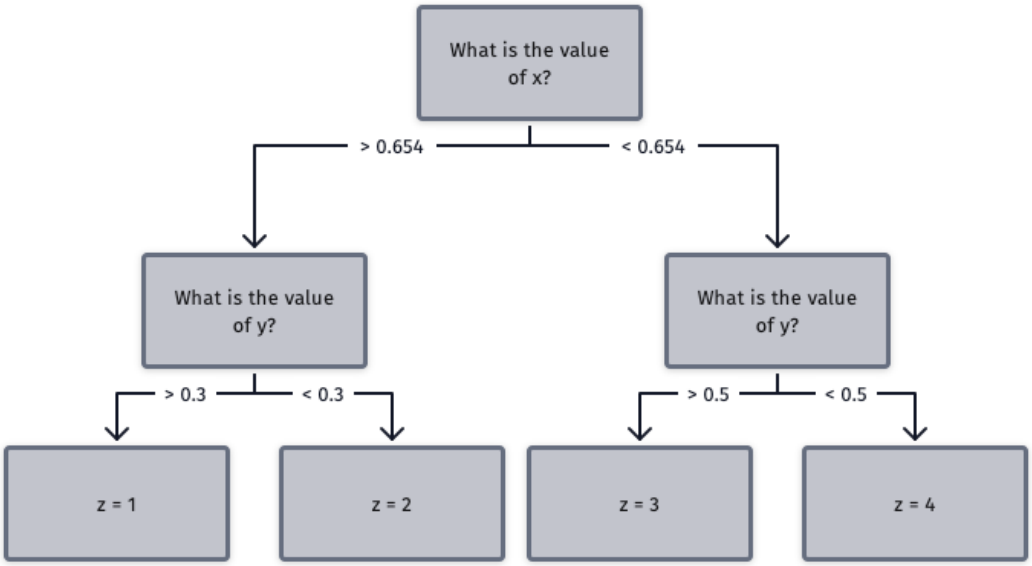
1: Department of Earth Sciences, University of Cambridge, Cambridge, UK

2: Department of Geology, University of Illinois at Urbana Champaign, Urbana, IL

*corresponding author: af606@cam.ac.uk

1 Machine Learning

1.1 Decision Tree Schematic



.

Figure S1: An example decision tree, in which for an input of (x=0.7, y=0.6) traversing the tree returns z=3.

1.2 The Gradient Boosting of Decision Trees

Tree boosting is a statistical learning approach whereby decision trees are iteratively added to an ensemble of such trees, with the aim of predicting a value (y) from a vector of inputs (x_i). At each level of the tree, there is an inequality presented, (for a schematic of this see Figure S1) and the model will traverse the tree based on whether that inequality is true or false for a given input. After the whole sequence of decisions has been evaluated, the tree has been traversed and arrives at an output—the prediction y. By itself, a single tree is a poor predictor of a complex system but tree boosting combines the weighted, averaged output of many trees, which does constitute a powerful predictor (Schapire, 1990). The gradient boosting of decision trees (Friedman, 2001, 2002) uses gradient descent to decide which tree is the best one to add to the ensemble to improve predicative capability. This works by defining an objective function (for example, root mean square error in the prediction) and calculating the tree that most reduces that objective function and adding that tree to the ensemble. By training such a GBT model on the synthetic dataset generated by the original RTM, we create a surrogate model (also known as an emulator) of the original system. These emulators encode all the process interactions that were described previously because they have been trained on a dataset that encompasses the complete range of the chosen variables

1.3 Description of XGBoost implementation

The hyperparameters reported in Table S1: were chosen by inspecting model training performance. We train the model on GPU and so use the GPU implementation of XGBoost, the gpu_hist tree method. We train using the linear regression objective function, as it gave the best results when compared to other objective functions offered in XGBoost. The reg:linear objective means that the squared error is the function being minimised.

The number of training rounds is the maximum number of iterations of the XGBoost algorithm that can be applied during training. The number of training rounds is partly a

function of η, the learning rate, which determines how quickly the emulator converges to a fit. If the learning rate is slow, then more training rounds are usually required, although this has not been the case here. The value of η was determined by experiment. The maximum number of training rounds here was selected to be large to ensure that the emulator was as accurate as possible, given its configuration. This does not mean that the models could not be trained further if more time and resources were allocated to training.

Having an unlimited tree depth, with a limited number of leaves (rather than is usual in XGBoost, which normally has a capped tree depth and an unlimited number of leaves) helped the GBT model hyperparameters generalise across case studies. Similarly, choosing a loss guided tree growth policy (so that nodes are split wherever they will make the best improvement to the loss, rather than always nearest the root of the tree, as is default) was the result of experimentation that showed it improved model performance and hyperparameter generalisability. The maximum number of bins was chosen as large to ensure high resolution when dealing with small changes in mineral precipitation that had to be captured by the emulator.

The maximum tree depth refers to the maximum number of consecutive decisions in any given tree within the ensemble (see Figure S1 for an example tree of depth two). L1 and L2 regularisations are factors that penalise complexity within the emulator. Complexity in this context refers to emulator behaviour that leads to overfitting, where the model learns the data in such a way that the model cannot generalise well to unseen data, while performing extremely well on the data that it has seen during training, which is undesirable when trying to draw conclusions from the emulator. By including factors that encourage the model to not become overly specific to the data shown during training, this overfitting can be avoided – in this case the default values for regularisation sufficed and so are not listed.

Hyperparameter	Hyperparameter meaning	Value
Tree method	Algorithm to construct the decision tree	gpu_hist
Number of training rounds	Maximum number of boosting rounds	50,000
Objective	Loss function to minimise	reg:linear
Growth policy	Choice of how new nodes are added to the tree	lossguide
Max tree depth	Maximum number of consecutive decisions allowed in a tree	Unlimited
Learning rate (η)	Multiplier to scale feature weights to prevent over fitting	0.01
Max leaves	Maximum number of nodes that can be added	6
Max bin	Number of bins for bucketing features	10000

Table S1: GBT model parameters used in each case study. Unlisted hyperparmeters are unchanged from the XGBoost default. For technical details on each parameter see Chen and He (2015).

1.4 Training curves

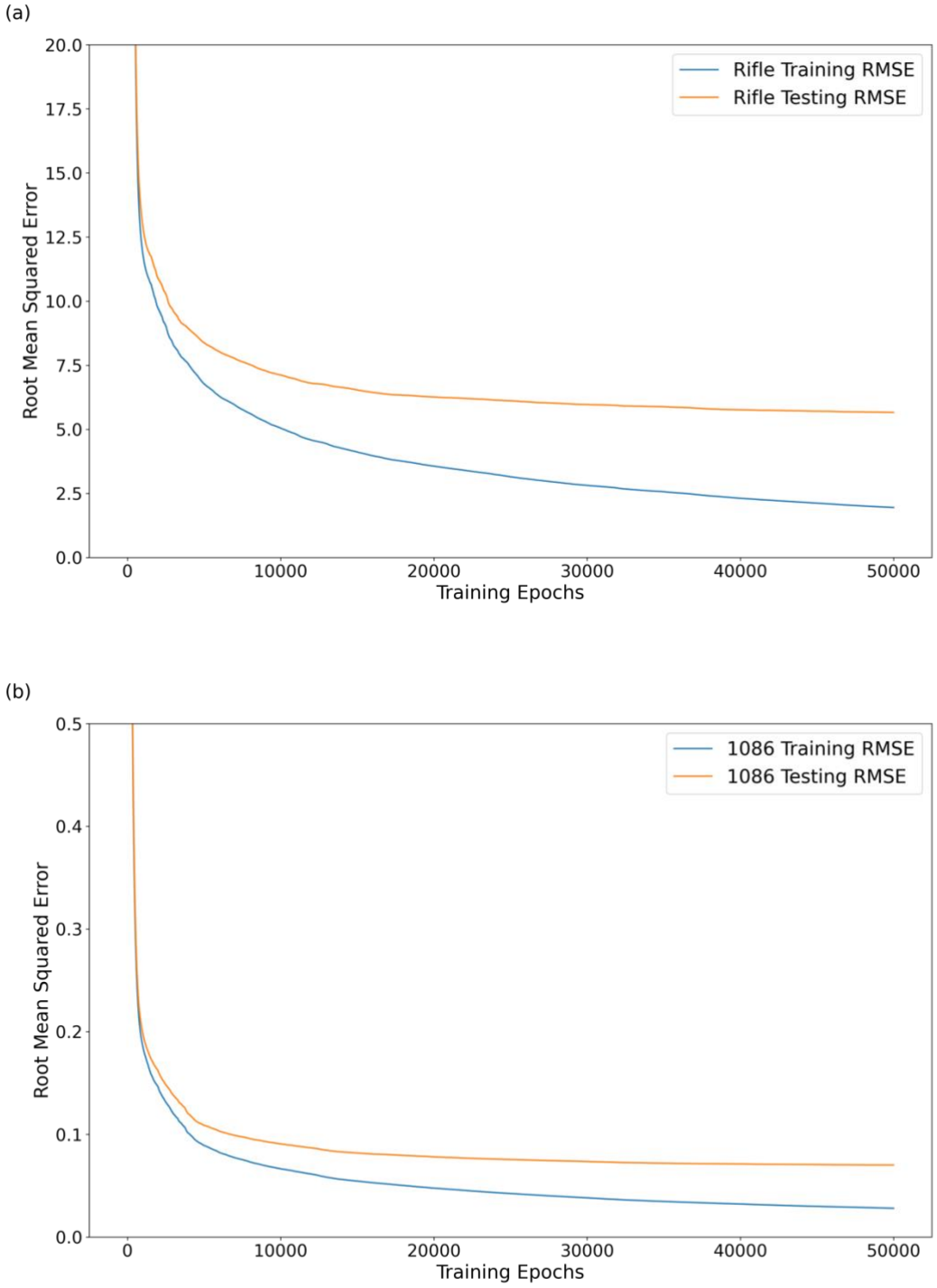


Figure S2: Training curves for XGBoost regressor models for both Old Rifle, (a), and ODP Site 1086, (b). Difference in scale between plots is a result of the difference in scale of the labels being predicted between the different environmental contexts.

2 Old Rifle

2.1 Old Rifle RTM Schematic

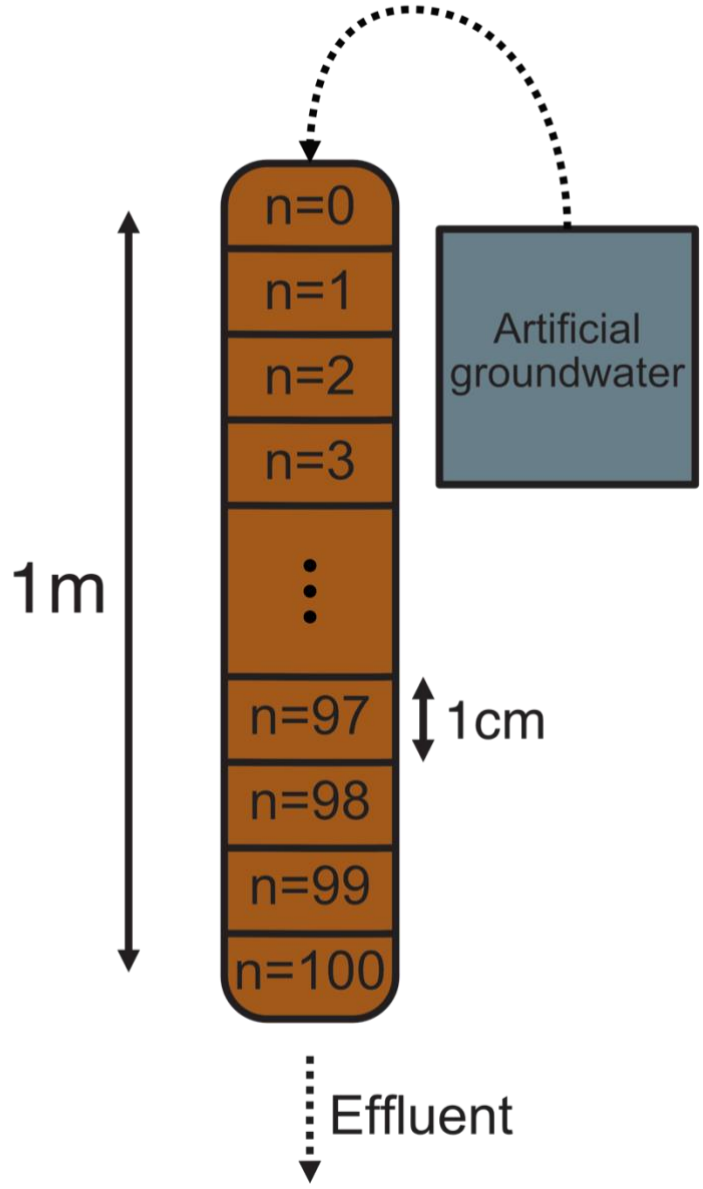


Figure S3: Schematic of the RTM which the emulator is trained on. The RTM describes the column of sediment taken from Old Rifle. The model was developed by Druhan et al., (2014). The RTM is made up of 100 grid cells, each sized 1 cm. The artificial groundwater is injected at the top, and flows through the sediment, eventually out of the bottom of the column.

2.2 Optimised values for pyrite precipitation at Old Rifle

Quantity	Optimised Value
NH_4^+	30 mM
SO_4^{2-}	30 mM
Ca^{2+}	27.5 mM
Acetate	30 mM
$CO_{2(g)}$	0 bar
GBT predicted net pyrite precipitation	0.143
RTM modelled pyrite precipitation	0.150
Difference in predicted vol. frac. change	7.0×10^{-3}
Difference in predicted vol. frac. change	4.7%

Table S2: Predicted injectate fluid composition that would maximize pyrite precipitation at Old Rifle with the associated volume fraction increase predicted by the emulator. We note that the range of predictable values is constrained by the trained range of the GBT model, so in the cases where the value predicted is equal to that maximum (30mM for Old Rifle) the true optimal value could be higher. The penultimate row gives the true net increase in volume fraction calculated by the RTM with these boundary conditions, which we take to be the ground truth.

2.3 Old Rifle co-dependency plots

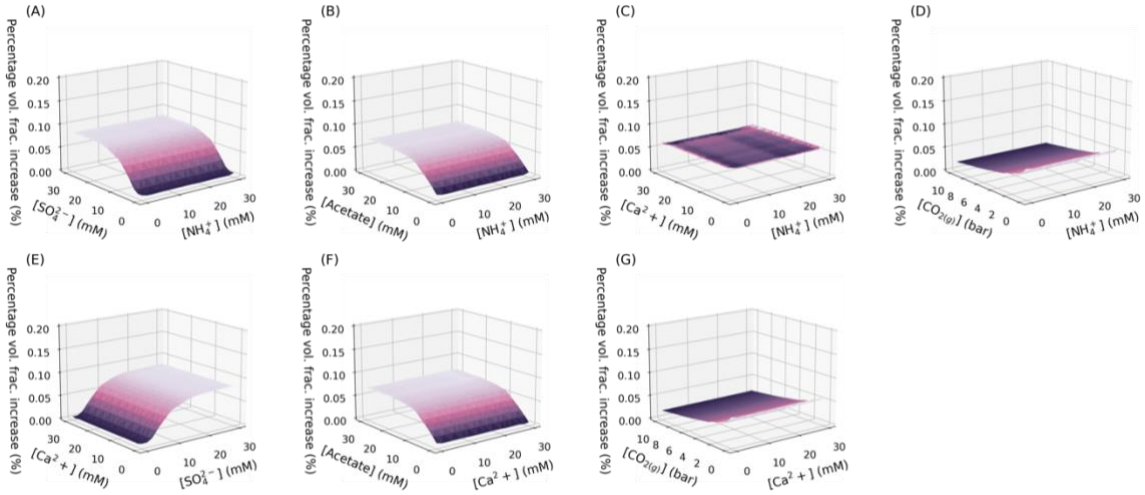


Figure S4: The additional co-dependency plots of the parameters varied for the Old Rifle injection fluid not shown in the main text.

2.4 Effect of rate law on CO_2 dependency in Old Rifle

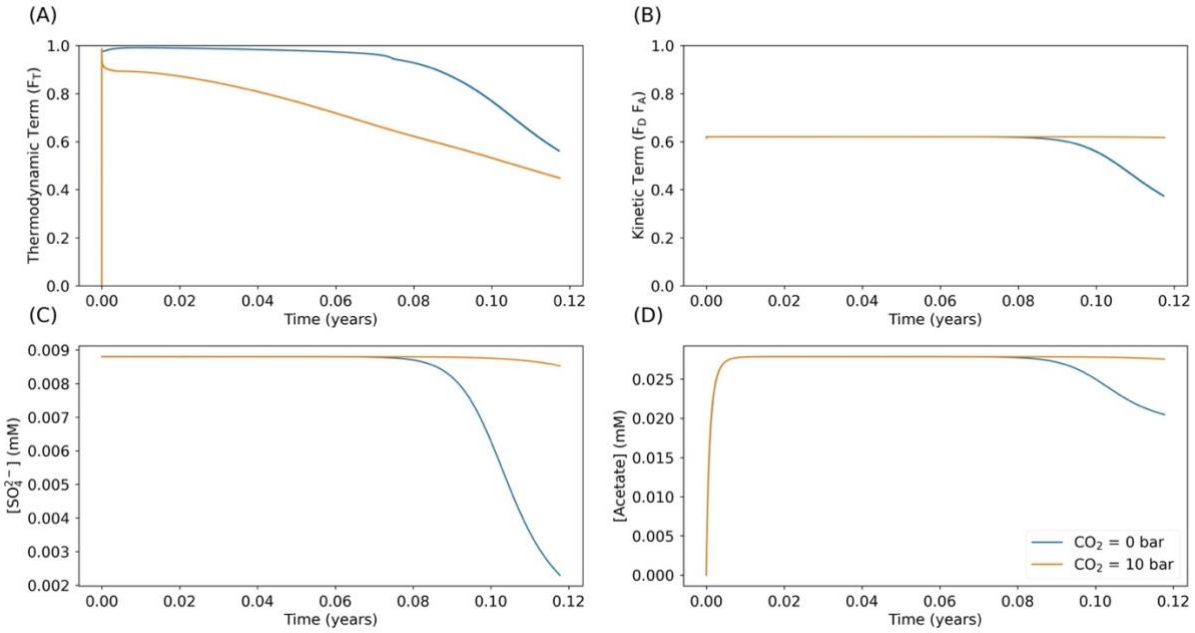


Figure S5: Plots comparing various properties and quantities related to microbial sulfate reduction at the Old Rifle Site over the course of two model runs to demonstrate the effect of CO_2 partial pressure in the injectate. All quantities are taken from the first grid cell of the simulation and plotted over the simulation duration. Blue curve represents the simulation when the injectate is equilibrated with 0 bars of $CO_{2(g)}$. Orange curve represents the simulation when the injectate is equilibrated with 10 bars of $CO_{2(g)}$. (A): the value of the thermodynamic factor in the Monod Biomass rate formulation for MSR (Jin and Bethke, 2003, 2005, 2007). (B): the value of the kinetic term in the Monod Biomass rate formulation for MSR (Jin and Bethke, 2003, 2005, 2007). (C): Sulfate concentrations over time in each case. (D): Acetate concentrations over time in each case. We note the inhibitory effect of increased $CO_{2(g)}$ on F_T and the accompanying slowing of consumption of both sulfate and acetate due to lower rates of MSR.

3 Supplementary Case Study – ODP Site 1086

3.1 Description of Case Study

Site 1086 is located off the coast of South Africa in the southernmost part of the Cape Basin, at 794 meters water depth (Wefer et al., 1998). The sediment comprises carbonate and clay minerals (~80% and ~20% respectively), with other minor constituents (see Pufahl et al., (1998) for a more detailed lithological description). A 200-meter core and associated pore-fluids were obtained during a drilling campaign in 1997 as part of Leg 175 of the Ocean Drilling Program (ODP). Relative to other sites in the Cape Basin, Site 1086 has less organic

carbon in the sediment, and therefore less overall subseafloor microbial activity. Microbial sulfate reduction does occur in this sediment, but sulfate concentrations decrease gradually and are not fully consumed until 180 meters below the seafloor, while nearby sites drilled in the same campaign show sulfate depletion within 20–50 meters of the seafloor (Wefer et al., 1998). Increased alkalinity produced via microbial sulfate reduction often leads to carbonate mineral precipitation within sediments (Meister, 2013), and modelling of the calcium isotopic composition of pore fluid calcium as well as calcium and strontium concentrations at Site 1086 has constrained the depth distribution of this carbonate mineral precipitation (Bradbury and Turchyn, 2018).

We have selected Site 1086 as an example for testing the RTE approach because the sediments and pore-fluids have been studied extensively (Diester-Haass et al., 2004; Weigelt and Uenzelmann-Neben, 2004; Udeze and Oboh-Ikuenobe, 2005; Higgins and Schrag, 2010), and a pre-existing published 1D RTM (Bradbury and Turchyn, 2018) has been used to model the depth distribution of carbonate mineral precipitation and dissolution in the sediment column. For this study, we have reimplemented this 1D RTM in CrunchTope, to feed into the RTE. By applying our emulation approach to this RTM, we explore coupled chemical and physical processes predicted by the RTM and discuss how these processes are encoded within the original parameterisation of the model by Bradbury and Turchyn (2018).

A rough schematic of the RTM is shown in Figure S6, left. We apply the RTE to explore how changes in overlying seawater chemical composition (the upper boundary condition) impact microbial activity and thus carbonate mineral precipitation in the sediment column and consider how sensitive the model is to different input parameters. We also use the model to explore the optimal overlying seawater composition that maximises sedimentary carbonate mineral formation. This is a hypothetical situation as seawater chemistry does not change

rapidly. However, there has been much discussion over how changing ocean chemistry impacts carbonate mineral precipitation in sediments (Kamber and Webb, 2001; Sumner and Grotzinger, 2004; Ridgwell, 2005; Higgins et al., 2009).

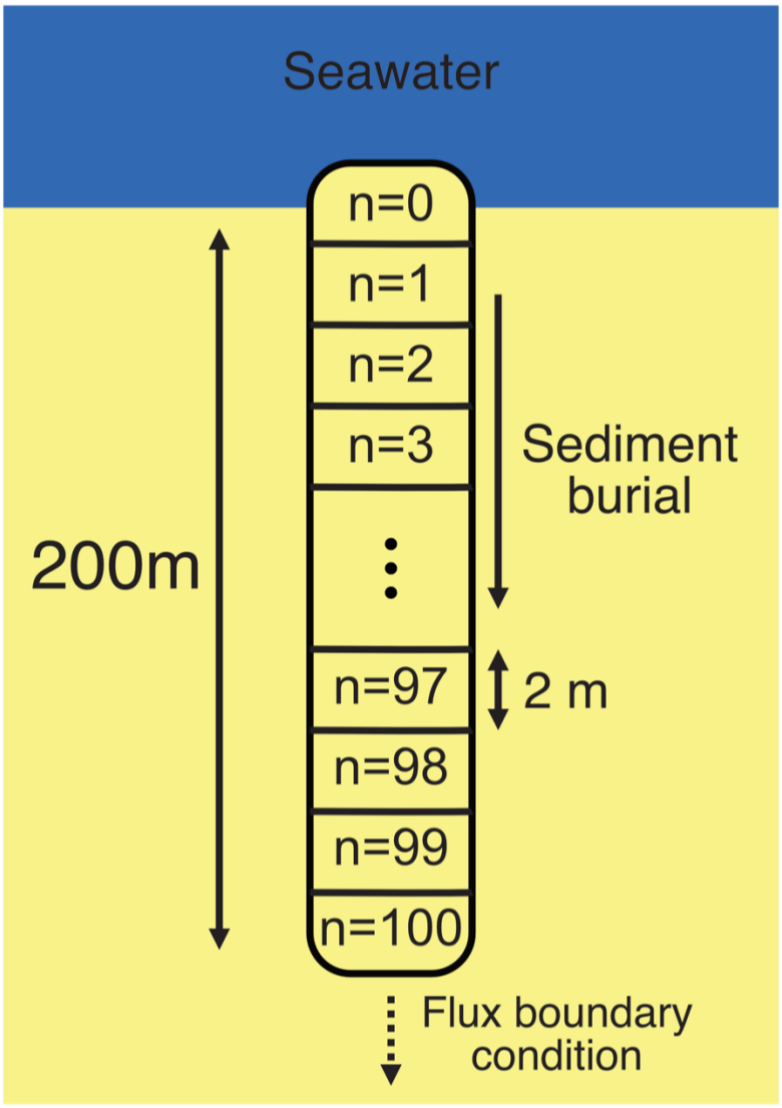


Figure S6: A schematic of the RTM describing the sediment column at ODP Site 1086 after Bradbury and Turchyn (2018).

3.2 Case Study results and interpretation

We begin by applying the emulation methodology to ODP Site 1086. The data for Site 1086 is produced by using Omphalos to create 10803 unique runs with random boundary conditions using concentration within the ranges given in Table S3: .The resulting dataset is

shown in Figure S7. The plots have been colour mapped by formaldehyde concentration (as a proxy for organic carbon available in the sediment column (Meister, 2013)). The colour mapping helps visualise how variability in the volume of calcite precipitated might be considered in conjunction with other model parameters.

Variable	Species	Range
y	Net calcite precipitation	-
x_1	SO_4^{2-}	0-50 mM
x_2	Ca^{2+}	0-50 mM
x_3	Formaldehyde	0-50 mM
x_4	$CO_{2(aq)}$	0-50 mM

Table S3: The attributes and labels (x_i, y respectively) used in the GBT model. Shown also are the ranges over which concentrations are drawn from a uniform distribution to create the boundary conditions that are used in the ODP Site 1086 RTM. The GBT model is trained to predict *y* values (top row) based on a set of input *x* values (next five rows).

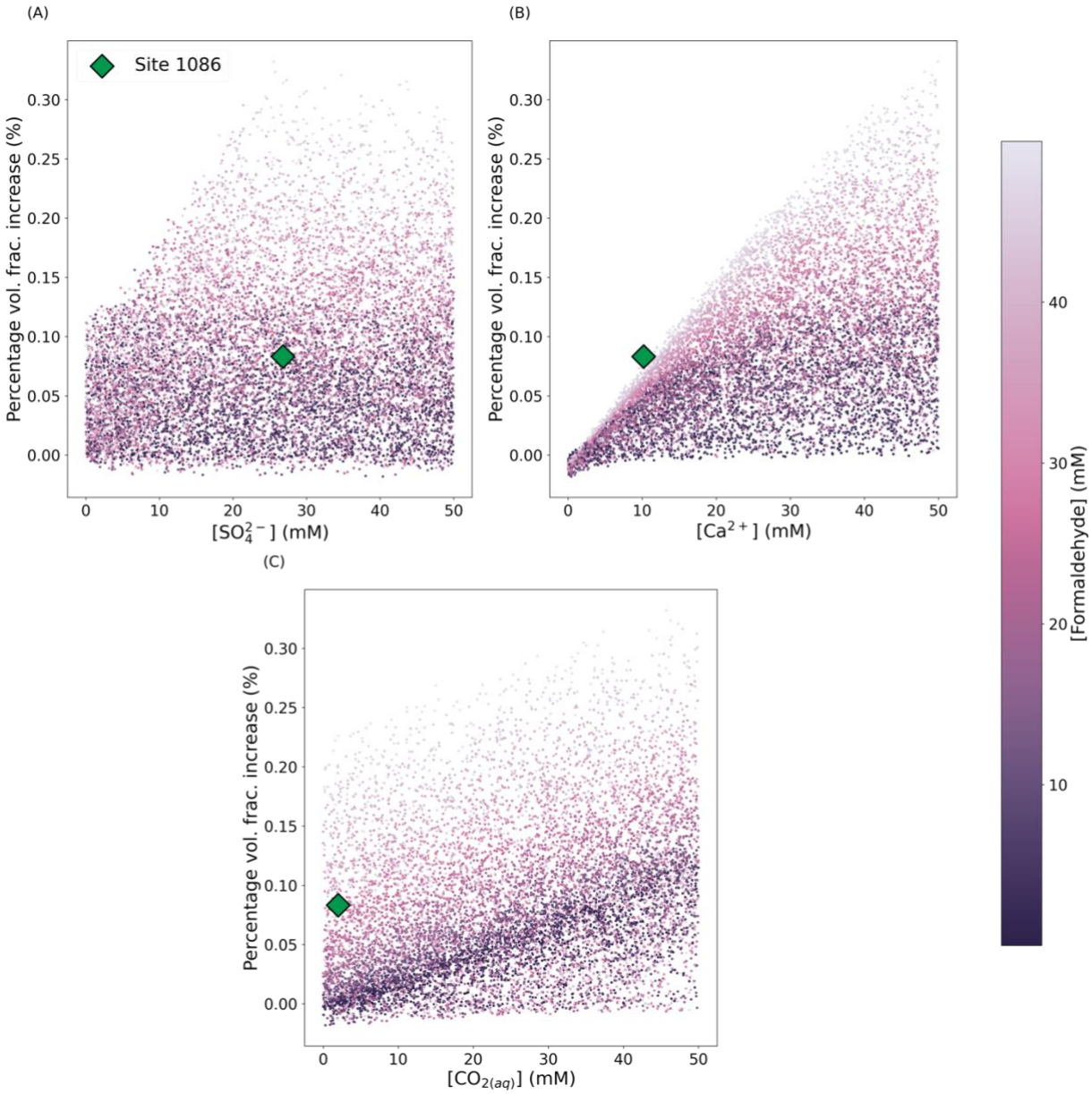


Figure S7: Scatter plots of the chemical concentration of modelled seawater above ODP Site 1086 against the RTM modelled volume of calcite precipitation, colour mapped by formaldehyde. The dataset comprises 10,803 points generated by drawing all four variables independently from uniform distributions, with the corresponding net calcite precipitation calculated by running the RTM described previously (Bradbury and Turchyn, 2018) with the randomised boundary conditions. The green diamond indicates the actual seawater composition above Site 1086. Negative values indicate net dissolution of calcite rather than precipitation.

The results visually suggest that SO_4^{2-}, Ca^{2+} and, $CO_{2(aq)}$ concentrations are all loosely correlated with the volume of calcite precipitated in the sediment. In particular, there are upper and lower bounds for the volume of calcite precipitated, most clearly seen in the plot of

Ca^{2+} against net calcite precipitation but also to a lesser extent in the plots of SO_4^{2-} and $CO_{2(aq)}$. We note that formaldehyde influences the precipitation of calcite with higher formaldehyde concentrations leading to more calcite precipitated; this is particularly visible in Figure S7B and Figure S7C, and to a lesser extent in Figure S7A. These results suggest higher amounts of the electron donor (formaldehyde) lead to higher amounts of calcite precipitation for a given concentration of calcium or carbon dioxide available in seawater. We acknowledge the caveat that the overall amount of calcite precipitated or dissolved will depend on the duration of the model run – this is held constant for all runs in this study. As the model is run for 1,500,000 years, we assume it reaches steady state.

We use the dataset presented in Figure S7 to train the GBT model (holding back 10% of the dataset for testing), and we use this emulator to plot the results shown in Figure S8. Here we show projections of the RTE along one axis accompanied by RTM modelled results. The volume of calcite precipitated is predicted by the GBT model over the range of parameter space (Table S3:), with the other parameters held at the real-world values for Site 1086 (see Bradbury and Turchyn (2018)). It is important to note again that these RTM modelled results (crosses in Figure S8) are not points present in the dataset shown in Figure S7. They are generated from the same initial, RTM describing Site 1086, but the GBT model has not been exposed to them during training. The ability of the RTE to correctly predict these unseen data demonstrates that, as intended, the GBT model is correctly learning the underlying RTM by training on the randomly generated points (Figure S7).

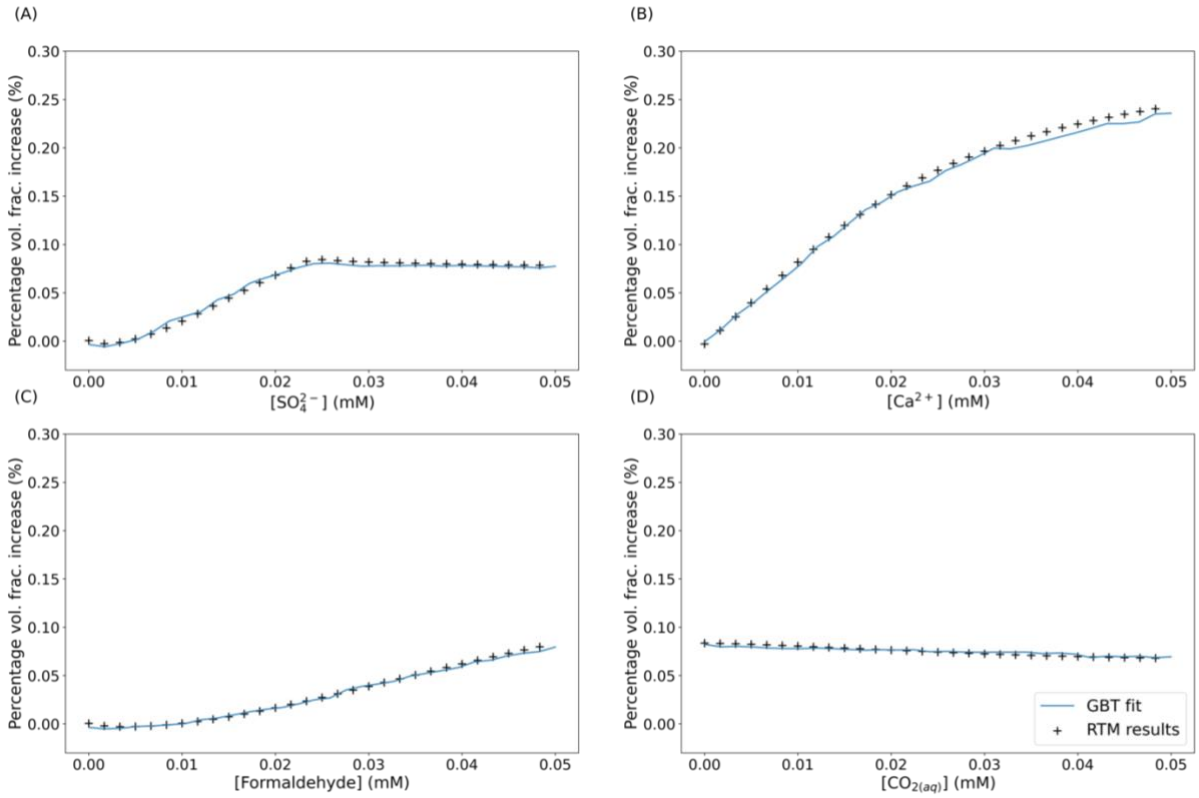


Figure S8: Plots of the GBT model fit compared to the RTM run discretely for the various changes in concentrations of the input (Bradbury and Turchyn, 2018). Each plot shows the percentage increase in volume fraction of calcite in the sediment as each concentration is varied while all other concentrations are held at modern seawater values. The blue line shows the RTE predicted values over this range and the black crosses show the RTM calculated value for the equivalent initial condition. We stress that the comparative points calculated by the RTM were not seen by the emulator during training.

Illustrating the model fit while varying two parameters simultaneously gives the plots shown in Figure S9. The surfaces show how the volume of calcite precipitation co-varies as a function of pairs of variables on which the emulator was trained. Taking the plot of formaldehyde versus Ca^{2+} (Figure S9A) for example, we note that as the concentration of formaldehyde increases, the gradient of the dependence of calcite precipitation on Ca^{2+} increases. This is an example of the complex non-linearities that result from the interplay between the many coupled processes accounted for within RTMs. Returning to the RTM, we can deduce that this is because increasing formaldehyde concentration increases the rate of microbial sulfate reduction in the sediment. This, in turn, increases the amount of dissolved

inorganic carbon (DIC) available for calcite precipitation. The rate law governing calcite precipitation in this model is a transition state theory (TST) formulation as shown in Equation (S1.6.1) (Lasaga and Kirkpatrick, 1981) and the associated chemical reaction for calcite precipitation is shown in Equation (S1.6.2). Both the available DIC and Ca^{2+} enter the equation in the same way: through the ion activity product. Therefore, as formaldehyde concentrations increase, resulting in greater rates of microbial sulfate reduction and hence more DIC in the porewater, the rate of calcite precipitation (Equation (S1.6.1)) for any given value of Ca^{2+} increases.

$$R = A_m k_m \exp\left(\frac{-E_a}{RT}\right)\left(1 - \frac{1}{K_{eq}}\frac{[Ca^{2+}][HCO_3^-]}{[H^+]}\right) \quad \text{(S1.6.1)}$$

$$Ca^{2+} + 2HCO_3^- \rightarrow CaCO_{3(s)} + H_2O + CO_2 \quad \text{(S1.6.2)}$$

There is comparable behaviour in Figure S9B when we consider Ca^{2+} and $CO_{2(aq)}$ although we have proportionally more calcite precipitated overall. This is because, when creating DIC through microbial sulfate reduction, we are also lowering the pH (Soetaert et al., 2007). Calcite precipitation is inhibited at low pH (see Equation (S1.6.1)) due to speciation of DIC to $CO_{2(aq)}$ at low pH. When we increase $CO_{2(aq)}$ in the boundary condition however, it speciates to HCO_3^- since the boundary condition pH is fixed. Therefore increasing DIC in the boundary condition drives more calcite precipitation due to increasing HCO_3^- concentrations.

Figure S9C suggests that the gradient of the amount of calcite precipitated as a function of Ca^{2+} increases with SO_4^{2-}. However, there is a region where, beyond a certain SO_4^{2-}, the amount of calcite precipitated no longer increases. We note this emerges because the equation governing microbial sulfate reduction in this RTM consumes twice as much formaldehyde as sulfate (Equation (S1.6.3)). Therefore, sulfate rapidly becomes in excess at high concentrations.

$$\frac{1}{8}SO_4^{2-} + \frac{1}{4}CH_2O + \frac{1}{4}H^+ \rightarrow \frac{1}{8}H_2S_{(aq)} + \frac{1}{4}CO_{2(aq)} + \frac{1}{4}H_2O \qquad \text{(S1.6.3)}$$

Finally, Figure S9D and Figure S9E show that when $CO_{2(aq)}$ is considered relative to formaldehyde or relative to SO_4^{2-} they have a largely independent effect on the amount of calcite precipitated. In each case, the effect of an increase in any one species appears approximately additive until the reaction is limited by other species. This reflects the fact that $CO_{2(aq)}$ versus SO_4^{2-} and formaldehyde together represent two different ways of adding DIC to the system (direct addition vs. microbial sulfate reduction). When the system achieves calcite supersaturation, these two sources of DIC then compete for the same pool of Ca^{2+} ions to form the solid phase.

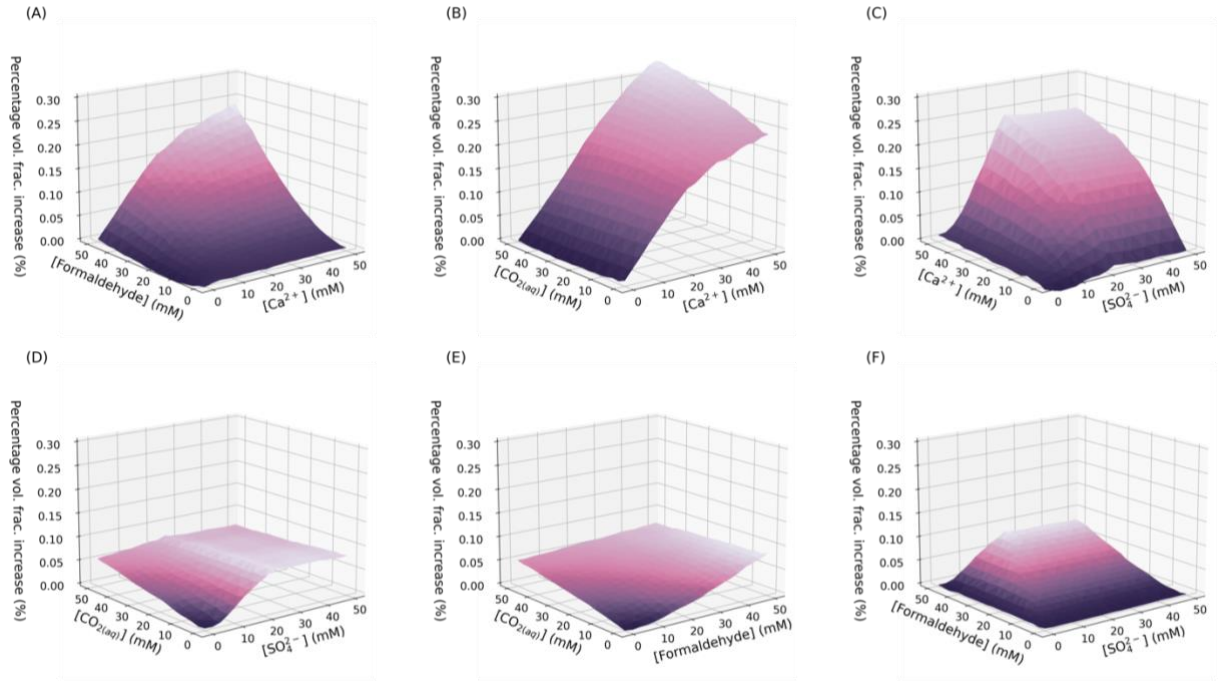


Figure S9: Plots of the emulator net predicted increase in volume fraction of calcite due to precipitation over the modelled period (1.5 Mya). Predictions are plotted in two dimensions for each possible pair of variables.

Now that we have analysed the behaviour of the emulator in the context of the original RTM formulation, we vary all four parameters simultaneously and use the emulator to determine

the maximum volume of calcite precipitated over the range covered by the model's 4D parameter space. By changing the chemistry of the overlying seawater it is theoretically possible to increase calcite precipitation (Table S4: S). While this result is not unexpected (increasing Ca^{2+}, and DIC should lead to more calcite precipitate) the emulator does correctly determine that increasing sulfate beyond 25 mM would be ineffective in increasing the total amount of calcite precipitated. This can be seen by examining the stoichiometry of the reaction for microbial sulfate reduction, shown in Equation (S1.6.3). This is consistent with our previous analysis that SO_4^{2-} is rapidly in excess in sedimentary pore fluids, and that otherwise adding large amount of DIC, either through modifying boundary conditions or via oxidation from microbial metabolism, in conjunction with large concentrations of Ca^{2+} will lead to the most sedimentary carbonate precipitation. The ability of the emulator to identify the point of diminishing returns when modifying the boundary condition offers an important demonstration of this approach's potential for future applications in predicting optimal geochemical amendments.

Quantity	Optimised Value
SO_4^{2-}	25 mM
Ca^{2+}	50 mM
Formaldehyde	50 mM
$CO_{2(aq)}$	45.8 mM
GBT predicted net calcite precipitation	0.00120
RTM modelled calcite precipitation	0.00122
Difference in predicted vol. frac.	1.49×10^{-5}
Percentage error in vol. frac. change	1.2%

Table S4: Seawater fluid composition that would theoretically maximise calcite precipitation at Site 1086 with the associated volume fraction increase predicted by the

emulator. We note that the range of predictable values is constrained by the trained range of the GBT model, so in the cases where the value predicted is equal to that maximum (50mM for Site 1086) the true optimal value could be higher. The penultimate row gives the true net increase in volume fraction calculated by the RTM with these boundary conditions, which we take to be the ground truth.

With respect to the first case study for ODP Site 1086, the emulator suggests a 4-fold increase in the amount of calcite precipitated when compared to the original RTM.

4 Bibliography

Biewald, L.: Experiment tracking with weights and biases, Software available from wandb. com, 2020.

Bradbury, H. J. and Turchyn, A. V.: Calcium isotope fractionation in sedimentary pore fluids from ODP Leg 175: Resolving carbonate recrystallization, Geochimica et Cosmochimica Acta, 236, 121–139, https://doi.org/10.1016/j.gca.2018.01.040, 2018.

Chen, T. and He, T.: xgboost: eXtreme Gradient Boosting, 4, 2015.

Diester-Haass, L., Meyers, P. A., and Bickert, T.: Carbonate crash and biogenic bloom in the late Miocene: Evidence from ODP Sites 1085, 1086, and 1087 in the Cape Basin, southeast Atlantic Ocean, 19, https://doi.org/10.1029/2003PA000933, 2004.

Druhan, J. L., Steefel, C. I., Conrad, M. E., and DePaolo, D. J.: A large column analog experiment of stable isotope variations during reactive transport: I. A comprehensive model of sulfur cycling and δ34S fractionation, Geochimica et Cosmochimica Acta, 124, 366–393, https://doi.org/10.1016/j.gca.2013.08.037, 2014.

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, 29, 1189–1232, 2001.

Friedman, J. H.: Stochastic gradient boosting, Computational Statistics & Data Analysis, 38, 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2, 2002.

Higgins, J. A. and Schrag, D. P.: Constraining magnesium cycling in marine sediments using magnesium isotopes, Geochimica et Cosmochimica Acta, 74, 5039–5053, https://doi.org/10.1016/j.gca.2010.05.019, 2010.

Higgins, J. A., Fischer, W. W., and Schrag, D. P.: Oxygenation of the ocean and sediments: Consequences for the seafloor carbonate factory, Earth and Planetary Science Letters, 284, 25–33, https://doi.org/10.1016/j.epsl.2009.03.039, 2009.

Jin, Q. and Bethke, C. M.: A New Rate Law Describing Microbial Respiration, Appl Environ Microbiol, 69, 2340–2348, https://doi.org/10.1128/AEM.69.4.2340-2348.2003, 2003.

Jin, Q. and Bethke, C. M.: Predicting the rate of microbial respiration in geochemical environments, 69, 1133–1143, 2005.

Jin, Q. and Bethke, C. M.: The thermodynamics and kinetics of microbial metabolism, American Journal of Science, 307, 643–677, https://doi.org/10.2475/04.2007.01, 2007.

Kamber, B. S. and Webb, G. E.: The geochemistry of late Archaean microbial carbonate: implications for ocean chemistry and continental erosion history, Geochimica et Cosmochimica Acta, 65, 2509–2525, https://doi.org/10.1016/S0016-7037(01)00613-5, 2001.

Lasaga, A. C. and Kirkpatrick, R. J.: Kinetics of geochemical processes, 1981.

Meister, P.: Two opposing effects of sulfate reduction on carbonate precipitation in normal marine, hypersaline, and alkaline environments, 41, 499–502, https://doi.org/10.1130/G34185.1, 2013.

Pufahl, P., Maslin, M., Anderson, L., Brüchert, V., Jansen, F., Lin, H., Perez, M., Vidal, L., and Party, S.: 18. Lithostratigraphic summary for Leg 175: Angola-Benguela upwelling system, Initial Reports, 175, https://doi.org/10.2973/odp.proc.ir.175.118.1998, 1998.

Ridgwell, A.: A Mid Mesozoic Revolution in the regulation of ocean chemistry, Marine Geology, 217, 339–357, https://doi.org/10.1016/j.margeo.2004.10.036, 2005.

Schapire, R. E.: The strength of weak learnability, 31, 1990.

Soetaert, K., Hofmann, A. F., Middelburg, J. J., Meysman, F. J. R., and Greenwood, J.: The effect of biogeochemical processes on pH, Marine Chemistry, 105, 30–51, https://doi.org/10.1016/j.marchem.2006.12.012, 2007.

Sumner, D. Y. and Grotzinger, J. P.: Implications for Neoarchaean ocean chemistry from primary carbonate mineralogy of the Campbellrand-Malmani Platform, South Africa, 51, 1273–1299, https://doi.org/10.1111/j.1365-3091.2004.00670.x, 2004.

Udeze, C. U. and Oboh-Ikuenobe, F. E.: Neogene palaeoceanographic and palaeoclimatic events inferred from palynological data: Cape Basin off South Africa, ODP Leg 175, 25, 2005.

Wefer, G., Berger, W. H., Richter, C., and et al. (Eds.): Proceedings of the Ocean Drilling Program 175 Initial Reports, Ocean Drilling Program, https://doi.org/10.2973/odp.proc.ir.175.1998, 1998.

Weigelt, E. and Uenzelmann-Neben, G.: Sediment deposits in the Cape Basin: Indications for shifting ocean currents?, 16, 2004.