Integrating Molecular-Biological Data and Process-Based Models of Nitrogen Cycling

DISSERTATION

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Anna Störiko
aus Kassel

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 27.06.2022
Dekan: Prof. Dr. Thilo Stehle
1. Berichterstatter: Prof. Dr.-Ing. Olaf A. Cirpka
2. Berichterstatter: Prof. Dr. Thilo Streck
3. Berichterstatter: Adjunct Associate Prof. Dr. Martin Thullner

Abstract

Advances in molecular-biological and omics tools have revolutionized microbiology, providing information about the vast abundance and diversity of previously difficult-to-study microorganisms in the environment. However, the obtained information is often qualitative in nature and it remains an open question whether and how molecular-biological data can be quantitatively linked to biogeochemical reaction rates. Mathematical models of biogeochemical processes enable quantitatively predicting reaction rates, but they require data to constrain model predictions. While molecular-biological data have the potential to provide information on microbial dynamics in biogeochemical models, this requires that models simulate the respective biological quantities, such as concentrations of functional genes, transcripts or enzymes. Such modeling approaches have been developed in recent years, introducing additional complexity and thus requiring a critical evaluation of their benefits and associated challenges. In this thesis, I advance the integration of molecular-biological data with biogeochemical modeling, focusing on the nitrogen cycle. Understanding the processes that transform and remove reactive nitrogen species in the environment is important because these compounds cause a number of environmental problems, including drinking water contamination, greenhouse gas emissions, and eutrophication of surface waters.

I address the question whether transcript and enzyme concentrations can directly serve as proxy variables for reaction rates. Further, I inquire whether accounting for the regulation of reaction rates by enzyme concentrations improves the prediction of denitrification rates. To this end, I developed an enzyme-based denitrification model that simulates concentrations of transcription factors, functional-gene transcripts, enzymes, and solutes. I calibrated it using experimental data from a well-controlled batch experiment. The model accurately predicts denitrification rates and the measured transcript dynamics. The relationship between simulated transcript concentrations and reaction rates exhibits strong non-linearity and hysteresis in time. The hysteresis is caused by the faster dynamics of gene transcription and substrate consumption, relative to enzyme production and decay. Hence, assuming a unique relationship between transcript-to-gene ratios and reaction rates, as frequently suggested, may be an erroneous simplification. Comparing model results of the enzyme-based model to those of a classical Monod-type model reveals that both formulations perform equally well with respect to nitrogen species, indicating only a low benefit of integrat-

ing molecular-biological data for estimating denitrification rates. Nonetheless, the enzyme-based model is a valuable tool to improve our mechanistic understanding of the relationship between biomolecular quantities and reaction rates. Furthermore, the results highlight that both enzyme kinetics (i. e., substrate limitation and inhibition) and gene expression or enzyme dynamics are important controls on denitrification rates.

In the second part of my thesis, I extend my analysis of the relationship between transcripts, enzymes and reaction rates to an environmental system by coupling the enzyme-based model to reactive transport at the river–groundwater interface. I evaluate the response of transcripts and enzymes to stable and dynamic hydrogeochemical regimes. While functional-gene transcripts respond to short-term (diurnal) fluctuations of substrate availability and oxygen concentrations, enzyme concentrations are stable over such time scales. The presence of functional-gene transcripts and enzymes globally coincides with the zones of active denitrification. However, transcript and enzyme concentrations do not directly translate into denitrification rates in a quantitative way because of non-linear effects and hysteresis caused by variable substrate availability and oxygen inhibition. I suggest that molecular-biological data should be combined with aqueous chemical data, which can typically be obtained at higher spatial and temporal resolution, in order to parameterize and calibrate reactive-transport models.

The third part of this thesis focuses on integrating functional gene data of several nitrogen cycling processes into a biogeochemical model. Adopting a gene-centric model approach, I simulated nitrogen cycling in hyporheic zone sediments during a flow-through column experiment. In order to quantify the uncertainty of reaction rates and parameters I calibrated the model with a Bayesian approach, using solute data and relative-quantitative data of functional genes from the literature. My results highlight that the functional-gene data reduce the uncertainty of reaction parameters, but not of reaction rates. The overall low uncertainty of reaction rates suggests that solute data alone can strongly constrain nitrogen cycling rates, but it could also be a sign of overconfident estimates due to model structural errors. To overcome convergence problems of the Bayesian sampler that are caused by highly correlated parameters, I introduced a reparametrization for the Monod reaction parameters that facilitates effective sampling from the posterior distribution.

Further research should aim to improve both measurement techniques that quantify microorganisms, and modeling approaches. Models that integrate molecular-biological data call for rigorous uncertainty quantification, and eventually need to be scaled to field applications.

Zusammenfassung

Fortschritte in der Molekularbiologie und die Entwicklung von »omics«-Methoden haben die Mikrobiologie revolutioniert und liefern Informationen über die enorme Fülle und Vielfalt von Mikroorganismen in der Umwelt, die zuvor nur schwer zu untersuchen waren. Die gewonnenen Informationen sind jedoch häufig qualitativer Natur, und es bleibt eine offene Frage, ob und wie molekularbiologische Daten quantitativ mit biogeochemischen Reaktionsraten verknüpft werden können. Mathematische Modelle biogeochemischer Prozesse ermöglichen quantitative Vorhersagen von Reaktionsraten, aber sie benötigen Daten, um die Modellvorhersagen einzuschränken. Zwar können molekularbiologische Daten potentiell Informationen über die mikrobielle Dynamik in biogeochemischen Modellen liefern, doch setzt dies voraus, dass die Modelle die entsprechenden biologischen Größen, wie die Konzentrationen funktioneller Gene, Transkripte oder Enzyme, auch simulieren. In den letzten Jahren wurden entsprechende Modellierungsansätze entwickelt, die jedoch zusätzliche Komplexität mit sich bringen. Daher ist es wichtig, sowohl ihren Nutzen als auch die damit verbundenen Herausforderungen kritisch auszuwerten. In dieser Arbeit entwickle ich die Einbindung molekularbiologischer Daten in biogeochemische Modelle weiter, wobei ich mich auf den Stickstoffkreislauf als Beispiel konzentriere. Reaktive Stickstoffspezies verursachen verschiedene Umweltprobleme, darunter die Verunreinigung von Trinkwasser, Treibhausgasemissionen und die Eutrophierung von Oberflächengewässern. Deshalb ist es wichtig, die mikrobiellen Prozesse, die diese Verbindungen umwandeln und aus der Umwelt entfernen, besser zu verstehen.

Im ersten Teil dieser Arbeit gehe ich der Frage nach, ob Transkript- und Enzymkonzentrationen direkt als Proxy-Variablen für Reaktionsraten dienen können. Außerdem untersuche ich, ob es die Vorhersage von Denitrifikationsraten verbessert, wenn die Regulierung von Reaktionsraten durch Enzymkonzentrationen berücksichtigt wird. Zu diesem Zweck habe ich ein enzymbasiertes Denitrifikationsmodell entwickelt, das die Konzentrationen von Transkriptionsfaktoren, Transkripten funktioneller Gene, Enzymen und gelösten Stoffen simuliert. Anhand von experimentellen Daten aus einem kontrollierten Batch-Experiment habe ich das Modell kalibriert. Das Modell kann die Denitrifikationsraten und die gemessene Transkriptdynamik gut abbilden. Die Beziehung zwischen den simulierten Transkriptkonzentrationen und den Reaktionsgeschwindigkeiten weist eine starke Nichtlinearität und zeitliche Hysterese auf. Die Hysterese ist auf die schnellere Dynamik der Gentranskription und des Substratver-

brauchs im Vergleich zur Produktion und dem Abbau von Enzymen zurückzuführen. Die Annahme einer eindeutigen Beziehung zwischen Transkriptkonzentrationen und Reaktionsraten, wie sie häufig vorgeschlagen wird, kann daher eine falsche Vereinfachung sein. Vergleicht man die Modellergebnisse des enzymbasierten Modells mit denen eines klassischen Monod-Modells, so zeigt sich, dass beide Formulierungen die Dynamik der Stickstoffspezies gleich gut vorhersagen können. Die Integration molekularbiologischer Daten in das Modell bringt für die Schätzung der Denitrifikationsraten also nur einen geringen Nutzen. Nichtsdestotrotz ist das enzymbasierte Modell ein wertvolles Instrument, um das mechanistische Verständnis der Beziehung zwischen molekularbiologischen Größen und Reaktionsraten zu verbessern. Darüber hinaus zeigen die Ergebnisse, dass sowohl die Enzymkinetik (also Substratlimitierung und Inhibierung) als auch Genexpression oder die Dynamik von Enzymkonzentrationen wichtige Einflussfaktoren auf die Denitrifikationsraten sind.

Im zweiten Teil meiner Arbeit erweitere ich meine Analyse der Beziehung zwischen Transkripten, Enzymen und Reaktionsraten auf ein Umweltsystem, indem ich das enzymbasierte Modell mit reaktivem Stofftransport an der Schnittstelle zwischen Grundwasser und Oberflächenwasser verknüpfe. Anhand verschiedener Szenarien untersuche ich, wie Transkript- und Enzymkonzentration auf stabile und dynamische hydrogeochemische Bedingungen reagieren. Während funktionelle Gentranskripte auf kurzfristige (tageszeitliche) Schwankungen der Substratverfügbarkeit und der Sauerstoffkonzentration reagieren, bleiben die Enzymkonzentrationen über diese Zeitskalen stabil. Qualitativ stimmen die Bereiche, in denen Transkripte funktioneller Gene und Enzyme vorhanden sind, mit den Zonen aktiver Denitrifikation überein. Die Transkript- und Enzymkonzentrationen lassen sich jedoch aufgrund nichtlinearer Effekte und Hysterese, die durch variable Substratverfügbarkeit und Sauerstoffhemmung verursacht werden, nicht direkt in quantitative Denitrifikationsraten umrechnen. Ich schlage vor, dass molekularbiologische Daten mit chemischen Daten gelöster Stoffe kombiniert werden sollten, die in der Regel mit höherer räumlicher und zeitlicher Auflösung gewonnen werden können, um reaktive Transportmodelle zu parametrisieren und zu kalibrieren.

Der dritte Teil dieser Arbeit konzentriert sich auf die Integration funktioneller Gendaten verschiedener Stickstoffkreislaufprozesse in ein biogeochemisches Modell. Unter Verwendung eines gen-zentrierten Modellansatzes habe ich den Stickstoffkreislaufreaktionen in Sedimenten aus der hyporheischen Zone während eines Durchfluss-Säulenexperiments simuliert. Um die Unsicherheit von Reaktionsraten und Parametern zu quantifizieren, habe ich das Modell mit einem bayesschen Ansatz mit Daten gelöster Stoffe und relativ-quantitativen Daten funktioneller Gene aus der Literatur kalibriert. Meine Ergebnisse zeigen, dass die funktionellen Gendaten die Unsicherheit der Reaktionsparameter verringern, nicht aber die der Reaktionsraten. Die insgesamt geringe Unsicherheit der Reaktionsraten deutet darauf hin, dass die

Daten gelöster Stoffen ausreichen, um die Stickstoffkreislaufraten stark einzuschränken. Das könnte jedoch auch ein Zeichen für eine Unterschätzung der Unsicherheit sein, die durch strukturelle Modellfehler verursacht wird. Starke Korrelationen zwischen Modellparametern verursachen Konvergenzprobleme des bayesschen Samplers. Eine Reparametrisierung der Monod-Reaktionsparameter hat es jedoch ermöglicht, dennoch effektiv eine Stichprobe aus der A-posteriori-Verteilung zu ziehen.

Weitere Forschung sollte anstreben, sowohl die Messverfahren zur Quantifizierung von Mikroorganismen als auch Modellierungsansätze zu verbessern. Modelle, die molekularbiologische Daten integrieren, erfordern eine rigorose Quantifizierung der Unsicherheit und müssen schlussendlich für Anwendungen im Feld skaliert werden.

Acknowledgements

Writing this dissertation was a great challenge for me. Fortunately, I was accompanied and supported by many people, to whom I would like to express my sincere gratitude.

First of all, I would like to thank my supervisor Olaf Cirpka who was a great mentor and teacher throughout my studies and PhD. You have sparked my enthusiasm for environmental modeling during the first year of my studies in your System's Analysis lecture. Looking back, my first modeling project in this course on analyzing the growth dynamics of bacteria in lab reactors seems to have pointed the way ahead for me.

I would like to extend my sincere thanks to my supervisor Holger Pagel, who has always taken the time to discuss my work and who motivated me to go on when I was frustrated.

I am also grateful to my supervisor Philippe Van Cappellen for inspiring discussions of my work, and for hosting me at the University of Waterloo for half a year.

Special thanks go to Adrian Mellage for being an unofficial forth member of my PhD committee. You are a great role model for me. I particularly appreciate your positive attitude and excellent way of providing feedback.

I am extremely grateful to my fellow PhD students in the research training group *Integrated Hydrosystems Modeling*. Your support and the feeling of being in this together have carried me through the difficult times of my PhD.

I would also like to thank the Ecohydrology group at the University of Waterloo, especially Christina Smeaton, Bingjie Shi, Stephanie Slowinski, Stephane Ngueleu, and Marianne Vandergriendt, who helped me to conduct my first real lab experiment. Even though the experimental results have not made it into this thesis, I have learned a lot from you.

I thank Linda Bergaust and Jianqiu Zheng for providing experimental data for the models in chapters 3 and 5, respectively. Thanks to the Deutsche Forschungsgemeinschaft for providing the funding of my PhD project.

I'd also like to acknowledge the many, many developers of all the great open source software that I have used for my work. The list of tools is long, and I can only name a few examples here that have been especially helpful for me: the Python packages xarray, PyMC, and Matplotlib; the Zotero reference manager; LaTeX in general, and the KOMA-script bundle and the biblatex package in particular.

Several people have provided feedback on parts of my thesis: thank you Luciana, Jonas, Cora, Theresa, and Lisa.

Last but not least I would like to thank my family and friends, whose support was essential for completing this project. I am deeply grateful to Adrian Seyboldt – for discussing my science over lunch or dinner and reading my early drafts, for teaching me a lot about Python and Bayesian statistics, for tolerating a long-distance relationship during my research stay in Waterloo as well as my frustration and long working hours in the past months, and for all your love and support.

Contents

List of Figures

List of Tables

Statement of Contributions

This thesis is based on a set of (published or yet unpublished) co-authored manuscripts. I was primarily responsible for the model design and analysis, and I am the first author on each of the papers.

Chapters 1 and 2 reproduce shorter sections from papers A and B. Chapter 3 is based on paper A, with a shortened introduction. Chapter 4 is adapted from paper B. A revised version of the paper has been published after the submission of this thesis. The contributions of all co-authors to the papers are outlined in the tables below.

Co-author contributions in % for paper A: "Does It Pay off to Explicitly Link Functional Gene Expression to Denitrification Rates in Reaction Models?".

	Author	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
1	Anna Störiko	65	100	70	70
2	Holger Pagel	10	0	10	10
3	Adrian Mellage	10	0	10	10
4	Olaf A. Cirpka	15	0	10	10
STATUS IN PUBLICATION PROCESS		Published (Störiko et al., 2021a).			

Co-author contributions in % for paper B: "Denitrification-Driven Transcription and Enzyme Production at the River-Groundwater Interface: Insights from Reactive-Transport Modeling".

	Author	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
1	Anna Störiko	60	100	70	70
2	Holger Pagel	10	0	9	9
3	Adrian Mellage	10	0	8	8
4	Philippe Van Cappellen	10	0	5	5
5	Olaf A. Cirpka	10	0	8	8
STATUS IN PUBLICATION PROCESS		Published (Störiko et al., 2022).			

1 Introduction

1.1 MICROBIAL NITROGEN CYCLING

This chapter contains sections from co-authored manuscripts as indicated in the Statement of Contributions on page xvii.

The increase of diffuse nitrogen inputs, mainly by agriculture, has led to elevated concentrations of reactive nitrogen species in groundwater and surface-water bodies, threatening drinking-water production, and causing eutrophication of rivers and lakes (Erisman et al., 2013). Microorganisms use reactive nitrogen compounds as substrates for redox reactions that fuel their energy metabolism, constituting the main attenuation process for nitrogen contamination in environmental systems (Kuypers et al., 2018). Understanding the factors that enhance microbial removal of reactive nitrogen species from the environment is therefore critical for contamination control and mitigation.

The most important reaction processes for nitrogen in the environment are (1) fixation of dinitrogen gas, (2) conversion of organic nitrogen into ammonium (*ammonification*), (3) oxidation of ammonium to nitrate (*nitrification*), (4) oxidation of ammonium with nitrite to N_2 (anaerobic ammonium oxidation, *anammox*), (5) dissimilatory nitrate reduction to ammonium (DNRA), and (6) the reduction of nitrate to dinitrogen gas (*denitrification*). Many of these processes consist of several reaction steps which can be linked to specific enzymes and their *functional genes*, that is, the genes responsible for coding an enzyme's production (see figure 1.1). Some microorganisms feature all the functional genes related to a certain pathway. Other organisms can only mediate a part of the full pathway and rely on the co-existence with other microorganisms that are capable of the complementary reactions.

Denitrification is the key reaction for the permanent removal of nitrogen species from the environment because it converts the reactive nitrogen species nitrate into inert N_2 gas rather than into another reactive nitrogen species. Even though microorganisms with the potential to readily reduce nitrate are widespread in the environment, nitrate persists in numerous aquifers around the world (Burri et al., 2019; Gutiérrez et al., 2018; Spalding and Exner, 1993). Investigating the factors that stimulate or inhibit the transformation of reactive nitrogen species therefore remains a relevant field of research.

Biogeochemical models have been developed as a tool to quantitatively describe and predict nitrogen cycling processes. As nitrogen cycling reactions are predominantly driven by microorganisms, it is important to capture the microbial dynamics

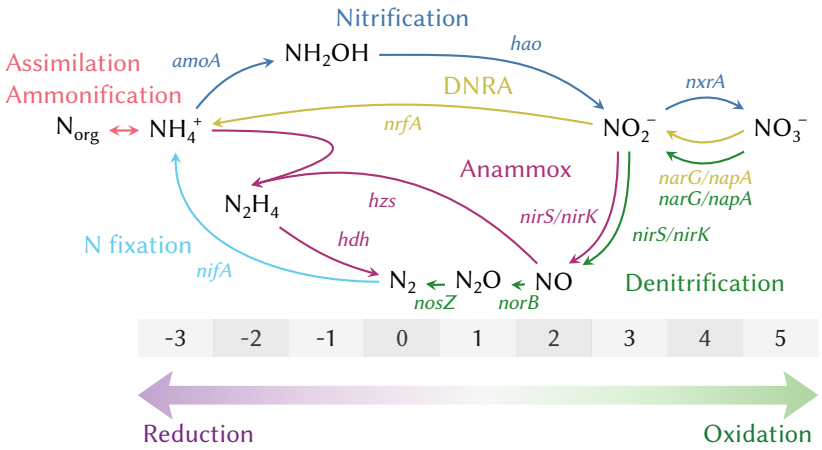


FIGURE 1.1: Nitrogen cycling reactions and the corresponding functional genes. The number line below the reaction scheme indicates the oxidation state of nitrogen compounds. Reaction steps and genes assigned to the same process have the same color. However, reaction steps of one process can be split up to several organisms. Figure adapted from Hallin et al. (2018) and Kuypers et al. (2018).

that regulate reaction rates. However, quantifying microbial biomass in samples is notoriously difficult with "traditional" methods such as adenosine triphosphate (ATP) analyses, chloroform fumigation extraction or plate counts (Gonzalez-Quiñones et al., 2011). Measurements are imprecise and cannot differentiate between microorganisms with different functions. Additionally, only a fraction of microbes can be cultured in the lab such that our knowledge is incomplete and biased towards a few model organisms (Stewart, 2012). Molecular-biological data that can now be readily applied to environmental samples have the potential of overcoming these limitations because they allow for targeting specific functional genes or organisms. This trend also calls for further development of biogeochemical modeling approaches as they need to catch up with the increasing availability of detailed information on microbial dynamics.

1.2 Integration of Molecular-Biological Data into Biogeochemical Models

Modern high-throughput molecular-biological techniques and omics methods provide insights into the abundance, activity, and metabolic function of microorganisms in environmental systems (Bouchez et al., 2016; Starke et al., 2019). Numerous laboratory and field studies (Anantharaman et al., 2016; Bælum et al., 2008; Bowen et al., 2014; Wegner et al., 2019) indicate that quantifying the abundance of genes and gene

transcripts has the potential to improve the description of microbially mediated reactions implemented in reactive-transport models. Many different approaches exist that we can distinguish based on the type of information that they provide:

1 Methods that target the DNA of microorganisms provide information about their *genetic potential*. However, the presence of DNA in the environment does not yet indicate that the related metabolic functions are active; for some environments it has been shown that large fractions of the DNA belong to inactive or dead microorganisms (Carini et al., 2016; Lennon et al., 2018). In contrast, methods that target the RNA or proteins present in the environment, provide insight about *active* microorganisms.

2 While some methods interrogate the *taxonomy* of microorganisms ("What species are present?") other methods can identify the metabolic *function* of microorganisms ("What reactions can they carry out?").

3 Methods such as quantitative PCR (qPCR) provide information about the *abundance* of microorganisms ("How many are there?") whereas other methods (e. g., amplicon sequencing) can be used to investigate their *diversity*.

4 Finally, we can distinguish methods that target *specific marker genes* (e. g., taxonomic or functional markers) from *global sequencing*, used to identify all the genes of all microbes present in a sample.

Transcript-to-gene ratios and relative gene expression levels have been suggested as a direct measure of microbial activity (Brow et al., 2013; Freitag and Prosser, 2009; Monard et al., 2013; Nicolaisen et al., 2008; Rohe et al., 2020). However, assuming a unique (e. g., linear) relationship neglects important factors that modulate reaction rates such as post-transcriptional and post-translational regulation, different time-scales of transcript and enzyme dynamics, and substrate limitation (Moran et al., 2013). In a meta-analysis of experimental studies, Rocca et al. (2015) found that transcript abundances were not significantly correlated with biogeochemical processes. They conclude that it is essential to investigate the factors controlling gene abundance, transcription, translation and enzyme activity to better interpret the patterns of gene and transcript abundances in the environment with respect to ecosystem processes. Therefore, integrating a mechanistic description of the production and decay of transcripts and enzymes into models could help to better establish a mechanistic link between transcript concentrations and reaction rates. Most reactive-transport models do not explicitly account for genes or transcription and thus cannot integrate molecular-biological data sets into quantitative validation frameworks. Addressing this challenge and integrating omics data into reactive-

transport modeling has been proposed as a way to improve our understanding of the dynamic behavior of biogeochemical systems (L. Li et al., 2017).

In a first approach to integrate genetic information, *genome-scale metabolic models* have been coupled to reactive transport, successfully predicting reaction rates under variable environmental conditions without the need for calibration (Scheibe et al., 2009). Though computationally intensive, genome-scale metabolic networks provide a powerful approach for cases in which reactions are dominated by a single, well-known organism or a small group of organisms, as in the case of uranium reduction. Other studies have focused on a single metabolic pathway of an organism, simulating the regulatory chain, including transcription factors, messenger RNA (mRNA), enzymes, substrate consumption, and growth kinetics (Bælum et al., 2013; Koutinas et al., 2011).

Natural biogeochemical cycles, such as the nitrogen cycle, are mediated by a diverse group of organisms, of which only a small percentage can be cultivated. Under these circumstances, genome-scale models and models focusing on the functionality of a single microbial strain become impractical. Instead, approaches that target metabolic pathways rather than processes at the scale of single cells may capture the explicit contribution of the non-cultivatable majority within a natural microbial community.

Thus, the alternative *gene-centric* approach focuses on the functionality represented by certain genes instead of specific organisms. By using functional-gene abundances as a proxy for biomass with a certain function, gene-centric approaches can easily be integrated into existing biogeochemical and reactive-transport models (Reed et al., 2014, 2015). Genomic data provide information on metabolic potential, but not on actual microbial activity. To account for activity, the gene-centric approach has been further developed to predict concentrations of functional transcripts and enzymes (M. Li et al., 2017b; Louca et al., 2016; Song et al., 2017). In one of the studies, the model was validated using metatranscriptomic and metaproteomic data (Louca et al., 2016). However, because the approach of these authors computes transcript and enzyme concentrations by postprocessing of reactions rates, it only provides a partially mechanistic link between rates, transcription, and enzyme production. Hence, this approach lacks the explicit integration of transcript and enzyme regulatory feedbacks into the biogeochemical model.

The *enzyme-based* approach, by contrast, simulates concentrations of specific enzymes that directly regulate reaction rates. Regulation of enzyme production can be based on energetic considerations (M. Li et al., 2017b) or a cybernetic approach (Song and C. Liu, 2015). In the latter, reaction networks are optimized with respect to a metabolic goal such as maximizing the microbial growth rate (Ramkrishna and Song, 2019). None of these modeling approaches represent the actual mechanisms of transcriptional regulation with transcription factors, having the advantage that they do not require special knowledge about the regulatory system. However, the validation of the enzyme-based modeling approach relies on quantitative enzyme

data which remains challenging to obtain, particularly for environmental samples. Conversely, the quantification of mRNA via reverse-transcription quantitative PCR (RT-qPCR) is well-established for both laboratory and field setups. Therefore, integrating transcript data into biogeochemical models is an approach that should be explored further.

1.3 Biogeochemical modeling and uncertainty quantification

Process-based reactive transport models simulate the transport and transformation of chemicals in the environment (Meile and Scheibe, 2019; Siade et al., 2021; Steefel et al., 2005). These models solve partial differential equations (PDEs) that describe the evolution of concentration distributions through reactions and transport by water flow (advection, dispersion) and diffusion. The level of system understanding encoded in model formulations varies between *mechanistic* models that represent physical, chemical and biological knowledge, and more empirical or data-driven approaches. In the past, many biogeochemical reactive transport models have employed simple rate laws such as first order or Monod kinetics that do not explicitly represent microorganisms (e. g., Y. Wang and Van Cappellen, 1996). More recent modeling approaches account for the dynamics of microbial populations and the regulation of the microbial metabolism (Meile and Scheibe, 2019). Biogeochemical models can be used to calculate spatially and temporally resolved reaction rates, analyze the effect of parameters and boundary conditions on the outcome, and test our conceptual understanding of how a system functions. This makes them a valuable tool both for addressing scientific problems as well as for guiding management decisions.

However, the uncertainty of model predictions needs to be taken into account. Several sources of uncertainty contribute to the overall uncertainty of model predictions. Many model parameters cannot be observed directly, and their exact values are unknown (*parameter uncertainty*). This holds particularly true for effective parameters that cannot be easily linked to an actual environmental process. The initial and boundary conditions are also often poorly determined. Data is subject to errors and noise (*measurement uncertainty*). Moreover, models of environmental processes are always a simplified representation of the reality and it is usually not clear which out of several competing model formulations is the "best" one to use (*conceptual* or *model uncertainty*). Forward uncertainty analysis methods such as Monte Carlo simulation can be used to evaluate how uncertainty in the model inputs propagates to model outputs (Linde et al., 2017). More formal sensitivity analysis methods (e. g., parameter screening, variance based sensitivity analysis) allow to identify influential parameters, and to apportion output uncertainty to different sources of uncertainty in the input (Saltelli et al., 2007).

However, the input parameter space is usually large if it is based on prior knowledge only, leading to high output uncertainty. For reliable and useful predictions, the parameter space needs to be restricted based on observations, that is, the model needs to be *calibrated*. In many modeling studies of microbial nitrogen cycling, a single set of hand-tuned parameter values is chosen, or deterministic least-squares optimization is applied to this end (e. g., Knights et al., 2017; M. Li et al., 2017b; Mellage et al., 2018; Reed et al., 2014). However, these methods can provide a linearized estimate of posterior uncertainty at best, which is often unsuitable for the highly non-linear rate laws of biogeochemical models (Siade et al., 2021). Bayesian inference, in contrast, allows to coherently quantify uncertainty while at the same time constraining the spaces of parameters and simulation outcomes with observations. Improved algorithms, more user-friendly software and the increase of computational resources have lead to the widespread use of Bayesian techniques in many disciplines of science (Rode et al., 2010; van de Schoot et al., 2021). Nevertheless, its application to real world environmental models remains challenging, and particularly biogeochemical models often do not exploit the potential of rigorous Bayesian parameter estimation and uncertainty quantification (Siade et al., 2021; G. Wang and Chen, 2012). Some examples of studies adopting a Bayesian approach for models of nitrogen cycling include simulations of nutrient cycling in lakes (Dietzel and Reichert, 2014; Wu et al., 2017), soil greenhouse gas emission (G. Wang and Chen, 2012, 2013; Ying et al., 2017) or nitrification in groundwater (Brunetti et al., 2020).

In the past years, innovations in microbiological and molecular-biological techniques have greatly expanded our knowledge about microorganisms and how they regulate environmental processes. Based on these insights, more detailed models have been developed that can more realistically describe the actual biological functioning of systems (Meile and Scheibe, 2019). However, more complex models are not necessarily better and a balance between model complexity and goodness of fit needs to be found in order to optimize the predictive power of a model (Höge et al., 2018). When multiple models can be formulated to explain the same set of data, science has long employed the principle of *Occam's Razor*, implied in Bayesian theory (MacKay, 2003). This principle advises choosing the simplest hypothesis or model that can explain the data. An important question, therefore, is if new data types can justify a higher level of detail and complexity for biogeochemical modeling. While more complex models allow to employ more mechanistic process descriptions, these formulations also introduce new parameters that are potentially difficult to identify. We thus need to test whether molecular-biological data can indeed improve predictions and reduce parametric uncertainty.

1.4 AIM AND OBJECTIVES

The increasing availability of molecular-biological and omics data from the environment calls for the development of reactive transport models that can make appropriate use of these data. In parallel, we need to critically assess the benefits and challenges that go along with more complex modeling approaches. The aim of my thesis is to develop and evaluate reactive transport models of microbial nitrogen cycling that integrate functional gene and transcript data. Nitrogen cycling is a suitable test application for such models because (1) genes and enzymes of nitrogen cycling reactions are well known (Simon and Klotz, 2013), (2) transcription of genes involved in nitrogen cycling reactions are known to be regulated by environmental conditions (Gaimster et al., 2018), so explicitly including enzymes into models might improve the process description, and (3) quantitatively understanding nitrogen cycling is relevant to a range of environmental problems such as drinking water contamination, eutrophication, and greenhouse gas emissions.

The specific objectives of this thesis are as follows:

1 Develop model formulations for microbial nitrogen cycling that can use molecular-biological data, specifically quantitative and semi-quantitative data of functional genes and functional-gene transcripts.

2 Evaluate the performance of more complex transcript- and enzyme-based model formulations compared to simpler biomass-based model formulations.

3 Evaluate the suitability of functional-gene transcript and enzyme concentrations as a proxy to predict nitrogen cycling rates.

4 Quantify the uncertainty of reactions rates of microbial N-cycling and the associated parameter uncertainty, and assess whether transcript or functional-gene data can reduce this uncertainty.

5 Identify sampling strategies for transcript and gene measurements that lead to a better quantification of reaction rates.

1.5 THESIS STRUCTURE

Chapter 2 summarizes the theory and methods applied in the following research chapters 3 to 5.

In chapter 3, I develop a model formulation that explicitly accounts for the dynamics of transcripts and enzymes related to denitrification, and their interaction with reaction kinetics. I apply the model to a well controlled batch experiment and compare it to a simpler biomass-based model formulation. The uncertainty of reaction

parameters and model outputs is assessed with Bayesian inference. Based on the simulations, I explore the relationship between transcript and enzyme concentrations, and denitrification rates.

Building on the results of chapter 3, I couple the previously developed model to solute transport at the river–groundwater interface in chapter 4, and evaluate the response of transcripts and enzymes for denitrification to stable and dynamic hydrogeochemical regimes. Based on the simulations, I further explore the relationship between enzyme concentrations and denitrification rates over space and time, and provide guidance on sampling strategies for transcript and enzyme concentrations at the river–groundwater interface.

In chapter 5, I combine gene-centric modeling of several nitrogen cycling pathways with Bayesian uncertainty quantification to evaluate the uncertainty of nitrogen cycling processes in hyporheic zone sediments.

The concluding chapter 6 summarizes the major findings of this thesis, highlighting how they advance research on biogeochemical modeling, and outlines future research directions.

2 Theory & Methods

2.1 Biogeochemical modeling of microbial reactions

2.1.1 Biomass-explicit models

Microorganisms catalyze a variety of redox reactions in the environment, transferring electrons from electron donors to substrates that serve as terminal electron acceptors, thus, generating energy for their growth and maintenance. The rates of biogeochemical reactions depend on the abundance and activity of microorganisms. In turn, microbial growth rates depend on the reactions taking place. Biomass-explicit reaction models account for this interplay between microbial growth and biogeochemical reactions by explicitly simulating microbial biomass and its influence on reaction rates.

The growth rate of microbial biomass is described by a rate law of first order with respect to the biomass concentration B [cells L^{-1}]:

$$r_{\text{growth}} = \mu B\,. \tag{2.1}$$

The specific growth rate μ [s^{-1}] is usually not constant because microbial growth depends on the concentration of limiting substrate c [mol dm^{-3}]. Monod (1949) introduced an empirical rate law to describe this dependency with a hyperbolic equation:

$$\mu = \mu_{\max}\frac{c}{c + K}\,, \tag{2.2}$$

where the parameter K [mol dm^{-3}] is called the half-saturation constant, and $\mu_{\max}$ [s^{-1}] is a maximum specific growth rate. The model has been extended to include multiple limiting substrates and inhibitors by considering the product of several terms. The specific growth rate is then given by

$$\begin{aligned}\mu &= \mu_{\max}\prod_{s\,\in\,\text{substrates}}\frac{c_s}{c_s + K_s}\prod_{k\,\in\,\text{inhibitors}}\frac{I_k}{I_k + c_k} \\ &= \mu_{\max}\, f_{\text{substrate}}\, f_{\text{inhibition}}\,.\end{aligned} \tag{2.3}$$

Here, K_s [mol dm^{-3}] and I_k [mol dm^{-3}] are the half-saturation constant of substrate s and the inhibition constant of inhibitor k, respectively.

To fuel their growth, microorganisms need energy that they obtain from the *catabolic* reactions, for example the oxidation of a carbon substrate with oxygen. Additionally, building up biomass also requires chemical building blocks such as a carbon substrate and electron donors that are transformed into more complex organic molecules in the *anabolic* reaction. Microbial growth therefore changes the concentrations of substrates according to the stoichiometry of the overall metabolic reaction. This is expressed by relating the substrate consumption rate proportionally to the growth rate with the growth yield Y [cell mol^{-1}] as proportionality factor. The rate of substrate consumption through the catabolic and anabolic reaction is

$$\begin{aligned} r_s &= -\frac{r_{\text{growth}}}{Y} \\ &= -\frac{\mu}{Y} B \\ &= -\frac{\mu_{\max}}{Y} f_{\text{substrate}}\, f_{\text{inhibition}}\, B\,. \end{aligned} \tag{2.4}$$

Instead of using a maximum specific growth rate $\mu_{\max}$ as parameter, we can also express the substrate consumption rate in terms of a maximum cell-specific consumption rate $\nu_{\max}$ [mol cell^{-1} s^{-1}] where $\mu_{\max} = \nu_{\max} Y$. The growth yield depends on the energy gained through the microbial reaction, which may vary over time because it is, in principal, influenced by environmental conditions. Theoretical growth yields can be derived with different bioenergetic approaches (e. g., Roden and Jin, 2011; Smeaton and Van Cappellen, 2018), but often the growth yield is determined experimentally. If we assume Y to be constant, both parameterizations for the growth rate are equivalent. Otherwise, we can either treat $\nu_{\max}$ as a constant or $\mu_{\max}$. The former is referred to as *Michaelis-Menten* formulation because the rate law then takes the form of the equation introduced by Michaelis and Menten (2011) for enzyme kinetics; the latter corresponds to the *Monod* formulation (Thullner and Regnier, 2019).

When a compound is produced or consumed in M reactions, the net reaction rate for compound i is given by

$$r_i^{\text{net}} = \sum_{j=1}^{M} \frac{\gamma_i}{\gamma_s} r_s^j\,, \tag{2.5}$$

where r_s^j is the substrate consumption rate in reaction j, and γ_i, γ_s are the stoichiometric coefficients of compound i and substrate s. In this notation, educts have a negative stoichiometric coefficient and reaction products have a positive coefficient.

2.1.2 Gene-explicit models

Functional genes are widely used as a proxy variable for functional biomass (e. g., Bouchez et al., 2016; Philippot, 2006). Reed et al. (2014) introduced a modeling

concept in which the concentrations of functional genes serve as the state variables. In comparison to looking at different organisms with potentially redundant function it has the advantage that it limits the number of variables to the number of reactions or functional genes associated to them. In this framework, the rate of gene production for gene j due to the reaction associated with it is analogous to equations (2.1) and (2.3):

$$r_{\Gamma_j} = \mu^j_{\max} \Gamma_j F_T \prod_{s \in \text{substrates}} \frac{c_s}{c_s + K_s} \prod_{k \in \text{inhibitors}} \frac{I_k}{I_k + c_k}, \tag{2.6}$$

where Γ_j is the concentration of gene j. Reed et al. (2014) also extend the rate law by the factor F_T to account for thermodynamic limitation (Jin and Bethke, 2005) in the studied energy-limited submarine environment.

Many microorganisms carry functional genes for several reaction steps. Since microbial growth always duplicates the entire genome, the growth of one gene always leads to an increase in all the genes that co-occur with it, even if the associated reactions are not active. Reed et al. (2014) account for this *metabolic plasticity* with a matrix of co-occurence probabilities $\boldsymbol{\sigma}$. The total production rate of gene i is then given by

$$r^{\text{tot}}_{\Gamma_j} = \sum_{k=1}^{M} \left(\frac{g_j}{g_k} \sigma_{jk} r_{\Gamma_k} \right), \tag{2.7}$$

in which g_j and g_k are the average copy numbers of gene j and k within a genome. Note that $\boldsymbol{\sigma}$ is not a symmetric matrix, i. e. the probability that an organism has gene j given that gene k is present (σ_{jk}) is not necessarily the same as the probability that an organism has gene k when it possesses gene j (σ_{kj}). For example, the *rpoB* gene coding for a subunit of bacterial RNA polymerase is so essential for the microbial survival that it is present in nearly all bacteria and, hence, also in denitrifiers carrying the *nirS* gene. That is, $\sigma_{rpoB,nirS} \approx 1$. Conversely, denitrifiers usually represent only a fraction of all bacteria and therefore the probability that *nirS* is present given that an organism carries the *rpoB* gene ($\sigma_{nirS,rpoB}$) is much smaller than 1.

Even though the gene-centric modeling approach introduced a concept to model metabolic plasticity, it has not been applied in practice – neither in the original study (Reed et al., 2014), nor in the follow-up studies that adopted a gene-centric modeling approach (Chavez Rodriguez et al., 2020; Guo et al., 2020; Hui et al., 2021). A drawback of the representation of metabolic plasticity introduced by Reed et al. (2014) is that it does not emerge naturally from the model. Rather, the co-occurrence probabilities need to be assigned a fixed value based on previous knowledge (e. g. metagenomic studies), or estimated from data. Gene-abundance and solute data alone may not provide enough information about the co-occurrence probabilities such that a simpler model not considering metabolic plasticity is preferable.

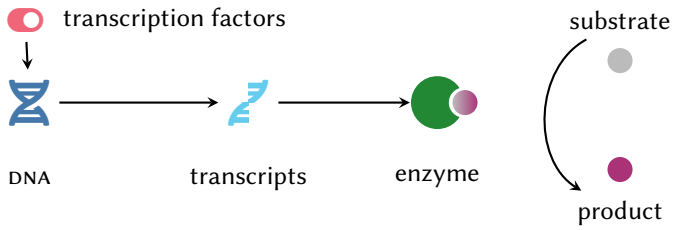


FIGURE 2.1: The production of enzymes from functional genes: Genes are transcribed into transcripts upon regulation by transcription factors. Transcripts are then translated into proteins. Proteins can serve as enzymes that catalyze reactions of the microbial metabolism.

2.1.3 Transcript- and enzyme-explicit models

Biomass- or gene-explicit models implicitly assume that the microbial capacity to transform substrates is constant over time. However, the microbial metabolism is in fact regulated by environmental conditions. Microorganisms use specific proteins, *enzymes*, to catalyze metabolic reactions. The information how to build these enzymes is encoded in the genome of microbes. When the environmental conditions are favorable for a certain process, regulatory proteins, so called transcription factors, initiate the *transcription* of the corresponding genes (figure 2.1). The produced transcripts consist of mRNA and serve as the "construction plan" of enzymes to ribosomes, specialized cell organelles that *translate* transcripts into enzymes by assembling them from smaller building blocks. Biogeochemical models can account for this regulation by explicitly simulating enzyme concentrations.

In an enzyme-based model formulation, the transformation rate of substrate s is based on the enzyme concentration E [mol L^{-1}] instead of functional biomass or gene abundance:

$$r_s = -k_{\max} E f_{\text{substrate}} f_{\text{inhibition}} , \tag{2.8}$$

where the turnover number $k_{\max}$ [s^{-1}] is the maximum amount of substrate that a single enzyme molecule can transform per unit of time. An enzyme-based model also needs to incorporate a description of the production and decay of enzymes in order to simulate their response to environmental conditions. The regulatory processes can be described with different levels of detail.

For example, the model can include a mechanistic description for the production of transcription factors and their regulatory effects (chapter 3; Bælum et al., 2013; Koutinas et al., 2011). In contrast, M. Li et al. (2017b) do not consider transcription factors or transcripts, and base the enzyme production term on the energy available from the corresponding catabolic reaction. Song et al. (2017) describe the production

of transcripts with a cybernetic approach, where a reaction network is regulated such that a metabolic goal (e. g., microbial growth) is maximized.

The production and decay of enzymes can be described by

$$\frac{\mathrm{d}E}{\mathrm{d}t} = k_{\mathrm{trnsl}} T - k^E_{\mathrm{dec}} E, \tag{2.9}$$

where k_{trnsl} [mol transcript^{-1} s^{-1}] is the per-mRNA translation rate, T [transcripts L^{-1}] is the concentration of transcripts, and k^E_{dec} [s^{-1}] is the enzyme decay rate constant (Ingalls, 2013). Note that when enzyme production and decay is described with equation (2.9), enzyme concentrations can be written as the convolution of transcript concentrations with an exponential function, plus a term accounting for initial conditions. This solution can be obtained by applying the Laplace transform to equation (2.9), solving for E, and back-transforming (see appendix A.1).

$$E(t) = k_{\mathrm{trnsl}} \int_0^t T(t) \exp\left(-k^E_{\mathrm{dec}}(t-\tau)\right) \mathrm{d}\tau + E_0 \exp\left(-k^E_{\mathrm{dec}} t\right). \tag{2.10}$$

That is, enzyme concentrations can be thought of as a weighted time-average of transcript concentrations where the weights decrease exponentially with the time scale of enzyme decay. We would therefore expect that enzyme concentrations at a given time point are mostly determined by the transcript concentrations of the preceding minutes to hours, depending on the enzyme half-life.

2.1.4 Reactive transport

Laboratory batch systems are usually well mixed and we can assume that concentrations are uniform in space. We can then treat the system as zero-dimensional in space and describe it with ordinary differential equations (ODEs), considering its evolution over time only. In subsurface environments, on the contrary, transport of solutes through water flow (advection, dispersion) and diffusion need to be considered. While these systems are in fact three-dimensional, there is often a dominant direction of flow and transport. If lateral exchange can be neglected, it is appropriate to adopt a lower-dimensional description of the system, assuming uniform concentrations in the directions perpendicular to the transport direction. For the applications in this thesis, transport and reactions of dissolved compounds (e. g., nitrate, nitrite, oxygen) are therefore described by means of the one-dimensional (1-D) advection-dispersion-reaction equation. The evolution of compound i's concentration c_i in space (x) and time (t) is given by

$$\frac{\partial c_i}{\partial t} + v\frac{\partial c_i}{\partial x} - D\frac{\partial^2 c_i}{\partial x^2} = r^i_{\mathrm{net}}, \tag{2.11}$$

where v [m s^{-1}] is the average linear flow velocity, D [m^2 s^{-1}] is the dispersion coefficient, and r^i_{net} is the net reaction rate of compound i. Dispersion is parameterized according to Scheidegger (1974), where

$$D = |v|\alpha_L + D_e\,. \tag{2.12}$$

Here, α_L [m] is the longitudinal dispersivity and D_e [m^2 s^{-1}] denotes the pore-diffusion coefficient.

In all simulations I neglect transport of bacterial cells because the majority (more than 99 % according to Griebler et al., 2002) of active microorganisms in the subsurface are attached to sediments (H. J. Smith et al., 2018). Transcripts and enzymes are assumed to be confined to the interior of bacterial cells and, thus, to be immobile.

2.2 Numerical methods

2.2.1 Solution of ordinary differential equations

The non-linear systems of ODEs that results from the rate laws of biogeochemical models cannot be solved analytically. Instead, numerical integration schemes can be applied to obtain an approximate solution. Several software libraries implement ODE solvers that only require the user to define the right-hand-side function. For the simulations in the following chapters a backwards differentiation formula (BDF) with variable order (Byrne and Hindmarsh, 1975; Jackson and Sacks-Davis, 1980) as implemented in the CVODES library (Hindmarsh et al., 2005) was used. CVODES also provides methods to compute gradients of the ODE solution with respect to parameters by solving the adjoint sensitivity equations. The models were written in Python using the package sunode (Seyboldt, 2020) that wraps CVODES and produces compiled code for the right-hand-side function using Numba (Lam et al., 2015).

2.2.2 Spatial discretization with the finite volume method

Equation (2.11) is a PDE that cannot be solved analytically, except for simple reaction rate laws such as first order decay. Instead, it needs to be solved numerically, applying discretization methods in time and space. For the applications in this thesis I used a semi-discretization approach: First, the equation is discretized in space using the cell-centered finite volume method (FVM), followed by integration of the resulting ODE system in time with an ODE solver. In the finite volume method, the governing PDE is integrated over a control volume, and the average of the concentration over the control volume $\tilde{c}_i$ becomes the state variable. The integrated transport equation for control volume i can be written as

$$V_i\phi_i\frac{\mathrm{d}\tilde{c}_i}{\mathrm{d}t} + \int_{\Gamma_i}\boldsymbol{n}\cdot\boldsymbol{J}\,\mathrm{d}\Gamma = V_i\phi_i\tilde{r}_i\,, \tag{2.13}$$

where V_i is the volume and ϕ_i is the porosity of the control volume. $\tilde{r}_i$ is the net reaction term for control volume i based on the discretized concentrations. The second term represents the mass flux across the boundary $_i$ of the control volume, where $\boldsymbol{J}$ is the flux density and $\boldsymbol{n}$ the normal vector. For the models in chapters 4 and 5, I implemented the finite volume method for advective-dispersive-reactive transport as a Python package that is publicly available (*adrpy*; Störiko, 2021).

2.3 Parameter estimation and uncertainty quantification

2.3.1 Bayesian inference

Bayesian inference is a powerful statistical framework for estimating parameters and their uncertainty from data. In this framework, probabilities express a degree of knowledge. By specifying prior distributions, the researchers must first lay open their assumptions about parameters to be inferred. Bayesian inference then allows to update the prior beliefs with information contained in the data (van de Schoot et al., 2021).

Previous knowledge about parameters $\boldsymbol{\theta}$ before seeing data is expressed in a probability distribution that is called the prior distribution $p(\boldsymbol{\theta})$, or simply prior. The choice of the prior distribution is usually based on physical constraints and parameter values found in the literature. The likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the probability distribution of the observed data $\boldsymbol{y}$ conditional to the parameters. As the observed data are fixed, it is a function of the parameters. The likelihood function represents the statistical model of the entire data generating process. This data generating model is often assumed to be composed of a deterministic model $\mathscr{M}$, containing the chemical or physical rate laws, and an additive error term ε to account for measurement and model structural errors (Linde et al., 2017; T. Smith et al., 2015):

$$\boldsymbol{y} = \mathscr{M}(\boldsymbol{\theta}) + \varepsilon\,. \tag{2.14}$$

The posterior distribution expresses one's knowledge about parameters based on both previous assumptions and observed data. Bayes' theorem relates it to the prior distribution and likelihood by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y})}\,. \tag{2.15}$$

The marginal likelihood of the data $p(\boldsymbol{y})$ in the denominator is a normalizing constant that is also called the Bayesian model evidence (BME). It is given by the integral of the likelihood over the whole parameter space which is, in most cases, analytically intractable. However, as $p(\boldsymbol{y})$ is a constant, it does not alter the shape of the posterior

distribution, and for many applications it is not necessary to determine its value. Equation (2.15) can then be simplified to

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\,. \tag{2.16}$$

2.3.2 Posterior sampling

The equations describing the posterior distribution are often high-dimensional and complicated. A closed form expression of the posterior distribution is therefore usually not available, prohibiting direct inference (van de Schoot et al., 2021). However, numerical techniques such as Markov chain Monte Carlo (MCMC) methods allow to draw samples from the posterior distribution even if the distribution can only be specified indirectly as in equation (2.16). For this thesis, two different algorithms for sampling from the posterior were used: sequential Monte Carlo (SMC) and the No-U-Turn sampler (NUTS), both as implemented in the software package PyMC (Salvatier et al., 2016). PyMC allows to easily define custom stochastic models, and automatically generates fast C-code for the likelihood function.

SMC, or particle filtering, is based on several MCMC chains that are resampled based on importance weights. SMC sampling can easily be parallelized because the Metropolis chains can run independently between the resampling steps. This makes it a suitable algorithm for Bayesian inference on reactive transport problems that involve evaluating a computationally expensive likelihood function.

When the number of parameters increases, traditional Metropolis-Hastings MCMC rapidly becomes inefficient owing to the geometry of high-dimensional spaces (Betancourt, 2018). In any high-dimensional space there is much more volume outside a given neighborhood than inside of it. When considering probability distributions this means that there is only little probability mass in the neighborhood of the distribution's mode because of its small volume, even though the probability density is highest there. In order to generate posterior samples, a sampler must efficiently explore the typical set, that is, the region of the parameter space where volume and density balance, and that contains the majority of the probability mass. The random proposal distribution used in traditional Metropolis sampling is biased towards the outside of the typical set because the outside volume is much larger. Therefore, most proposed points will lie outside the typical set and will be rejected, making the algorithm inefficient.

In contrast, Hamiltonian Monte Carlo (HMC) methods use gradient information of the probability density to exploit the geometry of the typical set and explore it particularly efficiently (Betancourt, 2018). HMC has been shown to outperform "traditional" MCMC methods (random walk Metropolis, Gibbs sampling, differential evolution Metropolis) in ecological (Monnahan et al., 2017) and hydrological (Krapu

et al., 2019) modeling, particularly for complex, high-dimensional models. NUTS (M. D. Hoffman and Gelman, 2014) is a variant of HMC that eliminates the need to hand-tune hyper-parameters of the sampling algorithm, improving the applicability of HMC in practice.

3 Does it pay off to link functional-gene expression to denitrification rates in reaction models?

3.1 INTRODUCTION

This chapter is based on a co-authored manuscript as indicated in the Statement of Contributions on page xvii.

Quantitative measurements of functional-gene transcripts have been proposed as a proxy variable for microbial reaction rates. However, several biological factors complicate the relationship between transcripts and reaction rate, and it remains an open question whether molecular-biological data can be quantitatively linked to reaction rates. To better establish this link, it could help to integrate a mechanistic description of the production and decay of transcripts and enzymes into biogeochemical models.

Enzyme-based biogeochemical models describe the regulation of microbial reaction rates by the concentrations of corresponding functional enzymes (M. Li et al., 2017b; Song et al., 2017). Compared to other biogeochemical models that use (functional) biomass or genes as a substitute for enzymes, they thus provide a more mechanistic description of microbial reaction processes. However, reactive-transport models that integrate molecular-biological data are unavoidably more complex than traditional Monod-type formulations. Whether the added complexity actually improves model predictability remains understudied.

In this study we integrate transcript data into a denitrification reaction model by linking the expression of functional genes to process rates. We explicitly account for transcriptional regulation of denitrification via transcription factors, translating the current conceptual understanding of the regulatory system in *Paracoccus denitrificans* into a quantitative model. The main pathways of nitrogen transformations – denitrification, nitrification, N-fixation, annamox and DNRA – have been extensively studied due to the relevance of reactive nitrogen compounds for ecosystem functioning (Steffen et al., 2015), groundwater contamination (Gutiérrez et al., 2018), eutrophication (Howarth, 2008), and greenhouse gas emissions (L. Liu and Greaver, 2009). The nitrogen cycle is thus an ideal test case for developing and testing new, enzyme-based models informed by measurements of functional genes and transcripts.

Previous gene-centric and enzyme-based modelling studies focused on systems with slow dynamics (Chavez Rodriguez et al., 2020; M. Li et al., 2017b; Song et al., 2017)

or at steady state (Louca et al., 2016; Reed et al., 2014). Here, we apply our model to a dynamic reactive system, that is, one with rapid shifts in the predominant electron accepting species, and inform it with a highly temporally-resolved dataset of transcript abundances. Previous studies highlighted a potentially hysteretic relationship between transcript concentrations and reaction rates (Bælum et al., 2008; Chavez Rodriguez et al., 2020). Our model allows us to further explore the relationship between transcripts, enzymes, and reaction rates and develop mechanistic interpretations of the observations. Via a comparison with a classical Monod-type model formulation that does not integrate transcriptional regulation we shed light on the potential added benefits of integrating transcript data into (denitrification) reaction models.

3.2 THEORY AND METHODS

3.2.1 Conceptual model description

We set up a model to simulate the experiments of Qu et al. (2015) performed in well-mixed batch reactors. Qu et al. (2015) monitored aerobic respiration and denitrification coupled to succinate oxidation by the denitrifying organism *P. denitrificans*.

Briefly, a series of batch reactors, inoculated with *P. denitrificans*, prepared in an aerobic medium, were amended with 5 mM succinate and 2 mM of NO_3^-. Automated headspace gas measurements (nitrogen gases, O_2) as well as aqueous-phase measurements of nitrite and optical density (OD) facilitated monitoring time series of concentrations driven by reaction and growth kinetics. In addition, and partly driven by headspace gas concentrations, cells were periodically harvested to measure the concentrations of the functional genes *narG*, *nirS*, *norB* and *nosZ*. Following mRNA extraction, transcript numbers were determined via RT-qPCR using the standard curve method. For further details regarding the experimental procedures, we refer to the original publication (Qu et al., 2015).

In the experiment, the contribution of the intermediates NO and N_2O to the mass balance was always less than 1‰. From this experimental observation we conclude that nitrite reduction was the rate-limiting step. Therefore, we set up a simplified model of denitrification simulating the specific experiments by assuming a two-step process (figure 3.1). Therein, the reduction of nitrite via NO and N_2O to N_2 was treated as a single reaction step. Denitrification was coupled to the oxidation of succinate as the sole carbon source and electron donor:

$$7\,NO_3^- + C_4H_6O_4 \xrightarrow{narG} 7\,NO_2^- + 4\,CO_2 + 3\,H_2O\,, \tag{3.1}$$

$$14\,NO_2^- + 14\,H^+ + 3\,C_4H_6O_4 \xrightarrow{nirS} 7\,N_2 + 12\,CO_2 + 16\,H_2O\,\cdot \tag{3.2}$$

In the reaction equations above, the names of the functional genes linked to the reduction of the nitrogen compounds by *P. denitrificans* are given above the arrows. In the presence of oxygen, aerobic respiration is energetically favored over denitrification:

$$2\,C_4H_6O_4 + 7\,O_2 \longrightarrow 8\,CO_2 + 6\,H_2O\,. \tag{3.3}$$

As shown in figure 3.1, the presence of oxygen is assumed to inhibit denitrification. The electron donor, succinate, was assumed to be primarily assimilated for energy conservation by *P. denitrificans*, and thus we do not consider the incorporation of carbon into biomass during growth. Furthermore, the model considers mass transfer between the liquid and the gas phases via a linear-driving-force approximation, and the dilution of the gas-phase concentrations by sampling (see appendix B.5 in the appendix).

We set up two models for comparison: (1) an enzyme-based and a (2) standard Monod-type model. The enzyme-based model considers transcripts and enzymes involved in denitrification reactions as state variables. Reaction rates are directly proportional to enzyme concentrations. The Monod-type model is much simpler and describes denitrification rates using the Monod equation. Therein, the regulation mechanisms and dynamics of catalyzing enzymes are not explicitly reflected, instead the reaction rates are assumed to be directly proportional to the biomass of the denitrifying bacteria.

3.2.2 Governing equations

Our model represents the transcriptional regulation of denitrification genes by simulating transcription factor concentrations in response to oxygen and nitrogen substrates and their effect on transcript concentrations. For comparison, we also set up a simplified model that omits the explicit representation of transcription factors so that transcription directly depends on the concentrations of signaling molecules (see appendix B.6). This simplified model can also be interpreted mechanistically (as discussed in appendix B.6.3), but the assumptions regarding transcriptional regulation differ slightly from the ones presented in the following section.

Several transcription factors regulate the transcription of denitrification genes in *P. denitrificans*, sensing oxygen and nitrogen oxides. The *narG* gene is activated by FnrP and NarR whereas transcription of *nirS* requires NNR (Gaimster et al., 2018). The FnrP protein directly reacts with oxygen which leads to its inactivation (Crack et al., 2016). NarR responds to nitrate and nitrite although the underlying mechanism is currently unknown and the sensing might be indirect (Spiro, 2017; Wood et al., 2001). NNR is activated by NO and deactivated by O_2 (Gaimster et al., 2018; Lee et al., 2006).

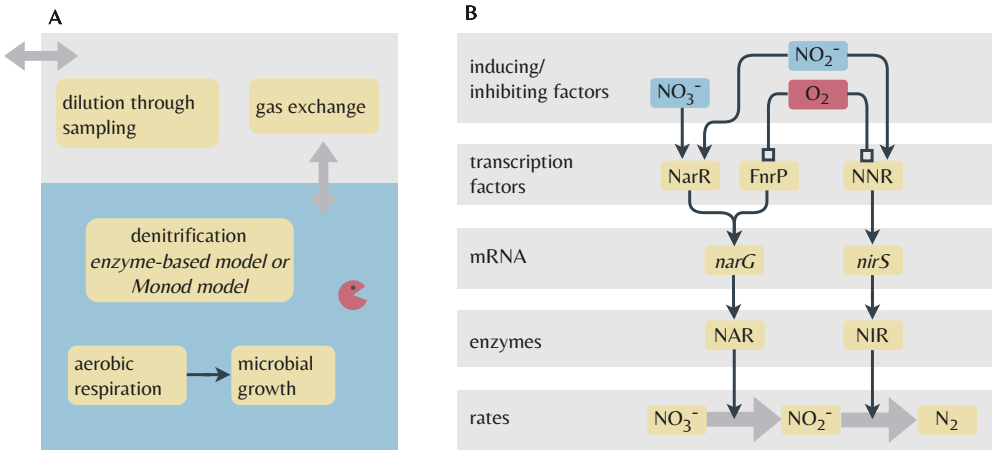


FIGURE 3.1: Schematic representation of processes considered in the models. **A** Model system and processes implemented common to both the Monod-type and enzyme-based models. **B** Simplified representation of gene expression and its link to reaction rates in the enzyme-based model. Narrow arrows represent inducing effects whereas lines ending in a square denote inhibition. In the model, NO_2^- instead of NO is used as activating compound for NNR since we do not simulate NO concentrations.

We assume that the sum of active and inactive transcription factors remains constant throughout the experiment and simulate all active transcription factors relative to the total concentration. Genes that encode the transcriptional regulators (*fnrP*, *narR* and *nnrR*) have been shown to be expressed at similar levels under aerobic and anaerobic conditions in *P. denitrificans* (Giannopoulos et al., 2017), supporting our assumption. We model the fraction of active NarR, X_{NarR} [−], as

$$\frac{\mathrm{d}X_{\mathrm{NarR}}}{\mathrm{d}t} = \left(a_{\mathrm{NarR}}^{\mathrm{NO_3^-}} c_{\mathrm{NO_3^-}} + a_{\mathrm{NarR}}^{\mathrm{NO_2^-}} c_{\mathrm{NO_2^-}}\right)(1 - X_{\mathrm{NarR}}) - k_{\mathrm{dec}}^{\mathrm{NarR}} X_{\mathrm{NarR}}\,, \tag{3.4}$$

where $a_{\mathrm{NarR}}^{\mathrm{NO_3^-}}$ and $a_{\mathrm{NarR}}^{\mathrm{NO_2^-}}$ [$\mathrm{M^{-1}\,s^{-1}}$] are the rate coefficients for activation of NarR by nitrate and nitrite, respectively, and $k_{\mathrm{dec}}^{\mathrm{NarR}}$ [$\mathrm{s^{-1}}$] is the dissociation constant of active NarR. The term $(1 - X_{\mathrm{NarR}})$ represents the inactive fraction of the transcription factor. Binding of oxygen to the transcription factors is described at equilibrium using a Hill function (Gesztelyi et al., 2012). The active fraction of FnrP is

$$X_{\mathrm{FnrP}} = \frac{{I_{\mathrm{FnrP}}}^p}{{I_{\mathrm{FnrP}}}^p + {c_{\mathrm{O_2}}}^p}\,, \tag{3.5}$$

with inhibition constant I_{FnrP} [M] and Hill coefficient p [−]. Experiments have shown that NNR quickly reacts with oxygen, but that its reaction to NO is slower (Lee et al.,

2006). We therefore model the inactivation process at equilibrium while accounting for temporal evolution in the activation process. Although NNR actually senses NO, we simulate its activation by nitrite as we do not explicitly model NO concentrations in our two-step representation of denitrification (see conceptual model description above and figure 3.1). The activation of NNR is described as follows:

$$\frac{\mathrm{d}\hat{X}_{\mathrm{NNR}}}{\mathrm{d}t} = a_{\mathrm{NNR}} c_{\mathrm{NO_2^-}} \left(1 - \hat{X}_{\mathrm{NNR}}\right) - k_{\mathrm{dec}}^{\mathrm{NNR}} \hat{X}_{\mathrm{NNR}}\,, \tag{3.6}$$

where $\hat{X}_{\mathrm{NNR}}$ [−] is the fraction of NNR activated by nitrite (without accounting for inactivation by oxygen), a_{NNR} [$\mathrm{M^{-1}\,s^{-1}}$] is the activation rate constant and $k_{\mathrm{dec}}^{\mathrm{NNR}}$ [$\mathrm{s^{-1}}$] is the dissociation constant of activated NNR. Taking oxygen inhibition into account, the active fraction of NNR X_{NNR} [−] is given by

$$X_{\mathrm{NNR}} = \frac{{I_{\mathrm{NNR}}}^q}{{I_{\mathrm{NNR}}}^q + {c_{\mathrm{O_2}}}^q} \hat{X}_{\mathrm{NNR}}\,, \tag{3.7}$$

with oxygen inhibition constant I_{NNR} [M] and Hill coefficient q.

We assume that the transcription rate for gene i scales with the fraction of operator sites where all activating transcription factors are bound f_{act}^i (Ingalls, 2013):

$$r_{\mathrm{trnscr}}^i = \alpha_i f_{\mathrm{act}}^i B\,, \tag{3.8}$$

where α_i [transcripts $\mathrm{cell^{-1}\,s^{-1}}$] is the maximum transcription rate for gene i and B is the cell density [cells $\mathrm{L^{-1}}$]. For *narG* and *nirS*, the fractions of active operator sites are given by

$$f_{\mathrm{act}}^{nar} = \frac{\frac{X_{\mathrm{FnrP}} X_{\mathrm{NarR}}}{K_{\mathrm{FnrP}} K_{\mathrm{NarR}}}}{1 + \frac{X_{\mathrm{FnrP}}}{K_{\mathrm{FnrP}}} + \frac{X_{\mathrm{NarR}}}{K_{\mathrm{NarR}}} + \frac{X_{\mathrm{FnrP}} X_{\mathrm{NarR}}}{K_{\mathrm{FnrP}} K_{\mathrm{NarR}}}}\,, \tag{3.9}$$

and

$$f_{\mathrm{act}}^{nir} = \frac{X_{\mathrm{NNR}}}{X_{\mathrm{NNR}} + K_{\mathrm{NNR}}}\,, \tag{3.10}$$

where K_{FnrP}, K_{NarR} and K_{NNR} (all dimensionless) are the half-saturation constants for transcription factor binding to the operator site relative to the total transcription factor concentration.

Translation of mRNA into enzymes is described by first-order kinetics:

$$r_{\mathrm{trnsl}}^i = k_{\mathrm{trnsl}} T_i\,, \tag{3.11}$$

in which k_{trnsl} [enzymes transcript^{-1} s^{-1}] is the first-order translation coefficient and T_i [transcripts L^{-1}] is the transcript concentration of gene i. Enzymes (E) and transcripts (T) undergo first-order decay with decay coefficients k^T_{dec} [s^{-1}] and k^E_{dec} [s^{-1}], respectively:

$$r^i_{\text{decay},T} = k^T_{\text{dec}} T_i \,, \tag{3.12}$$

$$r^i_{\text{decay},E} = k^E_{\text{dec}} E_i \,. \tag{3.13}$$

Dynamics of transcripts are usually fast, with transcript half-lives of only a few minutes (Bernstein et al., 2002; Härtig and Zumft, 1999), compared to enzyme half-lives and cell doubling times of several hours (Blaszczyk, 1993; Maier et al., 2011). Therefore, for transcript concentrations, we assumed a quasi-steady state (qss) (Ingalls, 2013):

$$\frac{\mathrm{d}T_i}{\mathrm{d}t} = r^i_{\text{trnscr}} - r^i_{\text{decay},T} = 0 \,, \tag{3.14}$$

which yields the quasi-steady state transcript concentrations T_i^{qss} [transcripts L^{-1}]:

$$T_i^{\text{qss}} = \beta^i_T f^i_{\text{act}} B \,, \tag{3.15}$$

where $\beta^i_T = \alpha_i / k^T_{\text{dec}}$ is the number of transcripts per cell at maximum transcription, that is when f^i_{act} equals 1. Enzyme concentrations are assumed to be governed by mRNA-translation as well as first-order decay:

$$\frac{\mathrm{d}E_i}{\mathrm{d}t} = r^i_{\text{trnsl}} - r^i_{\text{decay},E} \,. \tag{3.16}$$

The translation rate constant is difficult to measure and the prior range of this parameter is therefore essentially unconstrained. However, we can express it in terms of parameters that are easier to estimate by computing the quasi-steady state (qss) of enzyme concentrations, E_i^{qss} [enzymes L^{-1}], given by

$$E_i^{\text{qss}} = \frac{k^i_{\text{trnsl}} \beta^i_T}{k^E_{\text{dec}}} f^i_{\text{act}} B \,. \tag{3.17}$$

We defined $\beta_E = k^i_{\text{trnsl}} \beta^i_T / k^E_{\text{dec}}$ as the maximum quasi-steady state enzyme concentration per cell, a parameter that is easier to constrain based on literature data. We can then express the translation rate constant as

$$k_{\text{trnsl}} = k^E_{\text{dec}} \frac{\beta_E}{\beta^i_T} \,. \tag{3.18}$$

The transformation rate of nitrogen substrate j (NO_3^- or NO_2^-) with corresponding enzyme i (NAR or NIR) is described by Michaelis-Menten kinetics considering kinetic inhibition of the reaction by O_2:

$$r_j = r^i_{\max} \frac{c_j}{c_j + K_j} \frac{I^i_{\text{reac}}}{I^i_{\text{reac}} + c_{O_2}}, \tag{3.19}$$

in which the maximum rate constant $r^i_{\max}$ $\left[\text{mol L}^{-1}\,\text{s}^{-1}\right]$ is the theoretical rate at full substrate saturation without inhibition. The second and third term of equation (3.19) describe substrate limitation and oxygen inhibition of the enzyme, in which c_j [M] is the concentration of substrate j, K_j is the half-saturation constant [M] of substrate j and I^i_{reac} [M] is the oxygen-inhibition constant for the corresponding enzyme (i).

In a Monod-type model, the enzyme concentration inside a cell is assumed to be constant and $r_{\max}$ is proportional to the cell density:

$$r^i_{\max} = \nu^i_{\max} B, \tag{3.20}$$

in which $\nu^i_{\max}$ [mol cell^{-1} s^{-1}] represents a constant, cell-specific maximum substrate-consumption rate. In the enzyme-based model, the transformation rate of substrate i is given by

$$r^i_{\max} = k^i_{\max} E_i, \tag{3.21}$$

in which the turnover number $k^i_{\max}$ [s^{-1}] is the maximum amount of substrate that a single enzyme molecule can transform per unit time.

In both models (enzyme-based and Monod-type), aerobic respiration is described by a Monod rate law:

$$r_{O_2} = \nu^{O_2}_{\max} B \frac{c_{O_2}}{c_{O_2} + K_{O_2}}, \tag{3.22}$$

with the maximum cell-specific rate $\nu^{O_2}_{\max}$ [mol cell^{-1} s^{-1}] and half-saturation constant K_{O_2} [M]. *P. denitrificans* can utilize both aerobic respiration and denitrification for growth. However, typically reported growth yields for aerobic respiration are higher than those reported for denitrification (Boogerd et al., 1984) and, under the reported experimental conditions (Qu et al., 2015), growth using oxygen as the electron acceptor was the dominant process. Therefore, we assume that *P. denitrificans* only grows on aerobic respiration and that denitrification steps do not result in growth:

$$\frac{\mathrm{d}B}{\mathrm{d}t} = Y_{O_2} \frac{n_{C_4H_6O_4}}{n_{O_2}} r_{O_2}, \tag{3.23}$$

in which Y_{O_2} [cells mol^{-1} of succinate] is the growth yield and $n_{C_4H_6O_4} = 2$ and $n_{O_2} = 7$ are the stoichiometric coefficients of succinate and oxygen in the energy reaction, respectively.

The flasks used in the experiment had a headspace. We accounted for the mass transfer rates r_{tr}^{i} of gaseous compounds (N_2 and oxygen) between the water and the gas phase. Furthermore, we also considered dilution of the gas phase by sampling. The dynamics of solute concentrations of oxygen and nitrogen species are given by

Details on the description of gas dynamics can be found in appendix B.5.

$$\frac{\mathrm{d}c_{\mathrm{O_2}}}{\mathrm{d}t} = -r_{\mathrm{O_2}} + r_{\mathrm{tr}}^{\mathrm{O_2}}, \tag{3.24}$$

$$\frac{\mathrm{d}c_{\mathrm{NO_3^-}}}{\mathrm{d}t} = -r_{\mathrm{NO_3^-}}, \tag{3.25}$$

$$\frac{\mathrm{d}c_{\mathrm{NO_2^-}}}{\mathrm{d}t} = r_{\mathrm{NO_3^-}} - r_{\mathrm{NO_2^-}}, \tag{3.26}$$

$$\frac{\mathrm{d}c_{\mathrm{N_2}}}{\mathrm{d}t} = \frac{1}{2} r_{\mathrm{NO_2^-}} + r_{\mathrm{tr}}^{\mathrm{N_2}}. \tag{3.27}$$

3.2.3 Parameter identification

A subset of the model parameters was fixed, either because these parameters were known from the experimental setup, or they could be well constrained by literature values. In total, we estimated 14 parameters of the Monod-type model and 32 parameters of the enzyme-based model with a Bayesian approach. Given the evidence of observed data and prior knowledge, we obtain the *posterior* probability distribution of parameter values. Following Bayes' law (equation (2.16)), the conditional probability density of parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$, $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, is proportional to the prior probability density $p(\boldsymbol{\theta})$ of the parameters and the likelihood of the measured data $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

The prior distribution constrains the parameters on knowledge uninformed by the data, such as literature values or physical constraints. The likelihood describes the probability density of the measured data if the model parameters were correct. For a perfect model, the likelihood only describes measurement errors, but in most applications it also include the effects of conceptual errors on meeting the data. We used the likelihood function

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i,j} T(y_{ij} \mid \mu_{ij}, \sigma_{ij}, \nu), \tag{3.28}$$

in which T is the Student's t-distribution with scale σ_{ij}, degrees of freedom ν and location

$$\mu_{ij} = \mathrm{boxcox}(c_{ij} + b_i, \lambda_i). \tag{3.29}$$

The index i refers to the measured variable, j indicates the measurement number and boxcox is the Box-Cox transform (Box and Cox, 1964). All measured data from the experiment were Box-Cox transformed to ensure homoscedasticity of the residuals.

We added a constant background value b_i to the simulated concentration c_{ij} of oxygen and nitrite because measured concentrations never reached zero in the experiment, but stayed at a very low background value (<0.1 µM for NO_2^- and <0.2 µM for O_2).

The errors σ_{ij} are the sum of data errors computed from the standard deviations of triplicate measurements and a constant error that accounts for model structural errors which we choose to be an estimation parameter. We chose a Student's t-distribution with $\nu = 10$ instead of a normal distribution so that outliers do not get too much weight.

The model was implemented in the programming language Python and the code is freely available (Störiko et al., 2021b). We solved the system of ODEs using the Python package Sunode (Seyboldt, 2020), a Python wrapper to the CVODES library (Hindmarsh et al., 2005). Obtaining numerically stable results over a broad range of parameter values was difficult. Therefore, the ODE was log-transformed and solved using a backward differentiation formula of variable order (Byrne and Hindmarsh, 1975; Jackson and Sacks-Davis, 1980) with small tolerances (10^{-12} for the absolute and relative tolerance in the forward problem, 10^{-8} in the adjoint problem). Samples from the posterior distribution were drawn using the Python package PyMC3 (Salvatier et al., 2016) and analyzed with ArviZ (Kumar et al., 2019).

We sampled all models with NUTS (No-U-Turn sampler; M. D. Hoffman and Gelman, 2014), a Hamiltonian Monte Carlo algorithm implemented in PyMC3, with adjusted settings for the tuning phase (Foreman-Mackey, 2020). It uses gradient information which we obtained by solving the adjoint sensitivity equations. We used the $\widehat{R}$-criterion (Vehtari et al., 2021) that compares the variance between different chains to the in-chain-variance to assess convergence. The largest $\widehat{R}$ is 1.005 for the Monod-type model and 1.021 for the enzyme-based model – the values close to one indicate convergence.

Table B.2 in the appendix lists fixed parameter values, prior distributions, and statistics about the posterior distributions. Table B.1 provides initial values of the simulations.

Enzyme concentrations were not measured in the experiment and were only indirectly constrained through reaction rates. As all rate laws contain the product of enzyme concentrations and the maximum enzyme-specific turnover rate k_{max}, it is impossible to estimate both k_{max} and β_E without direct measurements of enzyme concentrations. Therefore we fixed β_E to an arbitrary value of 1125 enzymes cell^{-1} based on measured intracellular enzyme concentrations in the literature (Maier et al., 2011). This implies that the fitted values of k_{max} and the simulated enzyme concentrations within bacterial cells have to be interpreted with care, as they are conditioned on the arbitrary choice of β_E.

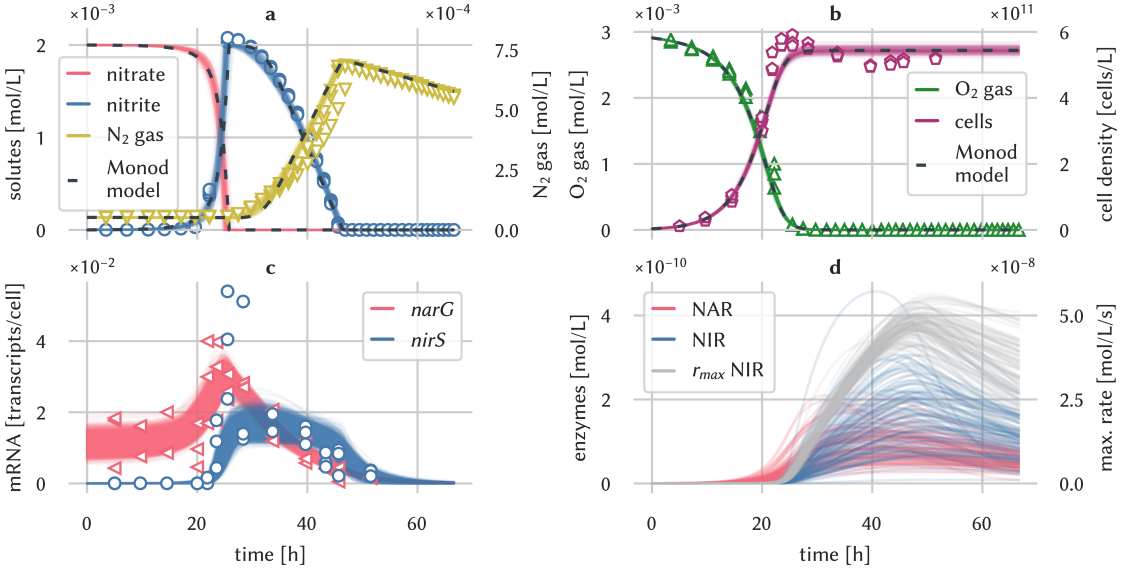


FIGURE 3.2: Simulated dynamics, 100 draws from the posterior of the enzyme-based model (solid lines) and posterior median of the Monod-type model (dashed lines), and measurements from three experimental replicates (open symbols). **a** Nitrogen compounds, **b** oxygen and cell densities, **c** transcripts, **d** enzymes and maximum rates.

3.3 RESULTS AND DISCUSSION

3.3.1 *Modeled concentration time series*

Figure 3.2 shows a comparison of the simulated concentrations of the enzyme-based and the Monod-type models with the experimental data (Qu et al., 2015). Figure 3.2a shows the concentrations of nitrogen species (nitrate, nitrite, N_2) of the models and the experiments, figure 3.2b the oxygen concentrations and cell densities, and figure 3.2c the transcripts of *narG* and *nirS*, respectively, whereas figure 3.2d contains simulated enzyme concentrations for which no measurements were available. Fitted parameter values and their uncertainties are outlined in table B.2, and simulated transcription factor concentrations are shown in figure B.4.

Concentration times series results in figure 3.2a and b are present as several draws of the posterior distribution for the enzyme based model and the median output for the Monod-type model. Results for the Monod-type model exhibited a similarly narrow spread of simulated concentrations, and thus multiple draws were omitted. Both model formulations fit the concentration data equally well. The excellent fit supports the validity of our conceptual model and the underlying assumptions outlined above.

Oxygen is consumed before nitrate and nitrite are reduced to molecular nitrogen, which confirms the inhibition of denitrification by oxygen. The simulations predict that nitrate is fully converted to nitrite before the onset of the second denitrification step. Both models capture the sharp increase in nitrite concentrations, peaking after 25 h at 2 mM, followed by a gradual decrease over the ensuing 20 h, and the subsequent production of N_2 gas, which is the final product of denitrification. The posterior distributions of solute and gas concentrations in both models were narrow, reflecting the small measurement variance. Even though nitrate concentrations were not measured by Qu et al. (2015), the nitrate simulation results are indirectly well constrained by nitrite data and the mass conservation assumption encoded in the model equations.

At the time of highest nitrite concentrations, the measured cell densities reached peak values and gradually decreased (figure 3.2b). Neither of the models was able to capture the slightly non-monotonic behavior in cell densities with a peak at 25 h and a subsequent modest decline. The slight discrepancy between simulated and measured cell densities could be due to several assumptions of our model. We assumed that *P. denitrificans* only grows on oxygen, even though the organism is known to grow on both oxygen and nitrate (Boogerd et al., 1984), which may explain growth beyond the time of oxygen availability.

The subsequent slight decrease in cell densities, if not a measurement artifact, may be a toxicity effect of nitrite or one of the non-modeled intermediates. Nitrite is known to be toxic to *P. denitrificans* and other microorganisms at millimolar concentrations (Stouthamer, 1980; van Verseveld et al., 1977; H. Zhang et al., 2013). Accounting for nitrite toxicity has been postulated as an explanation for delayed growth and increased cell lysis of *Shewanella oneidensis* while reducing nitrite to ammonium (Mellage et al., 2019). Altogether, we deemed the miss-fit in cell densities not significant enough to justify making the model more complex by introducing additional processes with difficult-to-constrain parameters.

3.3.2 Transcript and enzyme dynamics

Overall, the enzyme-based model formulation captured the dynamics of measured *narG* and *nirS* concentrations (figure 3.2c). Both simulated and measured *narG* transcript concentrations fluctuated around a baseline value of 0.01 transcripts per cell, and *nirS* levels were very low during the oxic phase. Following a drop in oxygen levels, simulated *narG* transcripts rose, followed by those of *nirS*. The model satisfactorily captured the peak of *narG* transcripts at 23 h and the following gradual decrease to low, but detectable transcript levels at the last measurement time point, at about 50 h. The model also predicted the very abrupt increase of *nirS* transcripts at 24 h. While peak levels of measured *nirS* transcripts at 25 h could not be matched,

the measurement uncertainty of the transcript data at that time point was very large (0.024 transcripts cell^{-1} to 0.054 transcripts cell^{-1}), and thus we considered these peak values as outliers. Our assumption is supported by additional data of a parallel experiment with the same setup, but using butyrate instead of succinate as the carbon source, where a peak in *nirS* transcripts was not detected (Qu et al., 2015). Analogous to measured data, both of the simulated transcripts remained at concentrations above the limit of detection for a few hours after complete denitrification.

As seen in figure 3.2d, enzyme concentrations were subject to considerable uncertainty – an expected result, as there are no data to constrain them. However, the maximum cell-specific turnover rate of NIR $r_{\max}^{\mathrm{NIR}}$, given by the product of enzyme concentrations and the maximum enzyme specific turnover rate, was well constrained, provided that substrate was being consumed. The solute data that we used for conditioning could apparently provide sufficient information about reaction rates to estimate the maximum rate.

We also setup a model version with a more simplified description of transcriptional regulation, for comparison. The model, described in appendix B.6.1, yielded similar results (figure B.6) to those outlined in figure 3.2. However, the simplified model formulation was unable to capture the sudden increase in *nirS* transcripts (at $t = 24$ h). Therefore, we opted for the more complex transcription-factor based formulation.

3.3.3 Posterior parameter uncertainty

Figure 3.3 shows prior and posterior distributions of selected parameters. In our application, most posterior parameter distributions were much narrower than the prior ranges (figure 3.3a), indicating that they were strongly constrained by the available data (e. g., Y_{O_2}). Where applicable, the Monod-type and enzyme-based models mostly resulted in similar parameter distributions. However, the estimated enzyme inhibition constants $I_{\mathrm{reac}}^{\mathrm{NAR}}$ and $I_{\mathrm{reac}}^{\mathrm{NIR}}$ were smaller in the Monod-type model compared to the enzyme based model because the Monod-type model compensates the lacking regulation of enzyme levels with a stronger inhibition of enzymes by oxygen. Some parameters exhibit considerable spread even after conditioning to the data, partly ranging over several orders of magnitude. This holds mainly for parameters that are related to transcription factor, transcript and enzyme concentrations such as transcription factor activation rate constants (e. g. $a_{\mathrm{NarR}}^{\mathrm{NO_3^-}}$) or the enzyme half-life $t_{1/2}^{E}$.

Marginal distributions of all parameters can be found in figure B.1 in the appendix.

In some cases, part of the spread can be attributed to the correlation among the conditional parameters, as illustrated in figure 3.3b for selected examples. For example, the enzyme half-life $t_{1/2}^{E}$ strongly correlates with the "efficiency" of NIR enzymes expressed by $k_{\max}^{\mathrm{NO_2^-}}$. The enzyme half-life influences the concentration of the enzymes. Because the total reaction rate depends on the product of the enzyme concentration with the efficiency of a single enzyme, changing the enzyme half-life can be compen-

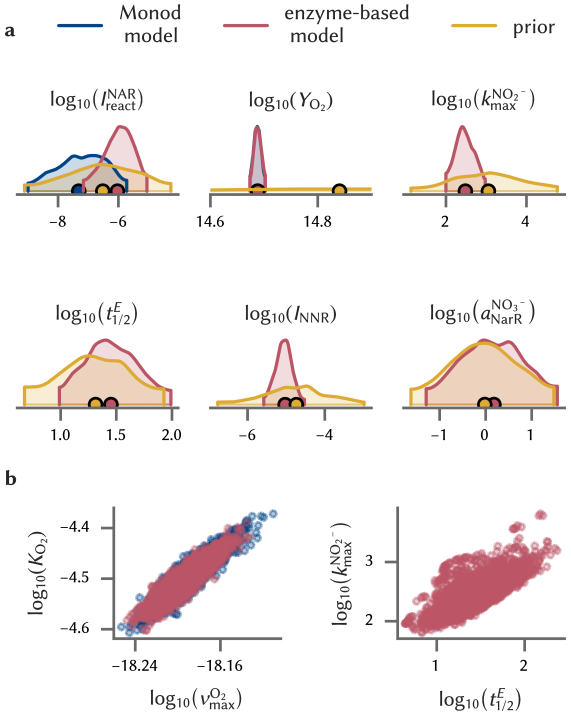


FIGURE 3.3: Selected marginal parameter distributions. **a** Kernel density estimates of the marginal distributions of selected parameters. Densities are cut off at the 94 % highest density intervals, circular markers indicate the mean. Parameters $k_{max}^{NO_2^-}$, $t_{1/2}^E$, I_{NNR} and $a_{NarR}^{NO_3^-}$ are not present in the Monod-type model such that the posterior is only shown for the enzyme-based formulation. Posterior distributions of Y_{O_2} are similar in both model formulations and cannot be distinguished in the plot. **b** Pairwise joint posterior distributions of parameters with strong correlation.

sated by also changing the enzyme efficiency. Thus, if we had enzyme data, we could significantly decrease the uncertainty in enzyme half-lives and the enzyme-specific maximum turnover rate $k_{\max}^{NO_2^-}$ of nitrite.

A strong correlation can also be observed between the Monod parameters related to aerobic respiration, which has been previously reported (C. Liu and Zachara, 2001) (figure 3.3b). Reparameterization alleviated the correlation between $v_{\max}^{O_2}$ and K_{O_2}. The parameter $v_{\max}^{O_2}$ is the cell-specific respiration rate at limit of the substrate concentration approaching infinity. Instead, we used the cell-specific respiration rate at a fixed concentration $v_{\text{fix}}^{O_2}$ as a parameter (see appendix B.4). $v_{\text{fix}}^{O_2}$ and K_{O_2} were less correlated than before the reparametrization, but the correlation of $v_{\text{fix}}^{O_2}$ with the growth yield was larger.

Correlation coefficients of all posterior distributions are plotted in figures B.2 and B.3 in the appendix.

3.3.4 Relationship between reaction rates and transcript concentrations

It has been previously suggested to use transcript concentrations as a proxy for reaction rates (Achermann et al., 2020; Brow et al., 2013; Freitag and Prosser, 2009; Nicolaisen et al., 2008; Rohe et al., 2020). To test if such a relationship would be a valid assumption, we plot the simulated transcript concentrations and corresponding reaction rates against each other in figure 3.4a and d. Cell-specific rates were calculated by applying the rate laws to the simulated concentrations and normalizing the result by the cell densities. Measured transcript concentrations plotted against the reaction rates based on the Monod-type model are shown for comparison.

Transcript concentrations of *narG* and *nirS* are upregulated before rates increase (figure 3.4a and d). Between 10 h to 25 h, *narG* transcript show a positive, but non-linear correlation with reaction rates, before the rates drop to zero once nitrate is depleted. Nitrite reduction rates, in contrast, only start increasing when *nirS* transcript levels have nearly reached their maximum (figure 3.4d). Rates continue to increase when *nirS* transcript levels start to drop such that rates and transcript levels are anti-correlated around 40 h, just before reaction rates drop abruptly upon substrate depletion. After the rapid drop in reaction rates, both transcripts remain present for several hours.

We also looked at the relationship of enzyme concentrations with denitrification rates (figure 3.4b and e). Enzyme concentrations are not well constrained by the data and their posterior uncertainty propagates to the relationship with reaction rates. We therefore plotted the product of enzyme concentrations and their "efficiency" $k_{\max}^{NO_2^-}$, corresponding to the maximum potential reaction rate, without accounting for enzyme inhibition or substrate limitation (figure 3.4c and f).

The maximum rate of NIR is well constrained and shows an almost linear relationship with actual reaction rates up to the point where nitrite depletion cuts the rate (figure 3.4f) The rate of nitrite reduction is essentially not limited by oxygen inhibition

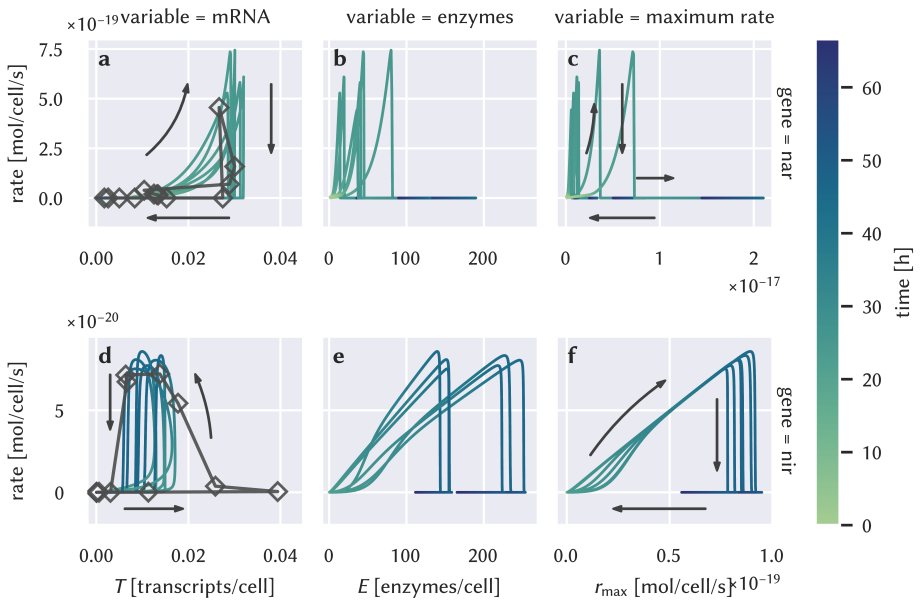


FIGURE 3.4: Cell-specific substrate turnover rates plotted against concentrations of the corresponding transcripts (panel **a** and **d**) or enzymes (panel **b** and **e**) and the maximum specific rate (panel **c** and **f**) linked to the reaction. The plot shows several draws from the posterior (colored lines). Measured transcript data are plotted against rates based on the Monod model (dark gray lines and rectangles).

or substrate limitation. The maximum rate is thus the determining factor such that an increase in enzyme concentrations directly increases the rate. The relationship of NAR maximum rates and actual nitrate reduction rates (figure 3.4c) is more complicated. The initial increase is strongly non-linear and the remaining posterior spread of the maximum rate leads to uncertainty in the scaling factor.

Both findings can be explained by the initial inhibition of NAR enzymes by oxygen: NAR enzymes are produced while oxygen is still present so that r_{max} increases, but the actual rate does not, driven by the weaker inhibition of *enzyme production* relative to that of the *enzymatic catalysis*. The posterior uncertainty of the relationship is due to correlation between the oxygen inhibition parameter and the maximum rate. Different combinations of the parameters can yield the same reaction rate and an unambiguous assignment is not possible for a case when the data only constrain the effective rate.

Our results show that under the dynamic conditions of the experiment, transcript concentrations are not a good predictor of reaction rates. The complicated relationships are mainly caused by the asynchronism and different time scales of substrate dynamics, transcript and enzyme production, and decay. Under environmental conditions, changes in concentrations of redox-active species may be much slower than in the laboratory experiment simulated in this study. In groundwater, for example, the typical time scales are on the order of several months (Arora et al., 2016, 2013).

If changes in redox-active species are slower than the enzyme dynamics (acting on the order of days), we would expect a better correlation between transcript and enzyme concentrations and reaction rates, potentially allowing for further simplifications of model formulations. However, many environments in which nitrogen cycling plays a crucial role are very dynamic. For example, rapid water-content driven changes in the redox state of soils (Pronk et al., 2020) or diurnal fluctuations in river biogeochemistry (Kunz et al., 2017) can control nitrogen turnover. The lack of obvious correlations between transcripts of functional genes and reaction rates observed in our modeling results are likely representative for such dynamic environments.

A non-linear, hysteretic relationship between transcript concentrations and reaction rates has also been found in a gene-centric model of pesticide degradation (Chavez Rodriguez et al., 2020). Interestingly, this is the case although pesticide degradation acts on much longer time scales than denitrification in this study and the observed relationship between rates and transcripts shows different patterns than the ones presented in figure 3.4 (e. g., direction of the hysteresis). This highlights the need to assess the relationship between gene expression and rates for each reaction system individually. Nevertheless, models that explicitly simulate transcript concentrations, as presented in this study, can help to understand the relationship between gene expression and rates and extrapolate it.

If enzyme concentrations were at quasi-steady state, the enzyme-based model could

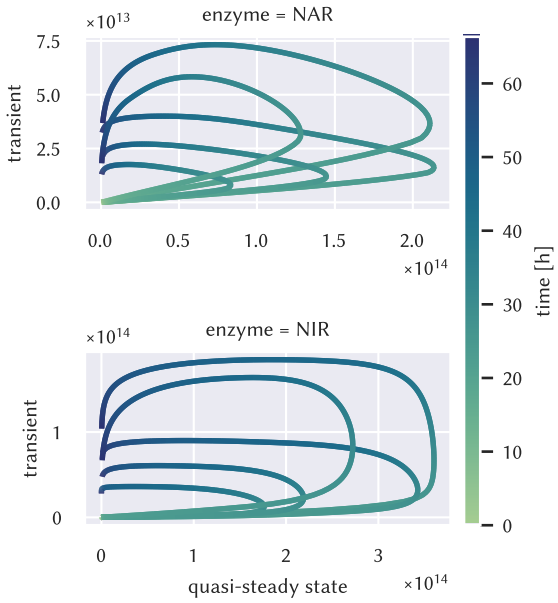


FIGURE 3.5: Quasi-steady enzyme state concentrations [enzymes L^{-1}] plotted against the concentrations simulated by the fully transient model. Although the exact relationship varies between different draws from the posterior (indicated by several lines), all of them are characterized by hysteresis in time (colors indicate the time point).

be simplified, reducing the number of state variables and parameters. To test if this assumption is valid, we ran the fully transient model version (as shown in figure 3.2) and computed the quasi-steady state concentrations based on the fully transient results. If the quasi-steady state assumption were valid, the two concentrations should cluster along the 1:1-line. As evident from figure 3.5, this is not the case – enzyme concentrations exhibit a strongly hysteretic relationship and remain much lower than the corresponding quasi-steady state concentrations. Enzyme dynamics, as measured by estimated enzyme half-lives, are considerably slower than substrate dynamics such that the substrate stimulus is too short for enzyme concentrations to reach their quasi-steady state value.

3.3.5 Overall model performance

Our parsimonious enzyme-based model describes expressional patterns well and simulates denitrification rates very accurately. The model quantitatively links concentrations of functional transcripts to turnover rates. In reproducing the solute-concentration data of Qu et al. (2015), however, the Monod-type model is as good as the enzyme-based model. Small data errors lead to narrow posterior distributions of the solute concentrations in both models, whereas the much more uncertain transcript data do not help to further constrain them. While the enzyme-based model can be

calibrated using more types of data than the Monod-type model, it also requires estimating more parameters, which appear to be weakly constrained.

The strength of the enzyme-based model lies in improving our mechanistic understanding in a quantitative way. For example, it helps to decompose the contributions of transcriptional regulation and enzyme inhibition to the down-regulation of denitrification rates by oxygen. It confirms that the conceptual model of transcriptional regulation of *narG* and *nirS* through FnrP, NarR and NNR can also explain expressional patterns quantitatively. Finally, the quantitative model helps to understand the non-linear relationship between rates and transcript or enzyme concentrations. It can be used to generate hypotheses about the behavior in other systems and develop sampling strategies for these quantities.

Transcription-factor regulation and enzyme dynamics need to be explicitly described as these processes take place at similar or even larger time scales than the actual enzymatic reactions and the corresponding substrate dynamics. The relation between transcript concentrations and reaction rates exhibits a strong hysteresis. Enzyme concentrations show a better correlation with reaction rates at early times but are uncorrelated at later times due to substrate limitations. Thus, enzyme kinetics and the abundance of enzymes present likewise have an impact on reaction rates and need to be taken into account.

3.3.6 Transferabilty to environmental systems

While the experimental conditions during the experiment we simulated are not representative for many environments (high succinate concentrations, single organism), the controlled conditions allowed us to set up a detailed and mechanistic model of the denitrification reactions. The presented model equations hereafter can be integrated in a model that better represents natural environments, for example by accounting for transport or carbon limitation. Representing all denitrifiers by a single organism in the model certainly does not account for all interactions that occur in natural microbial communities. Considering a single pool of denitrifiers in quantitative reaction models, however, is a common practice to keep the simulations computationally and conceptually tractable (Kinzelbach et al., 1991; Sanz-Prat et al., 2015; Yan et al., 2016).

In addition, *P. denitrificans* is a denitrifying organism that is actually present in natural environments like soils (Nokhal and Schlegel, 1983). The exact mechanism of transcriptional regulation is not the same in all denitrifiers, but some basic principles are similar in various studied organisms (Gaimster et al., 2018). Further, the exact pattern of gene expressional dynamics does not matter for reaction rates because they are smoothed out by the much slower enzyme dynamics. A simplified representation of the transcriptional regulation as we present in appendix B.6 might describe the behavior of a diverse microbial community sufficiently well, where the effective

expressional patterns would be some smoothed average of the individual patterns produced by different regulation mechanisms.

An alternative approach to modeling transcriptional regulation that does not require exact knowledge of the underlying mechanism has been presented previously by Song et al. (2017). The latter authors applied a cybernetic approach (Song and C. Liu, 2015) that relies on the principle of return-on-investment: Microorganisms will transcribe those genes that maximize their metabolic "profits" (though it is often unclear how to express that in quantitative terms).

Compared to the simulation results of the enzyme-based denitrification model presented earlier (Song et al., 2017), we can observe some key differences. While transcript concentrations respond relatively quickly to changes in substrate concentrations, Song et al. (2017) predict a long delay in the response of transcripts. At the same time, enzyme concentrations respond to transcripts quickly in their model. While this is one possibility to explain the metabolic lag they observed, it is not necessary to ascribe the delay to slow *transcription* dynamics. Transcript dynamics in their model are not constrained by data so the actual lag could also be due to delayed *translation*.

Differences between the model outcomes might be attributed to the different environmental systems. The laboratory system that we simulated exhibited faster kinetics and was more dynamic than the sediment that Song et al. (2017) analyzed (complete denitrification within two days with the pronounced accumulation of reactive intermediates, compared to a week). In contrast to the aforementioned study (Song et al., 2017), our simulations and transcript data support an immediate transcriptional response to denitrifying conditions, justifying the omission of a resource pool which first builds up before transcription starts in the model formulation.

The authors of the experimental data used here measured biomass in cell densities, which we also chose as state variable. In the natural environment, functional biomass is difficult to measure and a different approach is needed. A gene-centric approach which uses gene concentrations as proxies for functional biomass (Pagel et al., 2016; Reed et al., 2014) could be readily integrated into our model. Gene-centric approaches have been mostly applied to marine environments and their application to soils or groundwater might need extensions to account for relic DNA which has been shown to represent as much as 40 % (Carini et al., 2016).

3.3.7 Implications for biogeochemical modeling

We informed our model with highly resolved time series of all solute concentrations in a controlled experiment. Under these conditions, the quantitative prediction of reaction rates was not improved by integrating transcript data. Since substrate and product concentrations are easier to obtain than transcript concentrations, quantitative rate

predictions can be more effectively improved by additional chemical measurements than by integrating transcript concentrations. By contrast, transcript data can support the identification of key reactions, particularly if intermediate products have low concentrations and cannot be detected. In addition, transcript based models can fill a key gap in our predictive capabilities within natural systems where the difficulty in acquiring cell density information impedes the validation of biomass-explicit models.

The advancement of enzyme-based models requires appropriate data for evaluating models and testing hypotheses. Data sets with quantitative information about functional genes, transcripts *and* enzymes could help to find accurate and parsimonious descriptions of transcript and enzyme dynamics. Measuring enzyme concentrations (e. g., by targeted quantification of functional enzymes, see M. Li et al., 2017a) seems more promising than transcript concentrations as they more closely relate to actual turnover rates. However, measuring specific proteins in environmental samples may be restricted by major challenges, such as the efficient extraction of proteins from environmental samples, hindering accurate quantification and limiting the wider application of functional enzyme measurements (Starke et al., 2019).

We are convinced that further improvements of enzyme-based models in environmental systems can be achieved by integrating data from experiments carried out under different environmental conditions, expanding the sensitivity range of simulated processes and parameters. While batch experiments with high cell densities and nutrient concentrations usually exhibit fast dynamics there is a need to condition models to data from less dynamic systems such as chemostats and flow-through columns, which may more closely relate to natural environmental conditions.

4 Denitrification-driven transcription and enzyme production at the river–groundwater interface

4.1 INTRODUCTION

This chapter is based on a co-authored manuscript as indicated in the Statement of Contributions on page xvii.

The increase of diffuse nitrogen inputs, mainly by agriculture, has led to elevated concentrations of reactive-nitrogen species in groundwater and surface-water bodies, threatening drinking-water production, and causing eutrophication of rivers and lakes (Erisman et al., 2013). Microorganisms use reactive nitrogen compounds as substrates for redox reactions that fuel their energy metabolism, constituting the main attenuation process for nitrogen contamination in environmental systems (Kuypers et al., 2018). Understanding the factors that foster microbial removal of reactive nitrogen species from the environment is therefore critical for contamination control and mitigation. Denitrification is the key reaction for the permanent removal of nitrogen species from the environment because it converts the reactive-nitrogen species nitrate into inert N_2 gas rather than into another reactive-nitrogen species. The interface between surface waters and groundwater plays a key role for the turnover of nitrogen compounds because steep redox gradients (from oxic rivers to anoxic groundwater) and the availability of labile organic carbon as an electron donor, either in the river water or in the hyporheic and riparian zones, enhance microbial reactions (Krause et al., 2011, 2017).

Molecular-biological tools and so-called omics techniques, i. e., (meta)genomics, (meta)transcriptomics, (meta)proteomics analyses, have been used to characterize microbial nitrogen cycling in riparian zones (S. Wang et al., 2019), lake and river sediments (Reid et al., 2018; Stoliker et al., 2016), and the hyporheic zone (Danczak et al., 2016) by providing information about the microbial community composition, its functional and metabolic potential, and activity. Typically, genomic data are thought to provide information on metabolic *potential*, whereas transcript data are seen as an indicator of microbial *activity*.

While these methods can help to identify the relevant processes at a particular site and outline reactive zones, it remains a challenge to quantitatively relate molecular-biological measurements to turnover rates of nitrogen. Many sequencing studies

target taxonomy and diversity of organisms, without providing direct information about reactions rates. Meta-omics data primarily target the relative (qualitative) abundance of genes, transcripts, and proteins and are particularly difficult to convert into rate expressions. In contrast, measurements of functional genes, their transcripts, and the corresponding enzymes directly relate to the abundance of organisms capable of specific metabolic pathways and their activity. Several studies have suggested using transcript levels or transcript-to-gene ratios to estimate reaction rates of contaminant (Brow et al., 2013; Rahm and Richardson, 2008), pesticide (Monard et al., 2013) or nitrogen-species turnover (Rohe et al., 2020).

However, several factors may complicate the relationship between transcript concentrations and reaction rates: For example, enzyme levels can be disconnected from transcript levels due to their different half-lives, and concentrations of substrates in addition to enzyme levels regulate the catalysis rate (Moran et al., 2013; Störiko et al., 2021a). It therefore remains unclear, if concentrations of transcripts or enzymes in the environment can be used as a proxy for reaction rates. In a meta-analysis of experimental studies, Rocca et al. (2015) found that transcript abundances were not significantly correlated with biogeochemical processes. They conclude that it is essential to investigate the factors controlling gene abundance, transcription, translation and enzyme activity to better interpret the patterns of gene and transcript abundances in the environment with respect to ecosystem processes.

Due to the high analytical costs of molecular-biological analyses, spatially and temporally highly resolved measurements of gene, transcript, or enzyme concentrations to date hardly exist. Process-based modeling may help to bridge between a limited number of molecular-biological measurements, continuously logged physical and chemical parameters (e. g., using probes), and the need to understand the system's biological, chemical, and physical functioning at scales relevant for management. Modeling may also be used as a tool to investigate how transcript and enzyme concentration relate to reaction rates, and how different factors influence the relationship.

A few modeling approaches have been developed that explicitly simulate the levels of transcripts or enzymes, and how they regulate denitrification rates. For example, M. Li et al. (2017b) modeled production of enzymes for denitrification during the incubation of hyporheic zone sediments based on an energetic approach. Song et al. (2017) described the regulation of denitrification enzymes with a cybernetic approach. In the latter, reaction networks are optimized with respect to a metabolic goal such as maximizing the microbial growth rate (Ramkrishna and Song, 2019). Other models feature a mechanistic description of the regulatory chain, including the production of transcription factors triggered by signaling molecules, production of messenger RNA (mRNA), and translation into enzymes (Bælum et al., 2013; Koutinas et al., 2011; Störiko et al., 2021a).

Such enzyme-based modeling approaches provide a mechanistic link between transcript and enzyme concentrations, and reaction rates. We use an enzyme-based reactive transport model to assess whether measurements of functional-gene transcripts or enzymes can be used as a proxy for biogeochemical reaction rates. While our approach is generally applicable to other critical processes, we focus on denitrification at the river–groundwater interface. On the basis of this example, we analyze how biogeochemical and hydrological conditions influence the relationship between transcript concentrations and reaction rates.

4.2 METHODS

4.2.1 Model scenarios

We set up three model scenarios that represent different hydrological conditions at the river–groundwater interface (figure 4.1), ranging from steady-state hydrology and biogeochemistry to pronounced diurnal cycles. Biogeochemical and hydrological conditions that differ between the scenarios enable us to evaluate their impact on simulated transcript and enzyme concentrations and reaction rates. In all scenarios, we considered microbial aerobic respiration and denitrification. Both pathways were coupled to the oxidation of dissolved organic carbon (DOC) which was provided via hydrolysis of particulate organic carbon (POC) in the aquifer matrix and via inflow from the river water.

The first scenario simulated constant groundwater discharge (GD), where nitrate-rich water from the aquifer recharged into the river (figure 4.1a), a common situation in agricultural landscapes. In addition, we assumed that the reactive POC concentration was highest near the river and decreased with increasing distance away from the streambed. That is, we imposed a gradient in the electron donor availability that focused the denitrification activity near the river-aquifer interface. The formulation of the gradient is presented in the next section (see equations 4.9 and 4.10).

In the second scenario, we simulated oxic river water continuously entering the aquifer (figure 4.1b), mimicking a bank-filtration (BF) scenario that could be either induced by pumping or by the natural hydraulic gradient of the system. Oxygen concentrations in river water can be subject to strong daily fluctuations, reflecting the interplay between radiation-dependent photosynthesis, aerobic respiration, and gas exchange in the river (Hayashi et al., 2012; Kunz et al., 2017). We considered two subscenarios: In the first, the oxygen concentration in the river remained at a constant level of $8\,\mathrm{mg\,L^{-1}}$ (BFC, constant oxygen), whereas in the second the concentration sinusoidally fluctuated about the mean value, yielding dynamic redox conditions close to the river–groundwater interface (BFP, periodic oxygen).

In a third scenario, denoted bank storage (BS), we considered a flow-reversal,

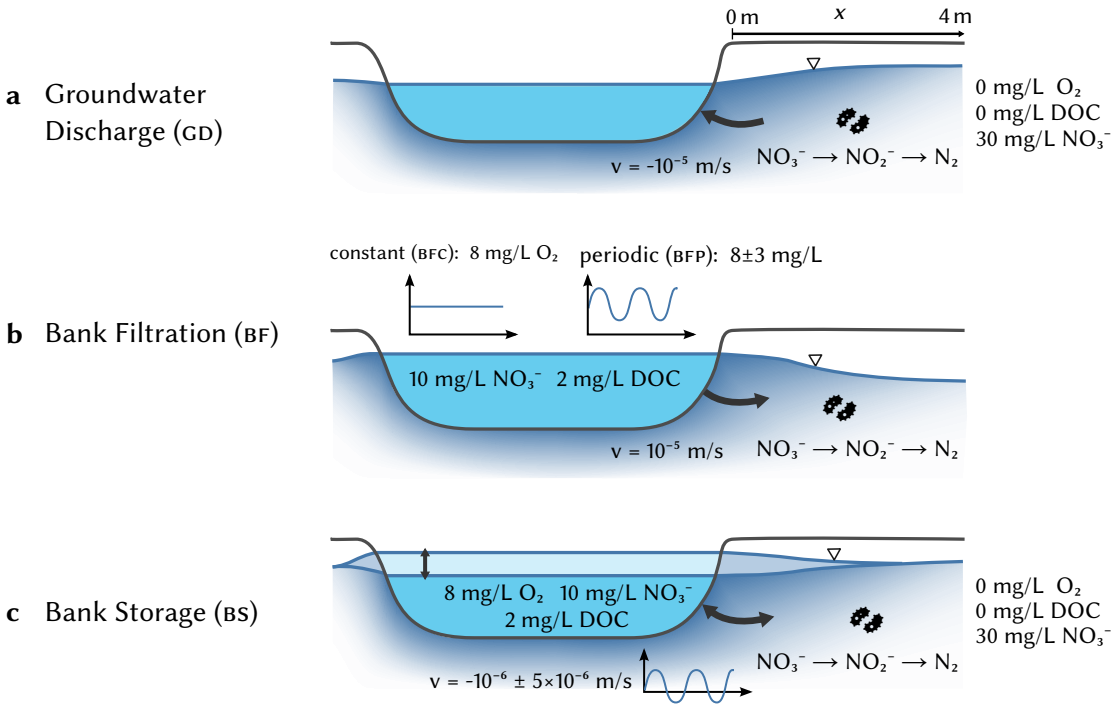


FIGURE 4.1: Schematic of the three simulation scenarios and the corresponding boundary conditions. The thick arrow at the river boundary indicates the groundwater flow direction. The inset plots in panel **b** schematically show the oxygen concentration in the river water used for the fixed concentration boundary condition over time. The inset plot in panel **c** illustrates the advective velocity v over time.

induced by dynamic river-stage fluctuations, reflecting, for instance, hydropeaking (Sawyer et al., 2009) or tidal influences (figure 4.1c). Close to the river–groundwater interface, the flow reversal caused alternating oxic and anoxic conditions.

4.2.2 Governing equations

Advective-dispersive-reactive transport We described transport and reaction of dissolved compounds (nitrate, nitrite, oxygen, DOC) via the one-dimensional (1-D) advection-dispersion-reaction equation. The evolution of compound i's concentration c_i in space (x) and time (t) is thus given by equations (2.11) and (2.12). We further assumed that flow is at quasi-steady state, in which v is uniform in space and reacts instantaneously to changes in boundary conditions. In scenarios GD and BF, the velocity is constant in time, whereas in scenario BS, we approximated v as a sinusoidal function

of time with mean velocity $\bar{v}$ [m s^{-1}], amplitude $\hat{v}$ [m s^{-1}] and frequency f_v [s^{-1}]:

$$v(t) = \bar{v} + \hat{v} \sin(2\pi f_v t) \tag{4.1}$$

In all simulations we neglected transport of bacterial cells because the majority (more than 99 % according to Griebler et al., 2002) of active microorganisms in the subsurface are attached to sediments (H. J. Smith et al., 2018). Transcripts and enzymes were assumed to be confined to the interior of bacterial cells and thus to be immobile.

Microbial reactions We used an enzyme-based model formulation of microbial denitrification (Störiko et al., 2021a) that reflects the biological regulation of reaction rates by simulating concentrations of transcription factors, functional-gene transcripts, and enzymes explicitly. The reaction model describes both aerobic respiration and reduction of nitrate to N_2 via NO_2^- as a reactive intermediate. Denitrification is coupled to the oxidation of organic carbon, formally expressed as succinate, serving as an electron donor and carbon source for the facultative anaerobe *Paracoccus denitrificans*. Herein, we applied the parameters in Störiko et al. (2021a) specific to *P. denitrificans* to simulate denitrification coupled to DOC oxidation (assuming that succinate acts as a generalized form of DOC) to the flow scenarios outlined in figure 4.1. Despite the parameters being specific to a pure-culture batch experiment (Qu et al., 2015), they provide an opportunity with which to probe the thus far poorly-characterized behavior of transcription and enzyme regulation in natural subsurface-transport settings, relevant for biogeochemical laboratory and field investigations. In the following we briefly summarize key model processes and refer the reader to the original publication for more detail.

The catabolic reactions were described by the following stoichiometric equations:

$$7\,NO_3^- + C_4H_6O_4 \xrightarrow{narG} 7\,NO_2^- + 4\,CO_2 + 3\,H_2O \tag{4.2}$$

$$14\,NO_2^- + 14\,H^+ + 3\,C_4H_6O_4 \xrightarrow{nirS} 7\,N_2 + 12\,CO_2 + 16\,H_2O \tag{4.3}$$

$$7\,O_2 + 2\,C_4H_6O_4 \longrightarrow 8\,CO_2 + 6\,H_2O \tag{4.4}$$

Gene expression is controlled by the transcription factors FnrP, sensitive to oxygen levels, NarR, regulated by nitrate and nitrite, and NNR, stimulated in the presence of nitrite and absence of oxygen. Transcription of the *narG* gene, coding for nitrate reductase (NAR), is initiated in the presence of FnrP and NarR, whereas the transcription of *nirS*, coding for nitrite reductase (NIR), requires NNR. The concentrations of transcripts were assumed to be at quasi-steady state with the transcription factor concentrations. The NAR and NIR enzymes are produced in response to *narG* and *nirS* levels and decay following a first-order rate law (Störiko et al., 2021a).

Denitrification rates are a function of the enzyme concentrations, a double Michaelis-Menten term for the limitation of electron donor (DOC) and electron acceptor (nitrate, nitrite) concentrations and an oxygen inhibition term:

$$r_{\mathrm{N}} = k^{j}_{\max} E_j \frac{c_{\mathrm{N}}}{K_{\mathrm{N}} + c_{\mathrm{N}}} \frac{c_{\mathrm{DOC}}}{K_{\mathrm{DOC}} + c_{\mathrm{DOC}}} \frac{I^{j}_{\mathrm{O_2}}}{c_{\mathrm{O_2}} + I^{j}_{\mathrm{O_2}}} \tag{4.5}$$

Here, $k^{j}_{\max}$ [$\mathrm{s^{-1}}$] is the amount of substrate that the enzyme j (NAR or NIR) can maximally turn over per time (also called turnover number), E_j [M] is the concentration of enzyme j that catalyzes the reaction of substrate N (nitrate or nitrite). K_{N} [M] and K_{DOC} [M] are the half-saturation concentrations for nitrate/nitrite and DOC, respectively, and $I^{j}_{\mathrm{O_2}}$ [M] is the oxygen inhibition constant for enzyme j. Aerobic respiration was described by a standard double Michaelis-Menten formulation with the maximum cell-specific respiration rate $v^{\mathrm{O_2}}_{\max}$ [$\mathrm{mol\,cell^{-1}\,s^{-1}}$] and biomass concentration B [$\mathrm{cells\,L^{-1}}$]:

$$r_{\mathrm{O_2}} = v^{\mathrm{O_2}}_{\max} B \frac{c_{\mathrm{O_2}}}{K_{\mathrm{O_2}} + c_{\mathrm{O_2}}} \frac{c_{\mathrm{DOC}}}{K_{\mathrm{DOC}} + c_{\mathrm{DOC}}} \tag{4.6}$$

To predict the dynamics of transcripts and enzymes under conditions similar to those found in natural environments, we modified and complemented the parts of the model that relate to DOC and biomass. Here, the model was expanded to include the release of DOC from POC in the aquifer matrix, and its consumption by both denitrification and aerobic respiration. The latter yielded a DOC consumption dependent on the electron-acceptor consumption rates (defined in equations 4.5 and 4.6) and their stoichiometric coefficients in the metabolic reaction:

$$r^{j}_{\mathrm{DOC}} = \frac{\gamma^{i}_{\mathrm{DOC}}}{\gamma^{j}_{A}} r^{j}_{A} \tag{4.7}$$

and

$$r_{\mathrm{DOC}} = \sum_{j} r^{j}_{\mathrm{DOC}}\,, \tag{4.8}$$

where γ^{j}_{A} and $\gamma^{j}_{\mathrm{DOC}}$ are the stoichiometric coefficients of the electron acceptor and DOC in reaction (j) and r^{j}_{A} is the corresponding electron-acceptor reaction rate. We modeled the release of DOC from the POC-containing aquifer matrix as a first-order mass transfer process (Gu et al., 2007; Kinzelbach et al., 1991; Knights et al., 2017), with the first-order coefficient $k^{\mathrm{DOC}}_{\mathrm{release}}$ [1/s]:

$$r_{\mathrm{release}} = k^{\mathrm{DOC}}_{\mathrm{release}} \left(c^{\mathrm{sat}}_{\mathrm{DOC}} - c_{\mathrm{DOC}}\right) \tag{4.9}$$

The DOC saturation concentration $c_{\mathrm{DOC}}^{\mathrm{sat}}$ [M] depends on the POC content of the sediment, which tends to decrease with distance from the river (Marmonier et al., 1995; Stelzer et al., 2011). Following Knights et al. (2017), we therefore assumed an exponential profile of $c_{\mathrm{DOC}}^{\mathrm{sat}}$:

$$c_{\mathrm{DOC}}^{\mathrm{sat}} = c_{\mathrm{DOC}}^{\mathrm{sat},0} \exp\left(-\frac{x}{l}\right), \tag{4.10}$$

where l [m] is the length scale for the concentration decrease.

In contrast to the original formulation, bacterial growth was parameterized as a function of the oxidation of organic carbon coupled to both oxygen and nitrogen oxide reduction. The synthesis of biomass, represented with the molecular formula $C_{10}H_{18}O_5N_2$, can formally be described by the reaction:

$$3\,C_4H_6O_4 + 2\,NH_4^+ \longrightarrow C_{10}H_{18}O_5N_2 + 2\,CO_2 + 3\,H_2O + 2\,H^+ \tag{4.11}$$

NH_4^+ is assumed to be non-limiting for microbial growth, and is not explicitly simulated. Equation 4.11 was then coupled to the energy-gaining reactions 4.2–4.4 to obtain the overall metabolic reaction. The stoichiometric coefficients in the metabolic reaction depend on the number of catabolic formula reactions that must be completed to generate the energy required for one anabolic formula reaction (and thus produce one mol of biomass). In turn, this number is directly related to the growth yield Y_i [cells $\mathrm{mol_C^{-1}}$], which corresponds to the amount of biomass that is produced per mole of organic carbon consumed. The growth yield relates the growth rate associated to the electron acceptor i to the corresponding DOC-consumption rate:

$$r_{\mathrm{growth}}^{i} = Y_i r_{\mathrm{DOC}}^{i} \tag{4.12}$$

Furthermore, we applied a logistic term to the biomass-growth expression (not to the substrate consumption rates) to limit biomass growth to a set maximum density (e. g., Grösbacher et al., 2018). This is in line with observations that biomass densities in porous media reach a "carrying capacity", even under non-growth-limiting conditions (Ding, 2010; Mellage et al., 2015).

The logistic growth term can be interpreted as a reduction in the maximum growth yield by the occupancy level:

$$Y_i = Y_{\mathrm{max}}^{i}\left(1 - \frac{B}{B_{\mathrm{max}}}\right), \tag{4.13}$$

where Y_{max}^{i} is the maximum growth yield and B_{max} is the carrying capacity. This implies that the growth yield and therefore the stoichiometric coefficients of the metabolic reactions depend on the biomass concentration.

The model also accounts for biomass decay via a first-order term with the decay coefficient k_{dec} [$\mathrm{s^{-1}}$]:

$$r_{\mathrm{decay}} = k_{\mathrm{dec}} B \tag{4.14}$$

TABLE 4.1: Boundary condition parameters used in the simulation.

Symbol	Parameter description	Scenarios	Value	Unit	Note
$c_{O_2}^{in}$	O_2 concentration in the river	BFC, BS	250	µM	a
$\overline{c_{O_2}}$	mean O_2 concentration in the river	BFP	250	µM	a
$\widehat{c_{O_2}}$	amplitude of oxygen fluctuations	BFP	94	µM	b
$c_{NO_3^-}^{river}$	NO_3^- concentration in the river	BFC, BFP, BS	161	µM	c
$c_{NO_3^-}^{GW}$	NO_3^- concentration in groundwater	GD, BS	484	µM	c
c_{DOC}^{river}	DOC concentration in the river	BFC, BFP, BS	167	µM	d

[a] Y. Liu et al. (2017) [b] Kunz et al. (2017) [c] Gu et al. (2007) and Y. Liu et al. (2017) [d] Bol et al. (2015) and Hayashi et al. (2012)

This leads to the build-up of dead biomass, which, in turn, decays in a first-order process with constant k_{min}, releasing DOC via mineralization.

Boundary conditions Fixed concentration (Dirichlet) boundary conditions were applied at the river and groundwater inflow boundaries. The river water was assumed to be saturated with respect to oxygen, and contained 10 mg L^{-1} of nitrate and 2 mg L^{-1} of DOC. These concentrations correspond to anthropogenically influenced but not excessively nutrient-enriched rivers. The inflowing groundwater was assumed to be anoxic but rich in nitrate (30 mg L^{-1}) and depleted in DOC. In scenario BFP, oxygen concentrations in the river were described by a sinusoidal function with amplitude $\widehat{c_{O_2}}$ [M], frequency $f_{O_2} = 1\ d^{-1}$ and mean value $\overline{c_{O_2}}$ [M]:

$$c_{O_2}^{in}(t) = \widehat{c_{O_2}} \sin\left(2\pi f_{O_2} t\right) + \overline{c_{O_2}} \tag{4.15}$$

All other concentrations at the inflow boundary were constant over time, with values given in table 4.1. At the outflow boundary, we assumed zero dispersive flux.

4.2.3 Simulation parameters

Parameters related to transcript and enzyme concentrations, denitrification and aerobic respiration were obtained from our previous study (Störiko et al., 2021a) in which we calibrated the enzyme-based model with the laboratory data of Qu et al. (2015). In the simulations presented here, the median values of the parameter distributions in Störiko et al. (2021a) were imposed (table 4.2). Values of new parameters, that is, those that were not included in the previous model (transport and DOC-related parameters) were chosen based on literature values.

TABLE 4.2: Reaction and transport parameters used in the simulation.

Symbol	Parameter description	Value	Unit	Note
v	linear velocity (GD, BF)	10^{-5}	$m\,s^{-1}$	a
$\bar{v}$	mean velocity (BS)	-10^{-6}	$m\,s^{-1}$	b
$\hat{v}$	velocity amplitude (BS)	10^{-5}	$m\,s^{-1}$	b
f_v	velocity frequency (BS)	1	d^{-1}	c
α_L	longitudinal dispersivity	0.1	m	d
D_e	effective diffusion coefficient	3×10^{8}	$m^2\,s^{-1}$	e
$k_{max}^{NO_3^-}$	NAR turnover number	4.4×10^{4}	s^{-1}	f
$k_{max}^{NO_2^-}$	NIR turnover number	2.9×10^{2}	s^{-1}	f
$K_{NO_3^-}$	NO_3^- half-saturation constant	5	µM	g
$K_{NO_2^-}$	NO_2^- half-saturation constant	5	µM	g
K_{DOC}	DOC half-saturation constant	40	$\mu mol_c\,L^{-1}$	h
$I_{O_2}^{NAR}$	O_2 inhibition constant for NAR	1	µM	f
$I_{O_2}^{NIR}$	O_2 inhibition constant for NIR	340	nM	f
$v_{max}^{O_2}$	max. cell-specific O_2 oxidation rate	6.4×10^{-19}	$mol\,cell^{-1}\,s^{-1}$	f
K_{O_2}	O_2 half-saturation constant	31	µM	f
$k_{release}^{DOC}$	DOC release rate constant	0.2	d^{-1}	i
$c_{DOC}^{sat,0}$	max. DOC saturation concentration	20.8	$mmol_c\,L^{-1}$	j
l	length scale for the decrease of the sediment POC content	0.2	m	k
$Y_{max}^{NO_3^-}$	maximum growth yield with NO_3^-	2.6×10^{13}	$cells\,mol_C^{-1}$	l
$Y_{max}^{NO_2^-}$	maximum growth yield with NO_2^-	1.6×10^{13}	$cells\,mol_C^{-1}$	l
$Y_{max}^{O_2}$	maximum growth yield with O_2	7.7×10^{13}	$cells\,mol_C^{-1}$	f, m
B_{max}	carrying capacity	3.3×10^{11}	$cells\,L^{-1}$	n
k_{dec}	biomass decay constant	10^{-7}	s^{-1}	o

[a] See Bertin and Bourg (1994) for bank filtration and C. D. Kennedy et al. (2009) for groundwater exfiltration. [b] Gerecht et al. (2011) and Y. Liu et al. (2017) [c] Diurnal cycles. [d] Gelhar et al. (1992) [e] Based on the approximation $D_e = D\phi$ where $D = 10^{-9}\,m^2\,s^{-1}$ is the molecular diffusion coefficient and $\phi = 0.3$ is porosity. [f] Median of the parameters in Störiko et al. (2021a). [g] Hassan et al. (2016) [h] Fixed to a value within reported ranges (Kinzelbach et al., 1991; Sanz-Prat et al., 2016). [i] Fixed to a value within reported ranges (Gu et al., 2007; Kinzelbach et al., 1991; Sanz-Prat et al., 2016; Sawyer, 2015). [j] Fixed to a value within reported ranges (Gu et al., 2007; Kinzelbach et al., 1991; Sawyer, 2015). [k] Knights et al. (2017) [l] Fixed to a value within reported ranges (Hassan et al., 2014, 2016). [m] Value corrected for the incorporation of organic carbon into biomass, which was not considered in Störiko et al. (2021a). [n] Fixed to a value within reported ranges (Ding, 2010). [o] Fixed to a value within reported ranges (Ding, 2010; Kinzelbach et al., 1991).

4.2.4 Numerical methods

We used the cell-centered finite volume method to discretize the reactive-transport equation 2.11 in space, applying a first-order upwind scheme for advection. The domain had a total length of 4 m and was divided into 200 cells with a uniform spacing of 2 cm. The resulting system of ordinary differential equations (ODEs) was solved with the backwards differentiation formula (BDF) as implemented in the CVODES solver in the SUNDIALS library (Hindmarsh et al., 2005). All code was written in Python 3.8, and the package Sunode (Seyboldt, 2021) that wraps CVODES was used for solving the ODEs. The simulations were run until reaching steady state (in the scenarios with constant boundary conditions) or dynamic steady state, that is, self repeating time cycles in the scenarios with periodic boundary conditions.

4.3 RESULTS AND DISCUSSION

4.3.1 Zonation of redox species and denitrifying bacteria

The three model scenarios result in distinct spatial distributions of N species, transcripts, enzymes, biomass, oxygen and DOC (figure 4.2, rows a–f). In the following we present and discuss the predicted steady-state concentrations scenario-wise in detail: GD (figure 4.2, left column), BF (figure 4.2, center column) and BS (figure 4.2, right column).

Scenario GD: groundwater discharge Nitrate enters the domain with in-flowing groundwater, and remains at high concentrations (i. e., close to the inflow value) over the first 2 m of the domain, where the aquifer matrix contains only little POC (electron donor limitation). At about 1.5 m from the river, nitrate begins to drop and is completely depleted at a distance of 0.25 m from the sediment-river interface. Nitrite concentrations increase, mirroring the drop in nitrate, until reaching a peak value of 340 µmol L^{-1} at 0.3 m and then decrease towards the river.

Our model-predicted nitrite concentrations are higher than typically observed in natural sediments. Profiles of pore-water nitrite in several studies indicate that the concentrations are usually below 30 µM (Akbarzadeh et al., 2018; Harvey et al., 2013; Stief et al., 2002). The parameter set used here is based on laboratory batch experiments with a single strain where strong nitrite accumulation was observed (Störiko et al., 2021a). Thus, while the high model-derived nitrite concentrations may be specific to the strain used in the experiments, we assume that the spatial trends in nitrogen species consumption and production are likely generalizable.

The concentration of DOC drops from 280 µmol_c L^{-1} at the sediment-river interface to below 40 nmol_c L^{-1} within 30 cm, driven by the prescribed exponentially decreasing content of POC in the sediment (the only source of DOC) away from the river boundary.

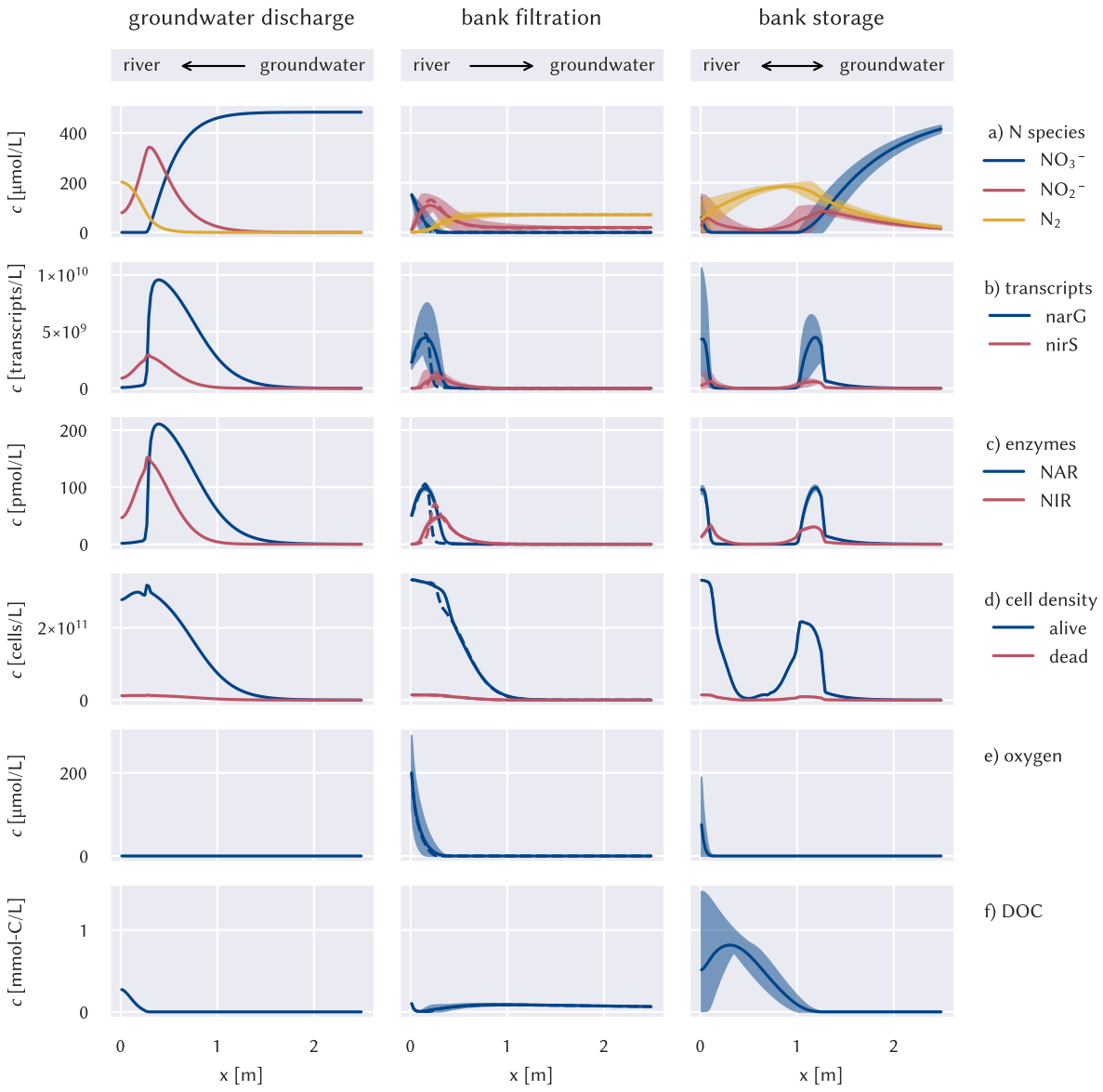


FIGURE 4.2: Spatial distributions of nitrogen compounds (a), transcript (b) and enzyme (c) concentrations, biomass (d), oxygen (e), and DOC (f) in the different scenarios. The steady-state solution in scenario BFC is indicated by a dashed line. For the periodic solution in scenarios BFP and BS, the minimum and maximum values over time are indicated by the shaded area, the mean value is plotted as a solid line. Concentrations between 2.5 m and the groundwater-side domain boundary at 4 m are omitted because they are almost constant.

The zones of nitrate and nitrite consumption coincide with elevated absolute concentrations of *narG* and *nirS* transcripts (that is, in units of transcripts L^{-1}) and NAR/NIR enzymes (figure 4.2c, left column). In contrast, cell-specific *narG* transcript and NAR enzyme concentrations are high in the DOC-limited section of the domain, despite the absence of denitrification (figure C.1). Nitrate triggers transcription but the low availability of the electron donor (DOC) results in low biomass concentrations strongly limiting denitrification. High biomass concentrations are only reached close to the river, where denitrification activity is the highest.

Scenario BF: bank filtration The center column in figure 4.2 shows the dynamics of the two bank-filtration scenarios with periodic (BFP) and constant (BFC) oxygen concentrations in the inflow. In the BFP scenario, concentrations do not reach a steady state but concentration time series converge to repeating diurnal cycles.

The model predicts a zonation of the redox processes starting with aerobic respiration at the inflow boundary, where oxygen-rich river water infiltrates. Nitrate, present in the incoming water, is subsequently reduced to nitrite and N_2. The fluctuating oxygen concentrations in the river (inflow) in scenario BFP leads to a periodic shift in the location of the denitrification zone, which oscillates back and forth over 0.1 m about 0.2 m. At a given location, nitrate and nitrite concentrations fluctuate considerably over the course of the day. For example, nitrate concentrations at 0.2 m vary between 60 µmol L^{-1} and total depletion. Nitrite is reduced to low, but non-zero "residual" concentrations (20 µM). The low concentration front subsequently penetrates deep into the aquifer.

Biomass concentrations are very stable over time in the scenario with a fluctuating inflow oxygen concentration and hardly differ from the steady-state scenario. Cell doubling times in the simulations range from a few hours to several days, which is in accordance with literature values (Mailloux and Fuller, 2003). Similarly, biomass decay is slow (with a half-life of about 80 d, see table 4.2), such that the biomass does not respond to daily cycles of substrate availability. Biomass concentrations are highest at the river inflow boundary where neither oxygen nor DOC are limiting and cell densities reach the maximum capacity B_{max}. At locations where oxygen and nitrate are consumed, the remaining low nitrite concentrations can only sustain the survival of a small biomass pool (starting at 1.3 m from the river boundary), which in turn reduces the denitrification rate to values close to zero.

Transcripts of the *narG* gene are abundant in the region where nitrate is available and *nirS* transcripts co-occur with nitrite. In the scenario with dynamic boundary conditions, the transcript concentrations of denitrification genes exhibit a distinct diurnal cycle with an amplitude of up to 70 % (*narG*) and 100 % (*nirS*) of the mean value, in some parts of the domain. Concentrations of NAR and NIR enzymes follow the patterns of *narG* and *nirS* transcripts, but are much more dampened, with amplitudes

that are one order of magnitude smaller than those of the corresponding transcripts. This difference stems from the different time scales of production and decay of transcript and enzymes. While transcripts usually decay within a few minutes (Bernstein et al., 2002; Härtig and Zumft, 1999) and are therefore assumed to be at quasi-steady state in our simulations, enzyme half-lives range on the order of several hours to days (Maier et al., 2011).

Because of the high DOC concentration (0.1 mM) imposed at the river boundary, the river water serves as a DOC source. The DOC concentration, however, drops sharply in the aquifer due to the high microbial electron-donor demand, driven by the presence of oxygen and nitrate. Outside of the zone of denitrification, the DOC concentration rises towards the groundwater boundary, driven by the hydrolysis of POC, reaching a maximum at about 1 m. The decreasing POC content away from the river yields a final gradual decline in DOC approaching the groundwater boundary.

Scenario BS: bank storage In the bank storage scenario, the alternating inflow of nitrate from the aquifer and from the river leads to the formation of two distinct zones of denitrification (figure 4.2, right column). The first one is located directly at the the river-aquifer interface. It is active only at the times when flow is from the river into the aquifer, hence supplying nitrate. We estimated the maximum penetration depth of the river water by integrating the positive part of the velocity function over one period. Via advection only, the water penetrates 0.23 m into the aquifer. Oxygen and nitrate reach that point only at very low concentrations because they are rapidly depleted after entering the aquifer.

The second zone of denitrification at about 1.1 m is fed by nitrate from the incoming groundwater. At the aquifer boundary, denitrification is mainly limited by carbon availability, such that nitrate concentrations remain at high values until the distance to the river is apprximately 1.5 m, after which they sharply decrease. Due to the flow reversal, this denitrification zone shifts between 1 m and 1.35 m over time. The response of concentrations to the dynamic flow is generally similar to scenario BFP where the dynamics are caused by fluctuating oxygen concentrations. Both solute concentrations and MRNA strongly fluctuate over time while enzyme concentrations and biomass are stable because of their longer time scales of production and decay.

Compared to the other two scenarios, the DOC concentration in the bank storage scenario is high in the 1.2 m adjacent to the river. On average, the magnitude of the advective velocity and therefore the influx of electron acceptors (nitrate and oxygen) is smaller in this scenario. This limits the consumption of DOC and leads to its overall high concentration.

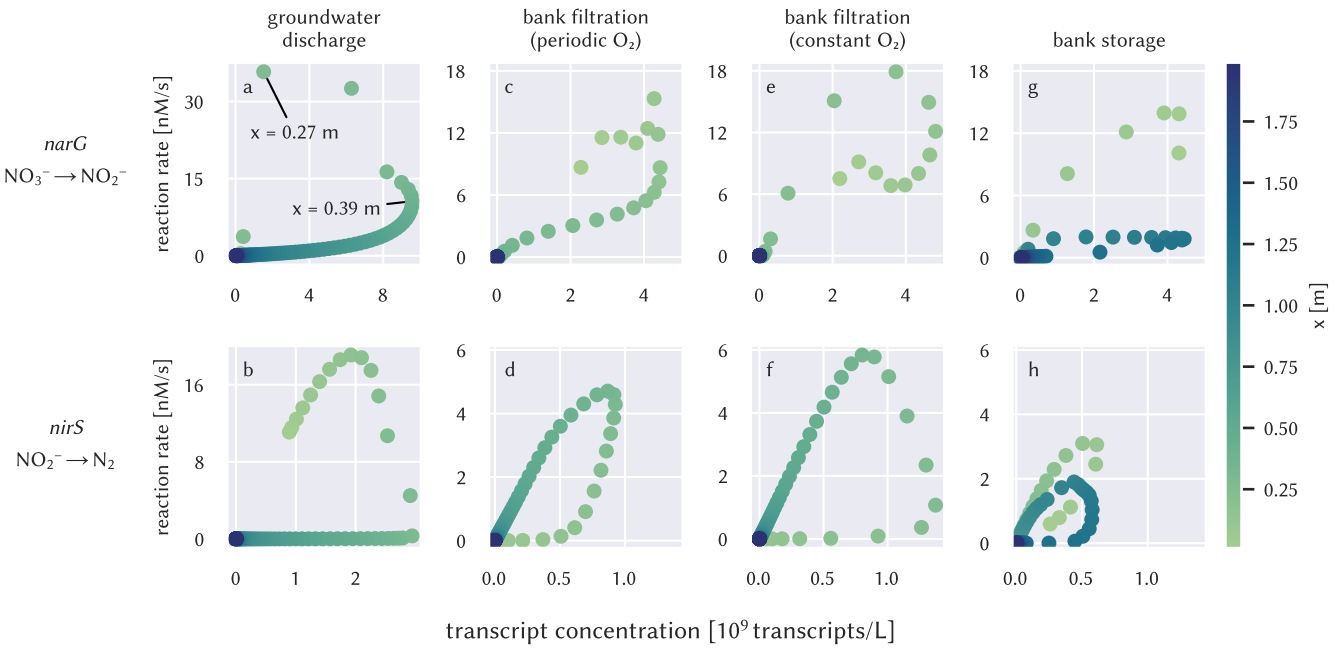


FIGURE 4.3: Relationships between the concentrations of functional-gene transcripts *narG* (upper row) and *nirS* (lower row) with the denitrification rates in the different scenarios. In the scenarios where concentrations do not reach constant steady state values but exhibit repeating diurnal cycles, daily averages of rates and concentrations are shown. The color indicates the spatial coordinate with dark blue corresponding to the groundwater inflow boundary and light green corresponding to the river boundary. Note that the axis scales are different for the groundwater discharge scenario.

4.3.2 Relationship between transcripts/enzymes and reaction rates

Based on our simulation results, we computed denitrification rates to explore how transcript and enzyme concentrations relate to the denitrification activity in the different scenarios (see figure 4.3 for transcripts and figure C.4 for enzymes).

Scenario GD: groundwater discharge In the groundwater-discharge scenario, the system reaches a steady state where the enzyme concentrations are proportional to transcript concentrations. Therefore, it is sufficient to analyze the relationship between reaction rates and transcripts *or* enzymes. For simplicity we compare rates to transcripts in figure 4.3a and 4.3b.

The relationships between rates and transcripts are non-linear and the correlation is positive in some parts of the domain, but negative or zero in other parts. At the

groundwater-inflow boundary (dark blue colors), both *narG* transcript concentrations and NO_3^- reduction rates are close to zero and increase towards the river (lighter colors). However, when the rates reach 10 nmol L^{-1} s^{-1} at 0.39 m, the trend reverses, that is, where transcript levels decrease reaction rates increase and reach their maximum at 0.27 m. At the points closest to the river boundary, both the nitrate reduction rate and *narG* transcript levels return to zero, closing the hysteresis loop.

The concentrations of *nirS* transcripts rise between 1 m and 0.3 m (figure 4.3b). However, their increase does not correspond to an increase in reaction rates, suggesting that under certain conditions, transcript concentrations and reaction rates may be completely decoupled. One may intuitively expect that increasing reaction rates would be accompanied by increasing transcript concentrations. However, the rise of reaction rates between 0.3 m and 0.17 m is concomitant with the opposite, a decrease in transcript concentrations. A positive correlation between *nirS* transcript concentrations and reaction rates is only observed in the 15 cm closest to the river. The strong non-linearity of the transcript-rate relationships (and partly negative correlations) can be explained by the limited availability of DOC over most of the domain (which in this scenario originates from the river and hydrolysis of POC). The latter limits denitrification, whereas transcript production is still triggered by the presence of nitrate and nitrite, irrespective of electron-donor availability.

Scenario BF: bank filtration Figure 4.4a and 4.4c show the relationship between transcript concentrations and denitrification rates for scenario BFP (bank filtration with a fluctuating oxygen inflow concentration). Reaction rates and transcript concentrations (and, to a lesser extent, also enzyme concentrations) both fluctuate over the course of the day, but the signals have a phase shift. This leads to a hysteresis in the relationship between transcript concentrations and reaction rates, with a different hysteretic pattern at different locations. Overall, transcript concentrations and denitrification rates do not show a clear (linear) relationship. These results suggest that it may not be possible to infer the denitrification activity at a given time and location from a single determination of the transcript concentration.

The relationship between enzyme concentrations and denitrification rates (figure 4.4b, 4.4d) is also highly non-linear and location-specific. However, it exhibits less pronounced hysteresis loops because, in contrast to transcripts, the characteristic times for enzyme production and decay are longer than the time scale of the fluctuations. As a consequence, in dynamic steady state with diurnal cycles, the enzyme concentrations remain almost constant throughout the day, whereas the reaction rates fluctuate in response to the periodic concentration changes of aqueous substrates. Thus enzyme distributions could, under the right conditions, be used as proxies for delineating the average denitrification activity.

For the mitigation of nitrate contamination in groundwater daily averages of reac-

tion rates are of greater interest than their diurnal fluctuations. To investigate whether repeated transcript measurements could be used as indicators of denitrification activity, we compare the daily averages of the denitrification rates and the transcript concentrations in Figures 4.3c and 4.3d. As can be seen, upon averaging more distinct positive correlations emerge, although they are still non-unique, particularly in the case of *nirS* transcripts, where the same transcript concentration can be associated with rates that differ by more than one order of magnitude. Different combinations of nitrite, oxygen and DOC concentrations can lead to the same transcript concentration, while the factors describing substrate limitation and oxygen inhibition affecting denitrification rates differ. The relationship looks very similar for transcripts and enzymes because daily averages of transcript concentrations are almost proportional to enzyme concentrations (see figure C.2).

The relationship between steady-state transcript concentrations and denitrification rates for BFC (bank filtration with constant oxygen input; figure 4.3e, 4.3f) slightly differs from the BFP scenario (figure 4.3c, 4.3d) but essentially mirrors the BFP characteristic features. For example, both bank filtration scenarios yield a positive, but non-unique, relationship of narG transcripts with the rates, whereas *nirS* transcripts exhibit a strong hysteretic behavior. It is to be expected that the relationships are generally similar for the steady-state solution and daily averages of the periodic solution as the simulated concentration profiles are nearly the same in both cases (figure 4.2, center), but non-linearity in the rate laws can lead to the observed differences.

Scenario BS: bank storage Similar to the BFP scenario, the periodic reversal of flow in the BS scenario results in complex relationships between the transcript or enzyme concentrations and the denitrification rates (figure C.3). However, in contrast to BFP, daily averages of transcript concentrations and reaction rates (figure 4.3g and 4.3h) show two clearly distinct patterns, corresponding to the two denitrification zones, and resembling to some extent the patterns of the pure groundwater-discharge and pure bank-filtration scenarios. In both zones, the relationships are non-linear and non-unique, analogous to all other scenarios. This is most evident for the *narG* transcripts (shown in figure 4.3d).

Comparison of the transcript-rate relationship between scenarios We compared the relationship between transcript concentrations and reaction rates between the different scenarios to evaluate how biogeochemical and hydrological conditions affect the relationship according to the three scenarios. Transcript concentrations and reactions rates differed between the scenarios. For example, we can observe that *narG* transcript concentrations are highest in the groundwater discharge scenario. This can be explained by the higher nitrate concentrations in the groundwater in comparison to the river water in the bank filtration scenario. In the bank storage scenario, transcript

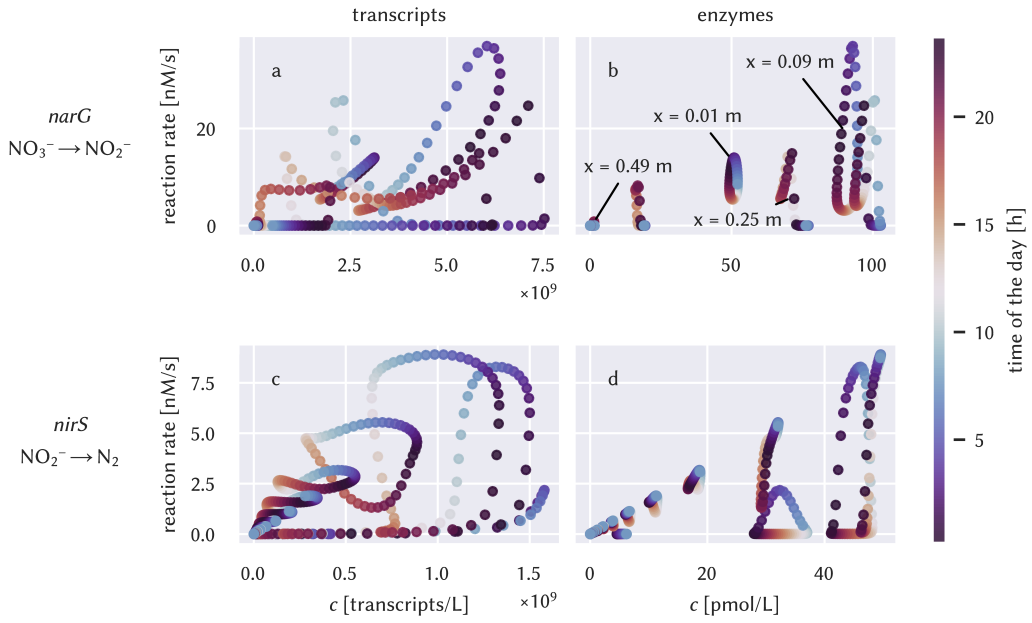


FIGURE 4.4: Relationships of transcript (left column) and enzyme (right column) concentrations with denitrification rates for scenario BFP (river water with fluctuating oxygen concentrations infiltrating groundwater). Colors indicate the time point within the diurnal cycle. Every point x in space shows a distinct pattern (with one "loop" corresponding to one point), and many of them are non-linear and hysteretic in time.

concentrations are lower because microorganisms receive nitrate input only half of the time. However, such absolute differences (e. g., in the slope or exact shape of the curve, or in magnitudes of reaction rates and transcript concentrations) depend on the choice of parameter values. Thus, model calibration with field data is required as a next step to generalize model-based conclusions in line whit empirical evidence.

Two patterns are, though, common to all scenarios, and arise rather from the model structure than from individual parameter values. First, we observe that in both of the dynamic scenarios, time-shifts between the dynamics of transcripts and reactions result in a complex relationship between transcript concentrations and reaction rates. Averaging transcript concentrations over one day significantly simplifies the relationship in both cases (compare figures 4.3 and 4.4). Even though the temporal uncoupling also depends on the values of reaction parameters that dictate the response times of transcripts and enzymes, the uncertainty of these parameters is relatively small in comparison to other parameters such as the half-saturation constants, where literature values range over several orders of magnitude (García-Ruiz et al., 1998). Typical time scales for the response time of transcripts and enzymes are minutes and hours, respectively (Bernstein et al., 2002; Maier et al., 2011). The second common pattern is that the relationship between (average) transcript concentrations and reaction rates is strongly non-linear and non-unique. This is the case because transcription also occurs under non-ideal conditions, that is, when reaction rates are limited by substrate availability or oxygen inhibition. Different combinations of limiting and inhibiting conditions at different locations produce the spatial hysteresis patterns.

4.3.3 Unraveling the relationship between transcript concentrations and reaction rates

The relationships between transcript concentrations and denitrification reaction rates, presented in the previous section, clearly show that transcript concentrations are not a reliable predictor of denitrification rates, even in cases where these are proportional to enzyme concentrations. Deviations from an expected linear relationship arise because denitrification rates are not only limited by enzyme concentrations (which, in turn, are ultimately determined by the nitrogen species triggering transcription), but also by substrate availability (in our study DOC and nitrogen species) and oxygen inhibition. In the following, we refer to the denitrification rates under in-situ conditions that are limited by substrate availability and oxygen inhibition as *effective* rates. In the model, we can eliminate these limitations by dividing the rate by the corresponding Michaelis-Menten or inhibition term. This then yields the *potential* denitrification rates. When these potential rates are compared to the transcript concentrations, clear positive relationships emerge (figure 4.5).

In the groundwater-discharge scenario (figure 4.5a, 4.5c), removing the DOC limitation yields a nearly linear relationship, showing that carbon limitation is the most

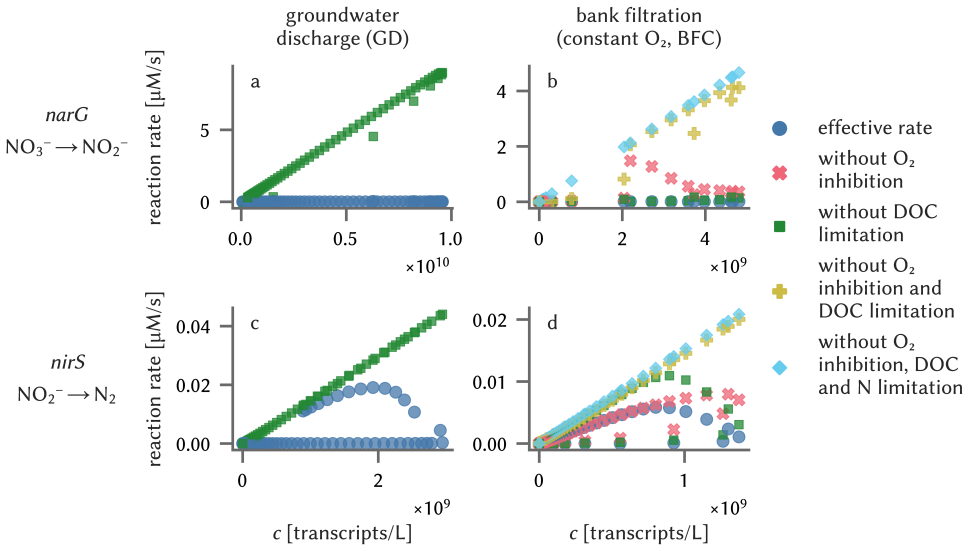


FIGURE 4.5: Relationship between the concentrations of functional-gene transcripts *narG* (upper row) and *nirS* (lower row) with potential denitrification rates after removing the effects of O_2 inhibition, DOC limitation, nitrogen substrate limitation, or combinations thereof. (Note: Scenarios where concentrations do not reach a steady state are omitted because correcting for the rate limitations based on time-averaged concentrations is not a valid approach).

important rate-limiting factor in this scenario. The remaining non-linearity of *narG*-transcripts at low reaction rates can be explained by the presence of nitrite near the river boundary, triggering *narG*-transcription even though nitrate levels and thus nitrate-removal rates are low.

The current model assumes that transcription of the denitrification genes is independent of DOC availability. While this approach is consistent with the current understanding of the targeted regulation of denitrification genes by nitrogen species and oxygen (Gaimster et al., 2018), our model formulation neglects unspecific mechanisms of gene regulation that act to shut down microbial metabolism at low carbon availability, thereby affecting denitrification genes. Accounting for transcription down-regulation of the denitrification genes under carbon limitation in our model formulation would likely yield relationships between transcripts and reaction rates closer to the potential rates without DOC limitation (figure 4.5). Non-linear effects of DOC limitation on the reaction rates would persist. However, the absolute deviation from a linear relationship would be negligible when transcript concentrations and, therefore, potential rates are close to zero. Under extreme electron-donor limitation, our model predicts very low absolute transcript concentrations even without explicitly accounting for DOC-controlled down-regulation of transcription because DOC-limitation restricts microbial growth, leading to low biomass and, thereby, low transcript concentrations. However, if there is evidence for a large abundance of inactive denitrifiers, the model might need to distinguish between the active and an inactive microbial pool, in which transcription is shut off (see, e. g., Chavez Rodriguez et al., 2020).

In the case of bank filtration with a constant oxygen concentration (figure 4.5b, 4.5d), accounting for the DOC limitation term alone does not remove the non-linearity because oxygen inhibition also exerts an important control on denitrification. Eliminating both DOC limitation and oxygen inhibition leads to an approximately linear relationship between transcripts and potential rates. However, the potential rates are orders of magnitude larger than the effective (substrate-limited and inhibited) reaction rates.

In the scenarios in which concentrations undergo periodic fluctuations in time (BFP and BS), applying the correction terms would only be permissible for the time-variable rates and concentrations, but not for the averages. This is so because the correction terms are nonlinear and the concentrations involved are strongly correlated in time. Under such conditions, the product of their time-averaged values is not the same as the time-average of their product. Hence, applying corrections to the time-averaged rates to obtain a more unique relationship of the time-averaged transcript concentrations is not permissible. Similar effects have been described for spatial correlations of degrader communities and substrate concentrations in carbon cycling models. Chakrawal et al. (2020) used scale-transition theory to analyze how spatial correlations among state

variables or between state variables and kinetic parameters affect upscaled reaction rates. In theory, the same method could be applied to obtain time-averaged rates based on average concentrations. However, it requires knowing the covariance terms of substrate and enzyme concentrations in time, which is not possible in practice because highly time-resolved measurements of transcript or substrate (DOC, nitrogen species) concentrations in groundwater are not available in the first place.

4.3.4 Implications for the design of field sampling and measurements

Our simulations show that transcripts of denitrification genes respond to short-term (diurnal) fluctuations of electron-acceptor concentrations, yielding highly temporally variable transcript concentrations at the river–groundwater interface. In such a dynamic system, analyses based on transcripts of functional genes would strongly depend on the time point of sampling. Transcripts exhibiting a low, even undetectable, abundance at a given time, may be present at much higher concentrations at other times of the day, and vice versa. Hence, interpretations on overall system behavior based on transcript concentrations obtained from sporadic sampling events, could be misleading in highly dynamic biogeochemical environments such as those found at the river–groundwater interface.

Based on our modeling results we simulated transcript measurements over time and space to illustrate, how different sampling frequencies and times can affect the outcome captured by measurement campaigns. Figure 4.6a shows time series of *nirS* transcript concentrations in the bank filtration scenario with fluctuating oxygen concentrations (BFP) at a distance of 0.17 m from the river, sampled at different frequencies (weekly samples, daily samples, three and ten samples per day). We added a small random time perturbation to the sampling times to represent a realistic situation.

The high sampling frequency of ten samples per day captures the diurnal signal quite well. Taking three samples per day also captures the dynamic behavior of the system, albeit with less accuracy, with many of the peaks cut off and a more irregular signal than it actually is. Daily and weekly sampling creates apparent patterns in the data that are not linked to any real process but that are due to sampling the diurnal signal at slightly different times each day or week.

Figure 4.6b shows a spatial profile of simulated transcript measurements, taken at two different times of the day. While the general shape of the two profiles is similar, the location of the peak is shifted by about 10 cm, and between 5 cm and 20 cm the concentrations between the two time points differ by up to two orders of magnitude. This example emphasizes the need to consider the relevant time scale of variation for transcripts when planning measurement campaigns. Simple tools like redox- or oxygen-sensitive probes could provide a first approximation of what the relevant

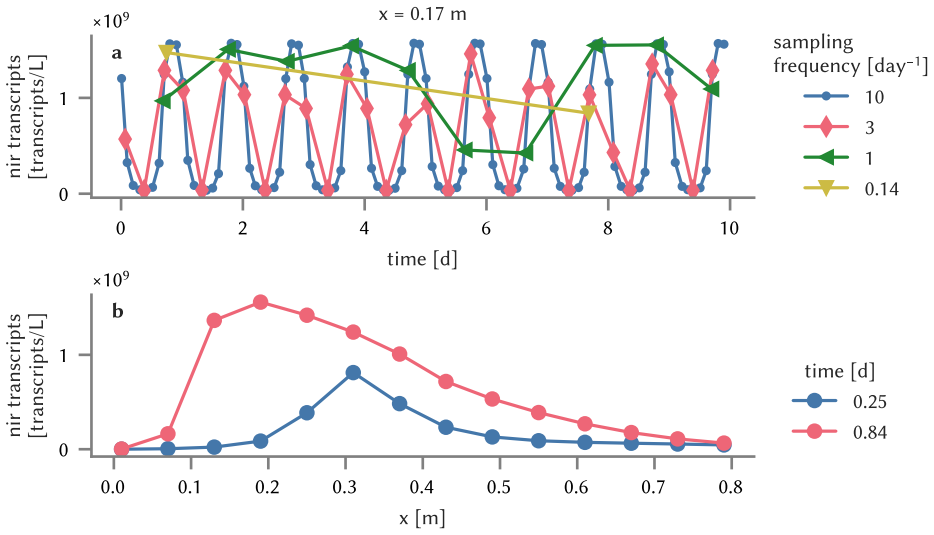


FIGURE 4.6: Simulated measurements of the *nirS* transcript concentrations. **a** Effect of different sampling frequencies on a time series measured at a fixed location ($x = 0.17$ m). **b** Dependence of a spatial profile on the time of the measurement.

time scale for transcript dynamics is.

In contrast to transcripts, the concentrations of functional enzymes and functional biomass (which can be estimated by functional-gene concentrations) are much more dampened and hardly respond to diurnal fluctuations of electron-acceptor availability because of their larger time scales of production and decay. As a consequence, DNA-based methods such as the quantification of functional genes or metagenomics can provide information less dependent on short-term fluctuations of electron acceptor or electron-donor concentrations.

However, a DNA-based approach, analogous to an enzyme-based approach, is subject to other uncertainties related to DNA's persistence and presence outside of active organisms (relic DNA) that can distort the characterization of the microbial community (Carini et al., 2016; Lennon et al., 2018; Nielsen et al., 2007), an effect not considered in this study. Different approaches to filter out the signals from relic DNA (viability PCR, e. g., Carini et al., 2016; Fittipaldi et al., 2012) and inactive microbes (BONCAT-FACS, selecting for translationally active cells, e. g., Couradeau et al., 2019) have been developed in the past years but are not yet applied routinely.

The unresponsiveness of enzyme concentrations and biomass in a system with short-term dynamics also implies that incorporating their time-variability into a biogeochemical model is not necessary and they can be assumed to be constant

in time (i. e., via a biomass-implicit rate formulation). However, spatial variations should be considered, for example by using spatially variable rate coefficients. In systems in which the concentrations of electron acceptors vary over larger time scales (seasonal dynamics, flood events with effects of several days), the temporal variability of functional biomass and particularly enzyme concentrations might also play a role. Measurements of functional enzymes would also provide a more robust picture of microbial activity compared to functional-gene transcripts, because they are less affected by short-term fluctuations. Unfortunately, the quantification of functional enzymes (as opposed to transcripts) is not yet an established measurement technique for environmental samples, even though some pioneering studies have been done (e. g., M. Li et al., 2017a).

Daily averaged transcript concentrations, however, are proportional to enzyme concentrations (for the scenarios investigated here), thus implying that several transcript measurements in time could replace the more difficult to measure enzymes in groundwater systems. Because only averages are required, mixing samples from several time points prior to RNA extraction could also help to reduce transcript measurement efforts.

The main challenge, however, lies in obtaining samples from the same location at several time points, as sampling for gene quantification is destructive. When reactions are much slower than advective transport (low Damköhler number), several samples along a flowpath at a single time point (representing water parcels infiltrated at different times) could replace samples from the same location at several time points. In our simulations, however, reactions deplete substrates within a few centimeters. Water parcels with a time difference of 12 h are separated by a distance of 0.43 m, such that averaging over the locations does not provide a replacement for the temporal average at a single location. Therefore, samples should be taken at adjacent locations, corresponding to the same distance along a flow path (heterogeneity would make that more difficult). The latter illustrates the difficulty of acquiring time-resolved field measurements of transcripts. However, column experiments in the laboratory that simulate conditions in the field (see, e. g., Y. Liu et al., 2017) provide a potential alternative, and would be a useful addition to capture higher-resolution dynamics.

Even after time-averaging, transcript or enzyme concentrations are not reliable predictors of reactions rates. The relationships in the simulated scenarios are non-unique and non-linear. Our analysis reveals that enzyme concentrations can be interpreted as a proxy for *potential* rates, which are hypothetical rates in the absence of specific limitations, such as substrate limitation and oxygen inhibition. These limitations reduce the potential rates towards the *effective* (in-situ) reaction rates.

Based on these findings we argue that an approach to predict denitrification rates directly from transcript or enzyme data would need to account for this, necessitating the following steps. As a first step, the relationship between transcript concentrations

and potential reaction rates needs to be determined. This could be achieved with lab incubations under non-limiting conditions. A caveat here is that under non-limiting conditions, a different part of the microbial community with a different physiology might be more active than under in-situ conditions (Hazard et al., 2021), modifying the relationship. In a system at steady state the relationship between transcript concentrations and potential reaction rates should ideally be linear. Measured transcript concentrations can subsequently serve as a predictor of potential rates which then need to be amended by rate limiting factors like substrate limitation to obtain the effective reaction rates. This correction step does not only require measurements of the involved solute concentrations, but also estimates of parameters describing rate-limiting factors of reaction kinetics (half-saturation and inhibition constants). Such parameter values are often not well known and reported values typically range over several orders of magnitude (see, e. g., García-Ruiz et al., 1998). Therefore, additional experiments to determine specific parameters of the studied system would be necessary.

A powerful integrative approach would be to use a process-based reactive-transport model to predict reaction rates, making use of molecular-biological data to determine model parameters. One advantage is that once a process-based model is calibrated it can deliver reaction rates at time-points and locations where no data are available. We therefore suggest the following strategy combining molecular-biological data, biogeochemical measurements and modeling to determine denitrification rates.

1 *Measure functional enzymes, genes or transcripts to determine temporally stable, spatial profiles of the active functional biomass.* Our simulations show that profiles of daily averaged transcript concentrations, enzymes, and functional biomass are very similar and may generally be linked to the denitrification activity. Given the challenges of measuring time-averages of transcript concentrations and excluding inactive biomass in DNA-based methods, enzyme measurements seem to be the most accurate proxy variable for active functional biomass. These data will provide a relative measure of the spatially variable maximum rate coefficient in a biomass-implicit rate formulation. Compared to an enzyme-explicit formulation (as used in this study), a biomass-implicit formulation has the advantage that it requires fewer parameters. The hypothesis that the active functional biomass maintains a constant spatial distribution should be verified with repeated measurements at different time points, and seasonal trends could potentially be accounted for using several coefficients. If a considerable time-variability of active functional biomass is observed, a biomass- or enzyme-explicit model formulation that provides a process-based explanation for the variability should replace the biomass-implicit formulation.

2 *Measure oxygen, nitrogen substrates and DOC at several locations with a high*

temporal resolution. These data are required to appropriately account for substrate limitations and oxygen inhibition. The required resolution depends on the typical length and time scales of the system and might need to be determined iteratively. Spatial gradients and dominant temporal dynamics should be resolved. In order to capture the short-term variability inherent to these variables, continuous logging with probes, if possible, is a good approach (e. g., for oxygen). Otherwise, manual measurements should also cover several temporal scales. For example, hourly measurements that capture diurnal dynamics on individual days could be combined with daily or weekly samples to provide information about longer terms dynamics.

3 *Use a process-based model to obtain temporally and spatially resolved predictions of concentrations and reaction rates.* The model integrates the different data types through the calibration of model parameters, yielding estimates of total in-situ denitrification rates, that are otherwise impossible to obtain with direct measurements.

The predictions of a reactive-transport model strongly depend on transport related parameters, such as flow velocities or solute fluxes at boundaries, governed by subsurface hydraulic conductivity. Therefore, at field sites, complementary hydrogeological data should accompany biogeochemical investigations. For example, tracer tests could provide information on the average flow velocity. If flow cannot reasonably be assumed to be uniform and one-dimensional, hydraulic head data at several locations and hydraulic conductivity tests are required to set up a groundwater flow model.

4.4 CONCLUSIONS

Our model exercise highlights some of the prospects and limitations of using functional-gene transcripts and enzymes to characterize biogeochemical reactions at the river–groundwater interface. Concentrations of functional-gene transcripts quickly respond to changes in substrate concentrations and oxygen levels, implying that dynamic systems need to be sampled at the appropriate temporal resolution. High transcript and enzyme concentrations spatially coincide with active denitrification and are therefore qualitative indicators of reactive zones. Substrate limitation and oxygen inhibition of the enzymes, however, lead to complex, non-unique relationships between transcript or enzyme concentrations and reaction rates. We based our study on a relatively simple model that describes only a part of the system (gene regulation of denitrification) in detail, with the advantage that it enables a straight-forward analysis of predicted patterns. However, even with our simplistic model formulation, the relationships between transcript or enzyme concentrations and denitrification

rates are not straightforward. We conclude that concentrations of functional-gene transcripts and enzymes should not be used as a proxy for reaction rates.

Our results highlight that a rigorous quantitative interpretation of transcript or enzyme data requires a process-based mathematical model that is able to reflect non-linear interactions between biogeochemical processes and the regulation of gene and enzyme abundances. While our purely numerical study provides predictions of expected transcript and enzyme behavior in dynamic natural systems, it does not replace laboratory and field investigations. In fact, we emphasize that further improvements in enzyme-explicit model development will depend on highly-temporally resolved measurement campaigns.

The qualitative conclusions from our analysis are in general transferable to other environmental systems. Soils, for instance, are a very dynamic environment where redox conditions can change abruptly through changes in the hydrologic conditions like drainage or flooding (Pronk et al., 2020; Z. Zhang and Furman, 2021). Such short-term fluctuations will lead to a disconnect of quickly-reacting transcript concentrations from enzyme concentrations and, consequently, reaction rates. Additionally, oxygen availability in soils can vary spatially over very short distances because the slow diffusion of oxygen into the matrix produces anoxic microsites (Z. Zhang and Furman, 2021). The spatial hysteresis patterns that our model predicts for larger spatial gradients of oxygen and nitrogen species (centimeter to meters) might then occur on very small spatial scales (millimeters).

In natural systems, other nitrogen-cycling processes (nitrification, anammox, dissimilatory nitrate reduction to ammonium (DNRA)), alternative electron donors (e. g., reduced sulfur and iron species/minerals), and the temperature dependence of the reaction kinetics may affect denitrification rates. Nitrification acts as an additional source of nitrate for denitrification whereas DNRA competes with denitrification for nitrite. Adding these processes to the model will potentially modify the observed concentration profiles and reaction rates, and also affect the exact shape of the transcript-rate relationships. However, we expect that the general features that we observed – time-variable non-linear and non-unique relationships – will stay valid.

5 Quantifying uncertainty of hyporheic nitrogen-cycling rates with gene-based modeling

5.1 INTRODUCTION

The zone where rivers interact with groundwater (hyporheic zone) is a biogeochemically highly reactive environment where strong redox gradients and mixing of chemicals originating from groundwater and surface water promote microbial activity (Fischer et al., 2005; Krause et al., 2011). The microbial turnover of nitrogen compounds in the hyporheic zone can vitally affect the water quality in the connected river (Lewandowski et al., 2019).

Microbial nitrogen cycling comprises a complex network of reactions (Kuypers et al., 2018), where some reactions provide the substrates for the next reaction, and several reactions compete for the same substrates. For example, nitrification produces nitrate which can be either reduced to N_2 gas through denitrification or to ammonium via dissimilatory nitrate reduction to ammonium (DNRA). Nitrogen-cycling processes do not occur in an orderly fashion but different thermodynamically favorable reactions can happen simultaneously through a diverse community of versatile microorganism (Kuypers et al., 2018). Additionally, small-scale variations of chemical concentrations can lead to an apparent concomitance of reactions that require contrasting conditions, like nitrification and denitrification. This can lead to either a net release or consumption of nitrate and nitrite in the hyporheic zone, depending on the residence time of the water, nutrient availability, and water oxygenation (Akbarzadeh et al., 2018; Zarnetske et al., 2012). Because nitrogen species are both produced and consumed through microbial nitrogen cycling, it is difficult, or even impossible, to infer the reaction rate of a specific nitrogen-cycling process from solute concentrations only.

Biogeochemical models enable representing one's conceptual understanding of the processes occurring in a particular environment such as the hyporheic zone in a quantitative manner. A quantitative model that is not consistent with observations falsifies the underlying conceptual model. Process-based mathematical models can hence be used to test hypotheses. However, biogeochemical models are subject to several sources of uncertainty (Arhonditsis et al., 2018). Different parameter combinations

may lead to a comparably acceptable fit of the data, and measured concentrations are noisy. Further, the limited information contained in chemical concentrations may not be enough to disentangle the contributions of different nitrogen-cycling processes and reliably determine reaction rates. It is therefore essential to quantify uncertainty of the simulations, given concentration measurements.

Functional gene data of nitrogen-cycling genes provide information about the reactions that a microbial community can potentially carry out and, thus, may help to resolve the prevailing nitrogen-cycling pathways. Abundances of functional genes can be interpreted as a proxy for the biomass of functional groups of microorganisms that are capable of the reactions corresponding to the respective gene. Incorporating functional gene data into biogeochemical models therefore has the potential to improve predictions of microbial nitrogen-cycling rates. Gene-centric modeling (Reed et al., 2014) has been proposed as a straight-forward approach to incorporate functional gene data into biogeochemical models. Even though several review articles have found the approach promising (Dick, 2016; Grossart et al., 2020; Zhu et al., 2017), only few studies have applied it to date. These studies simulated biogeochemical cycling in a marine oxygen minimum zone (Louca et al., 2016), a deep-sea hydrothermal plume (Reed et al., 2015), and a river channel (Hui et al., 2021).

Parameters in these studies were fixed based on literature values and the simulation results were qualitatively compared to the measured gene data. Automatic calibration based on maximum likelihood estimates and Bayesian uncertainty quantification, however, require to encode a quantitative comparison of model predictions and data in the likelihood function. Data as obtained from metagenomics or relative-quantitative polymerase chain reaction (PCR) can only provide relative measures of gene abundance and, hence, Reed et al. (2014) suggested to compute shares of functional genes that can be compared between model outputs and data. A traditional least-squares approach (corresponding to the assumption of independent normally distributed data) does not adequately reflect the likelihood of gene shares because they have to sum up to unity and, thus, are not independent from each other. Using relative gene data for quantitative data-model integration therefore requires developing new measures of the goodness of fit for these data.

Applying a gene-centric approach, we assessed the uncertainty of microbial reaction rates of nitrogen in a system where several reaction pathways (and thus, microorganisms) compete for the same substrates. Our aim was to quantify the contributions of different nitrogen-cycling processes to the overall turnover. Further we wanted to assess the benefit that the functional gene data provide for identifying the relevant nitrogen-cycling processes and quantifying their rates. Finally, we investigated the most influential factors for nitrogen removal in the hyporheic zone.

To this end we set up a gene-centric biogeochemical model and calibrated it with data from an experimental study by Y. Liu et al. (2017). These authors conducted

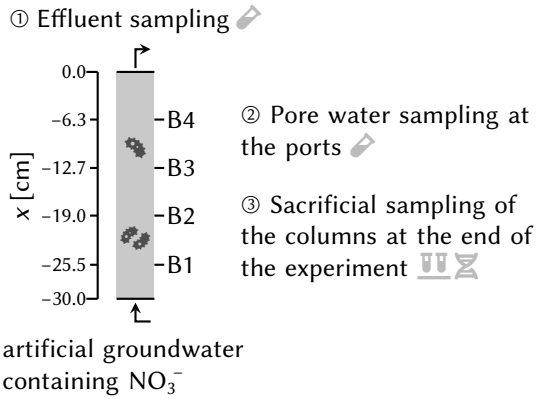


FIGURE 5.1: Set-up of the flow-through column experiment conducted by Y. Liu et al. (2017). Symbols indicate sampling of solute concentrations (), soluble or extractable inorganic nitrogen () and functional genes ().

flow-through column experiments with riverbed sediments that provide a rich data set of chemical concentrations and functional gene data for nitrogen-cycling genes. To see if and how the gene data reduce the uncertainty of parameter values and reaction rates, we calibrated the model both with and without the functional gene data. Based on the findings of Y. Liu et al. (2017) we hypothesize that residence time and organic carbon dynamics mainly determine the reaction rates.

5.2 MATERIAL AND METHODS

5.2.1 Simulated experiments

We set up a model that simulates the flow-through column experiments conducted by Y. Liu et al. (2017). In these experiments, Y. Liu et al. (2017) investigated nitrogen transformation processes in riverbed sediments of the Columbia River. The columns were packed with the homogenized sediments and flushed either with anoxic artificial groundwater rich in nitrate, or with oxic artificial surface water with lower nitrate contents. Before the start of the flow-through phase, the columns were presaturated with nitrate-free artificial groundwater or surface water. Concentrations of nitrogen species (NO_3^-, NO_2^-, N_2O, NH_4^+), dissolved organic carbon (DOC) and dissolved inorganic carbon (DIC) were monitored in the effluent, and in porewater taken from four ports along the column (figure 5.1).

At the end of the experiment, the columns were sacrificed to determine profiles of soluble NO_3^- and NO_2^-, extractable NH_4^+ and relative concentrations of functional genes. Functional genes of nitrogen cycling (*narG*, *napA*, *nirS*, *nosZ*, *nrfA*, archaeal and bacterial *amoA*, *nxrA*, and *hzo*) were quantified with quantitative PCR (qPCR). The groundwater columns were operated at a flow velocity of 2.2 m d^{-1} for 4 days and the surface water column was run at 2.0 m d^{-1} for 5 days. Afterwards, the velocity was

reduced to 1.2 m d^{-1} (groundwater column) and 1.1 m d^{-1} (surface water column) until the end of the experiment on day 11 and 16, respectively.

The model is currently limited to the anoxic groundwater column because initial analyses showed that even with a limited set of anoxic reactions the system is quite complex and the uncertainty estimation is challenging. It would be interesting, though, to extend the analysis to the oxic system that features more potential nitrogen-cycling reactions but requires substantially higher model complexity.

5.2.2 Microbial reactions

Reaction stoichiometry In the anaerobic system, we consider three denitrification steps (with corresponding genes *narG*, *nirS*, and *nosZ*), and reduction of nitrite to ammonium in DNRA, corresponding to the gene *nrfA*. Anammox is also a potential nitrogen-cycling reaction that can take place under anoxic conditions. However, anammox genes (*hzo*) were the least abundant ones in the sediment. We set up a model variant that includes the anammox reaction and genes, and the results indicated that anammox reaction rates are negligible. Therefore, we excluded anammox from the further analysis.

DOC serves both as the carbon source for heterotrophic organisms and as the electron donor for denitrification and DNRA. We use the molecular formula CH_2O to describe a single pool of generic organic carbon. Biomass is described generically as a compound with the molecular formula $C_5H_9O_{2.5}N$ based on data from seven microorganism (Roels, 1983). The catabolic and anabolic reactions considered in the model are given in table 5.1.

The overall metabolic reaction is obtained by coupling a catabolic reaction j with the corresponding anabolic reaction. The factor coupling the two reactions, and therefore the stoichiometric coefficients γ_{ji} of reactant i in the metabolic reaction, depend on the growth yield. Appendix D.1 provides details on how to calculate the stoichiometric coefficients from the stoichiometry of the catabolic and anabolic reaction, and the growth yield.

Reaction kinetics The reaction kinetics are described by multiple Monod terms multiplied by the concentrations of the functional gene linked to reaction j (Reed et al., 2014). The rate of the substrate in reaction j is

$$r_s^j = v_{\max}^j \frac{\Gamma_j}{g_j} \prod_{i \in \text{substrates}} \frac{c_i}{c_i + K_i^j} \,. \tag{5.1}$$

Here, $v_{\max}^j$ [mol cell^{-1} d^{-1}] is the maximum cell specific reaction rate of substrate s, and K_i^j [mol L^{-1}] is the half-saturation constant of substrate i in reaction j. Γ_j [genes L^{-1}]

TABLE 5.1: Stoichiometry of the catabolic and anabolic reactions considered in the model.

Reaction	Reaction stoichiometry	Gene
Catabolic reactions		
Nitrate reduction	$2\,NO_3^- + CH_2O \longrightarrow 2\,NO_2^- + H_2O + CO_2$	*narG*[a]
Nitrite reduction to N_2O	$2\,NO_2^- + 2\,H^+ + CH_2O \longrightarrow N_2O + 2\,H_2O + CO_2$	*nirS*[b]
N_2O reduction	$2\,N_2O + CH_2O \longrightarrow 2\,N_2 + H_2O + CO_2$	*nosZ*
Nitrite reduction to NH_4^+	$2\,NO_2^- + 4\,H^+ + 3\,CH_2O \longrightarrow 2\,NH_4^+ + H_2O + 3\,CO_2$	*nrfA*
Anabolic reactions		
Heterotrophic growth	$21\,CH_2O + 4\,NH_4^+ \longrightarrow 4\,C_5H_9O_{2.5}N + 9\,H_2O + CO_2 + 4\,H^+$	

[a] We use *narG* as the functional gene for nitrate reduction, and sum the measured *napA* and narG genes. [b] We use *nirS* as the functional gene for nitrite reduction to N_2O, and sum the measured *nirK* and nirS genes.

denotes the concentration of functional gene j, and g_j is the average number of genes j per cell. We assumed that there is a single gene copy per cell for all genes, as only a minority of microorganisms is known to have multiple copies of genes related to denitrification or DNRA per genome (Jones et al., 2008; Welsh et al., 2014). In contrast to Reed et al. (2014), we neglect inhibition by oxygen because the system is anoxic.

The net reaction rates for the microbial transformations are given by

$$\boldsymbol{r}_{\text{net}}^{\text{mic}} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}, \tag{5.2}$$

where $\boldsymbol{r} \in \mathbb{R}^M$ is a vector containing the reaction rates r_s^j (with M being the number of reactions). Given a system with N reactants, $\boldsymbol{A}$ is a $M \times N$ matrix containing the normalized stoichiometric coefficients. The coefficient for reactant i in reaction j is

$$A_{j,i} = \frac{\gamma_{ji}}{|\gamma_{j,s}|}, \tag{5.3}$$

where $\gamma_{j,s}$ is the stoichiometric coefficient of the substrate s in reaction j.

Instead of simulating CO_2 concentrations we use DIC as state variable because, DIC rather than CO_2 concentrations were measured, and because this eliminates the need to account for pH-dependent speciation of inorganic carbon.

Gene production and decay The growth rate of microbial cells due to reaction j, r_{growth}^j [cells $L^{-1}\,s^{-1}$], is linked to the reaction rate r_s^j by the stoichiometric coefficients

of biomass in the metabolic reaction:

$$r^j_{\text{growth}} = \frac{M_{\text{bio}}}{w_{\text{bio}}} \frac{\gamma_{j,\text{bio}}}{|\gamma_{j,s}|} r^j_s \,, \tag{5.4}$$

where M_{bio} [g mol^{-1}] is the molar weight of biomass and w_{bio} [g cell^{-1}] is the cell weight. The production rate of gene j is then given by

$$r^j_{\text{gene}} = g_j \, r^j_{\text{growth}} \,. \tag{5.5}$$

Here, we do not account for metabolic plasticity as Reed et al. (2014) suggested, that is, we assume that each microorganism can carry out only one metabolic reaction rather than several. Decay of genes is described with a first-order rate law with the same decay coefficient k_{dec} [d^{-1}]:

$$r^j_{\text{dec}} = k_{\text{dec}} \, \Gamma_j \,. \tag{5.6}$$

5.2.3 Release of organic carbon and nitrogen

The hydrolysis of organic compounds in the sediment was the only source of DOC in the column. We describe release of DOC from the sediment with a linear driving-force approach (Gu et al., 2007):

$$r^{\text{DOC}}_{\text{release}} = k_{\text{release}} \left(\frac{c_{\text{POC}}}{K_d^{\text{POC}}} - c_{\text{DOC}} \right) , \tag{5.7}$$

where k_{release} [d^{-1}] is the release rate constant, c_{POC} [g g^{-1}] is the content of bioavailable particulate organic carbon (POC) in the sediment, K_d^{POC} [L g^{-1}] is the distribution coefficient and c_{DOC} is the aqueous concentration of DOC. The change in bioavailable POC content is given by

$$r_{\text{POC}} = -\frac{\phi}{\rho_b} r^{\text{DOC}}_{\text{release}} + r_{\text{hydroysis}} \,, \tag{5.8}$$

where ϕ is the porosity and ρ_b [g L^{-1}] is the bulk density of the sediment. We considered a constant production rate for bioavailable POC, $r_{\text{hydroysis}}$, that conceptually represents the breakdown of complex organic material into smaller, bioavailable compounds. We tested model variants both with and without this production term and rejected the model with a single pool of POC (i. e., without $r_{\text{hydroysis}}$) because it yielded a poorer fit for the nitrate data (figure D.9).

Organic material in the sediment also contains organic nitrogen that is converted to ammonium during the hydrolysis. We assume that the production of ammonium linearly relates to the release of DOC by the inverse of the C:N-ratio $\kappa_{\text{C:N}}$ of the organic material:

$$r^{\text{NH}_4^+}_{\text{release}} = \frac{1}{\kappa_{\text{C:N}}} r^{\text{DOC}}_{\text{release}} \tag{5.9}$$

5.2.4 Advective-dispersive transport

Transport of solutes is described by the advection-dispersion-reaction equation as given in equation (2.11) in chapter 2. We describe the decrease of the flow velocity during the experiment by a logistic function that has the shape of a smoothed step:

$$v(t) = v_0 + (v_1 - v_0)\,\frac{1}{1 + \exp\left(-b(t - t_{\text{step}})\right)}\,, \tag{5.10}$$

where v_0 and v_1 [$\mathrm{m\,d^{-1}}$] are the constant flow velocities before and after the time of switching the flow rate, t_{step}[d], and b [$\mathrm{d^{-1}}$] defines the steepness of the step.

We assume that microbial cells are attached to the sediment, and that, as a consequence, the transport of functional genes can be neglected.

5.2.5 Ammonium sorption

Ammonium transport can be substantially retarded by cation exchange (Böhlke et al., 2006; Ceazan et al., 1989; Triska et al., 1994). The levels of ammonium extracted from the sediment with KCl at the end of the experiment were nearly two orders of magnitude higher than the measured aqueous ammonium concentrations. This suggests that ammonium sorption is an important process for the fate of ammonium. We describe kinetic ammonium sorption with a rate law similar to equation (5.7):

$$r_{\text{sorption}} = k_{\text{sorption}}\left(K_d^{\mathrm{NH_4^+}} c_{\mathrm{NH_4^+}} - s_{\mathrm{NH_4^+}}\right)\,, \tag{5.11}$$

where k_{sorption} [$\mathrm{d^{-1}}$] is the kinetic rate constant, $K_d^{\mathrm{NH_4^+}}$ [$\mathrm{L\,g^{-1}}$] is the equilibrium sorption coefficient, and $s_{\mathrm{NH_4^+}}$ [$\mathrm{mol\,g^{-1}}$] is the concentration of ammonium sorbed to the solid phase. The related change of concentration of ammonium in the liquid phase is

$$\left.\frac{\partial c_{\mathrm{NH_4^+}}}{\partial t}\right|_{\text{sorption}} = \frac{\rho_b}{\phi} r_{\text{sorption}}\,. \tag{5.12}$$

5.2.6 Initial and boundary conditions

At the inflow boundary we set a constant concentration condition based on the composition of the inflow solution (table 5.2). At the outflow boundary, the dispersive flux was assumed to be zero. The advective flux at the outflow boundary was given by the chosen upwind differentiation scheme. Before the start of the flow-through period, the columns were presaturated with nitrate-free artificial surface water or groundwater. We assumed that solute concentrations at the end of this incubation phase were uniform throughout the column. We then used the concentrations measured at the effluent during the first 0.2 pore volumes as the initial conditions (table 5.2).

TABLE 5.2: Initial and inflow concentrations used in the simulations. Inflow concentrations of species that are not listed are zero.

Parameter	Value	Units
Inflow NO_3^- concentration	4.48×10^{-4}	M
Inflow DIC concentration	1.6×10^{-3}	M
Initial NO_3^- concentration	0	M
Initial NO_2^- concentration	0	M
Initial N_2O concentration	0	M
Initial NH_4^+ concentration	1.4×10^{-5}	M
Initial DOC concentration	4.4×10^{-4}	M
Initial DIC concentration	3.1×10^{-3}	M
Initial *narG* concentration	4.4×10^{11}	genes L^{-1}
Initial *nirS* concentration	2.0×10^{11}	genes L^{-1}
Initial *nosZ* concentration	2.1×10^{9}	genes L^{-1}
Initial *nrfA* concentration	1.0×10^{9}	genes L^{-1}

Because of the presaturation phase it was assumed that the sorbed ammonium and POC were in equilibrium with their respective liquid concentrations. We therefore set their initial values based on the initial aqueous concentrations and the equilibrium constants:

$$s^0_{NH_4^+} = c^0_{NH_4^+} K_d^{NH_4^+}, \tag{5.13}$$

$$s^0_{POC} = c^0_{DOC} M_c K_d^{POC}. \tag{5.14}$$

Functional genes were assumed to be homogeneously distributed in the column at the start of the experiment because the columns were packed with homogenized sediments. The qPCR data did not provide absolute concentrations of the functional genes but only data relative to the abundance of 16S ribosomal RNA (rRNA) genes. To obtain absolute gene concentrations in genes L^{-1} we scaled the relative concentrations with a parameter indicating the total concentration of all considered functional genes. This parameter is not identifiable from the data because the data do not contain information about the absolute magnitudes of gene concentrations. Therefore we fixed the parameter at a value of 1.5×10^{12} genes L^{-1}, yielding the gene concentrations listed in table 5.2. Note that, as a consequence, estimated parameter values, in particular the maximum specific consumption rates v^j_{max}, will be conditional on this value, and caution is advised when comparing them to estimates obtained with different methods.

5.2.7 Parameter estimation and uncertainty quantification

We used Bayesian inference to fit simulated values to measurements and quantify parametric uncertainty. A Bayesian workflow comprises three basic steps (van de Schoot et al., 2021): (1) defining prior parameter distributions that express expert knowledge about the parameter values in ignorance of the data, (2) quantifying the goodness of fit between model outputs and measurements with a likelihood function, and (3) combining the knowledge about the parameters from the prior and the likelihood in the posterior distribution.

Prior distributions The choice of prior distributions is always subjective which has frequently been used as argument against Bayesian methods. Conversely, defining prior distributions forces scientist to state their assumptions explicitly, and helps ensuring physically meaningful results that are consistent with current expert knowledge. Because the choice of priors can substantially influence the results of a Bayesian analysis, the prior distributions should be chosen carefully, especially when the data do not contain enough information to constrain a parameter. We therefore started the process of selecting priors by creating a database of reference parameter values from the literature.

More than 40 parameter values from the literature are tabulated in appendix D.3 (tables D.3 to D.7 on page 136).

Nearly all reaction parameters are positive, real-valued variables. These parameters were log-transformed, and a normal distribution was chosen as the prior. The log-transformation ensures that parameter values are positive. Furthermore it allows to express uncertainty about the order of magnitude of a parameter, given the large variability of literature values. The mean values and standard deviations of the normal distributions were chosen such that the distributions cover the full range of literature values. Many of the standard deviations for the log parameters (using a natural logarithm) are larger than 1.5, corresponding to a factor of about half an order of magnitude for the non-logarithmic values. The prior distributions, thus, range over several orders of magnitude and are only weakly informative. The growth yields were expressed in relative terms (see appendix D.1) such that they range between 0 and 1. We therefore chose a Beta distribution for the prior of relative growth yields. Prior distributions of all parameters are listed in table 5.3.

Likelihood Solute data $\boldsymbol{y}_c^{ij}$ of the measured chemical species i at measurement number j (accounting for several measurements in time and space) were used to define the likelihood. We chose a Student's t distribution with location μ_{ij}, standard deviation σ_i and degrees of freedom d to not overly weight outliers in the data:

$$\boldsymbol{y}_c^{ij} \sim T(\mu_{ij}, \sigma_i, d) \tag{5.15}$$

TABLE 5.3: Prior distributions of the model parameters.

Description	Symbol	Prior distribution[a]		
		type	θ_1	θ_2
Log half-saturation constant for...				
DOC in NO_3^--reduction to NO_2^-	$\log\left(K_{\mathrm{DOC}}^{narG}/_{\mathrm{M}}\right)$	N	$\log\left(5\times10^{-5}\right)$	1
DOC in NO_2^--reduction to N_2O	$\log\left(K_{\mathrm{DOC}}^{nirS}/_{\mathrm{M}}\right)$	N	$\log\left(5\times10^{-5}\right)$	1
DOC in N_2O-reduction to N_2	$\log\left(K_{\mathrm{DOC}}^{nosZ}/_{\mathrm{M}}\right)$	N	$\log\left(5\times10^{-5}\right)$	1
DOC in NO_2^--reduction to NH_4^+	$\log\left(K_{\mathrm{DOC}}^{nrfA}/_{\mathrm{M}}\right)$	N	$\log\left(5\times10^{-5}\right)$	1
NO_3^- in NO_3^--reduction to NO_2^-	$\log\left(K_{\mathrm{NO_3^-}}^{nirS}/_{\mathrm{M}}\right)$	N	$\log\left(3\times10^{-5}\right)$	1
NO_2^- in NO_2^--reduction to N_2O	$\log\left(K_{\mathrm{NO_2^-}}^{nirS}/_{\mathrm{M}}\right)$	N	$\log\left(4\times10^{-6}\right)$	1
N_2O in N_2O-reduction to N_2	$\log\left(K_{\mathrm{N_2O}}^{nosZ}/_{\mathrm{M}}\right)$	N	$\log\left(10^{-6}\right)$	1
NO_2^- in NO_2^--reduction to NH_4^+	$\log\left(K_{\mathrm{NO_2^-}}^{nrfA}/_{\mathrm{M}}\right)$	N	$\log\left(10^{-5}\right)$	1
Maximum cell-specific reaction rate for...				
NO_3^--reduction to NO_2^-	$\log\left(v_{\max}^{narG}/_{(\mathrm{mol\,cell^{-1}\,d^{-1}})}\right)$	N	$\log\left(10^{-14}\right)$	1.5
NO_2^--reduction to N_2O	$\log\left(v_{\max}^{nirS}/_{(\mathrm{mol\,cell^{-1}\,d^{-1}})}\right)$	N	$\log\left(10^{-14}\right)$	1.5
N_2O-reduction to N_2	$\log\left(v_{\max}^{nosZ}/_{(\mathrm{mol\,cell^{-1}\,d^{-1}})}\right)$	N	$\log\left(10^{-14}\right)$	1.5
NO_2^--reduction to NH_4^+	$\log\left(v_{\max}^{nrfA}/_{(\mathrm{mol\,cell^{-1}\,d^{-1}})}\right)$	N	$\log\left(10^{-14}\right)$	1.5
DOC release rate constant	$\log(k_{\mathrm{release}}/\mathrm{d^{-1}})$	N	$\log(0.1)$	2.5
POC distribution coefficient	$\log(K_d^{\mathrm{POC}}/_{(\mathrm{L\,g^{-1}})})$	N	$\log(0.042)$	1
POC production rate	$\log(r_{\mathrm{hydroysis}}/_{(\mathrm{g\,g^{-1}\,d^{-1}})})$	N	$\log(10^{-6})$	1
NH_4^+ equilibrium sorption constant	$\log\left(K_d^{\mathrm{NH_4^+}}/_{(\mathrm{L\,g^{-1}})}\right)$	N	$\log(0.01)$	2.5
NH_4^+ kinetic sorption coefficient	$\log\left(k_{\mathrm{sorption}}/\mathrm{d^{-1}}\right)$	N	$\log(1.0)$	1.0
Relative growth yield for...				
NO_3^--reduction to NO_2^-	Y_{narG}^{rel}	Beta	1.5	20
NO_2^--reduction to N_2O	Y_{nirS}^{rel}	Beta	1.5	20
N_2O-reduction to N_2	Y_{nosZ}^{rel}	Beta	1.5	20
NO_2^--reduction to NH_4^+	Y_{nrfA}^{rel}	Beta	1.5	20
Scale parameter for the gene likelihood	$\log(s_\chi)$	N	$\log(80)$	0.5

[a] N indicates a normal distribution where the first parameter θ_1 is the mean value μ, and θ_2 is the standard deviation σ. Beta indicates a Beta distribution with shape parameters $\theta_1 = \alpha$ and $\theta_2 = \beta$.

The μ_{ij} are given by the Box-Cox transformed simulated concentrations. We applied the Box-Cox transformation (Box and Cox, 1964) to account for heteroscedasticity of the residuals. The parameters λ_i of the Box-Cox transformation and the standard deviations σ_i were adjusted manually.

Y. Liu et al. (2017) quantified functional genes of nitrogen cycling only relative to gene counts of the 16S rRNA gene. In contrast, the simulation yields absolute concentrations of functional genes. In order to compare the simulation results to the data, we computed the fraction of each functional gene i at time point j with respect to the sum of all genes for which data are available. It is defined as

$$\chi_{ij} = \frac{\Gamma_{ij}}{\sum_k \Gamma_{kj}} \,. \tag{5.16}$$

Since $\chi_{ij} \in (0, 1)$ and $\sum_i \chi_{ij} = 1$ we chose a Dirichlet distribution to define the likelihood of the measured gene fractions. Let $\boldsymbol{y}_\chi^j$ be the vector of all measured gene fractions for measurement j. We assume that

$$\boldsymbol{y}_\chi^j \sim \mathrm{Dir}(\boldsymbol{a}_j) \,, \tag{5.17}$$

where the vector of concentration parameters $\boldsymbol{a}_j$ of the Dirichlet distribution is defined by $a_{ij} = \chi_{ij} s_\chi$. This means that the simulated gene fractions χ_{ij} define the expected value of the distribution. The spread of the distribution is given by the scaling parameter s_χ. The larger s_χ is, the narrower is the distribution. This is illustrated for a hypothetical example of three genes in figure 5.2. We estimated s_χ from the data, using a lognormal distribution as the prior (see table 5.3).

Fixed parameters Transport parameters were set to fixed values because we expected their uncertainty to be small compared to the uncertainty of reaction parameters. The dispersion coefficient was chosen by manually fitting the breakthrough curve of a conservative tracer (bromide). For the parameters of the velocity function, we fitted equation (5.10) to velocity data from flow-rate measurements. The bulk density was calculated as $\rho_b = (1 - \phi)\rho_s$ where we used the density of quartz ($2.65\,\mathrm{g\,cm^{-3}}$) for the solid density ρ_s.

Some of the reaction parameters were fixed because they were poorly identifiable given the available chemical and microbial data. This applies to the absolute value of initial gene concentrations, the weight of a microbial cell, and the microbial decay constant because the qPCR data provided relative measures of functional gene abundance only. We also fixed the C:N-ratio of organic material $\kappa_{\mathrm{C:N}}$ based on measured ratios from Y. Liu et al. (2017). The fixed parameters and their values are listed in table 5.4.

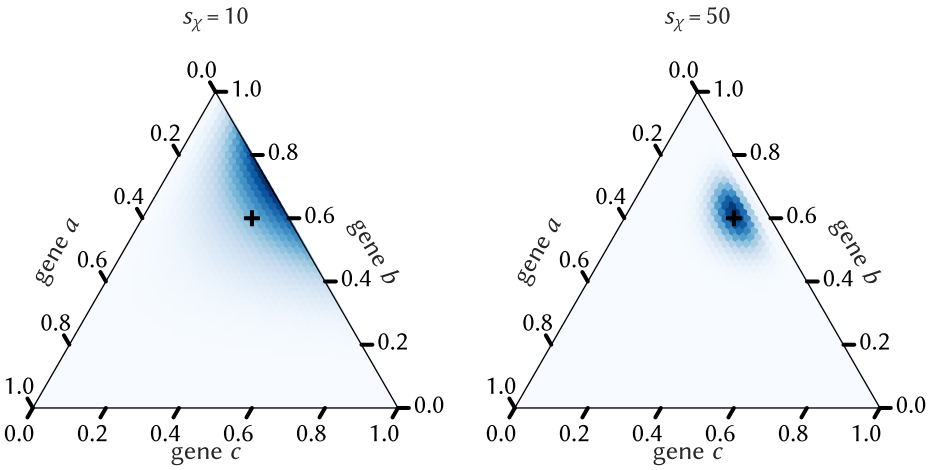


FIGURE 5.2: Illustration of the Dirichlet distribution for defining the likelihood of gene fractions. The example is simplified to a measurement of three genes for illustration purposes. The simulated fractions of the three genes represent the mean of the distribution and are marked by a cross. Note that the mean of the distribution does not equal its mode. Colors indicate the likelihood of measured gene fractions. The two subplots show the effect of the scaling parameter s_χ. The ternary plots are created with python-ternary (Harper, 2021).

TABLE 5.4: Model parameters set to fixed values.

Description	Symbol	Value	Units	Reference
Grid spacing	Δx	0.299	cm	
Length of the column	L	29.9	cm	1
Porosity	ϕ	0.39		1
Bulk density	ρ_b	1612	$g\,L^{-1}$	quartz density
Time of velocity reduction	t_{step}	3.9	d	flow rate data
Velocity before t_{step}	v_0	2.2	$m\,d^{-1}$	flow rate data
Velocity after t_{step}	v_1	1.2	$m\,d^{-1}$	flow rate data
Steepness parameter for v	b	20	d^{-1}	flow rate data
Dispersivity	α_L	1.2	cm	tracer data
Molecular diffusion coefficient	D_m	8.64×10^{-5}	$m^2\,d^{-1}$	2
C:N-ratio	$\kappa_{C:N}$	8.5	$mol\,mol^{-1}$	1
Microbial decay coefficient	k_{dec}	0.01	d^{-1}	see table D.4
Cell weight	w_{bio}	0.1	$pg\,cell^{-1}$	3
NO_3^- inflow concentration	$c^{in}_{NO_3^-}$	0.448	mM	1
DIC inflow concentration	c^{in}_{DIC}	1.6	mM	1

[1] Y. Liu et al. (2017) [2] Picioreanu et al. (1997) [3] Loferer-Krößbacher et al. (1998)

Calibration without functional gene data To investigate the influence of functional gene data on parameter inference we set up two model variants. In the variant *with gene data*, initial gene concentrations were based on measured gene fractions at the beginning of the experiment, and profiles of gene fractions at the end of the experiment informed the likelihood. In the alternative model variant *without gene data*, the initial gene concentrations become parameters that are estimated by calibration. As the prior we chose a log-normal distribution with the same mean value for all genes, that is, with equal gene fractions. For the likelihood, only solute-concentration data were used such that gene concentrations were not directly constrained.

5.2.8 Numerical methods

The reaction model was implemented in Python (Störiko, 2022). We applied a semi-discretization approach to discretize the reactive-transport equation in space using the Python package adrpy (Störiko, 2021). The resulting system of ordinary differential equations (ODEs) was solved numerically with the backwards differentiation formula (BDF) using the Python package sunode (Seyboldt, 2021). It automatically generates compiled code for the right-hand-side function and wraps the C-library SUNDIALS (Hindmarsh et al., 2005) for solving the ODEs.

We set up the statistical model with the help of the Python package PyMC (Salvatier et al., 2016). Samples from the posterior distribution were generated with PyMC's sequential Monte Carlo (SMC) algorithm, using a custom SMC kernel. We ran two independent chains to assess convergence of the sampler based on a visual comparison of posterior distributions, and on the rank-normalized split $\hat{R}$ convergence criterion (Vehtari et al., 2021). It compares between-chain and in-chain variance, and values of $\hat{R}$ greater than 1 indicate non-convergence. In the final simulations, deviations of $\hat{R}$ from 1 were smaller than 0.04 for all parameters.

Details on the sampling algorithm can be found in appendix D.2 (page 135).

5.2.9 Integrated reaction rates

Based on simulated concentrations and equation (5.1) we computed reaction rates for the different nitrogen-cycling processes. This allows us to see how reaction rates evolve in space and time, and compare the reaction rates of different processes. To estimate the contributions of different processes to overall nitrogen cycling in the column, we integrated the reaction rates along the advective flow path of a water parcel.

The position $x(t)$ of a water parcel that enters the column at time point t_0 can be described by the following ODE:

$$\frac{\mathrm{d}x_p(t)}{\mathrm{d}t} = v\Big(x_p(t), t\Big) = v(t)\,, \tag{5.18}$$

where the second equality holds because the advective velocity v is constant in space in our application. The initial condition is given by $x_p(t_0) = x_0$, where x_0 is the coordinate of the inflow boundary. We integrated equation (5.18) with an ODE solver (SciPy's `solve_ivp`, Virtanen et al., 2020) until the particle reaches the outflow boundary at x_{out}. This results in pairs of particle positions $x_p(t_0, t)$ and time points $t_p(t_0, x)$ at which the particle reaches the corresponding location. The total removal rate R_{ij} in the water parcel that enters the column at t_0 is then given by

$$R_{ij}(t_0) = \int_{x_0}^{x_{\text{out}}} r_{ij}\left(x, t_p(t_0, x)\right) \mathrm{d}x, \tag{5.19}$$

where $r_{ij}(x, t)$ is the reaction rate of compound i by reaction j at a given point in space and time. We computed the integral in equation (5.19) numerically with a trapezoidal rule and a constant temporal resolution of $1/100$ of the total travel time.

5.3 RESULTS

5.3.1 Solute concentrations and gene profiles

Nitrate time series Figure 5.3 shows the simulated and measured response of solute concentrations over time at the outflow of the column, and at the ports at 4 cm and 11 cm from the inlet. Overall, the model captures spatial and temporal trends of the data very well. Nitrate breaks through first at the ports and then also at the outflow, reaching concentrations that range between 60 % and 90 % of the inflow concentration (figure 5.3a). After the initial breakthrough, nitrate concentrations slowly increase. The reduction of the flow velocity leads to a sudden drop of the nitrate level, before concentrations slowly rise over time again.

Nitrate profiles Although the model slightly overestimates the nitrate concentration at the first two ports, it captures the decrease of the nitrate concentration in space, with the lowest concentrations being measured at the outflow. The decrease of the nitrate concentration in space is also evident from the nitrate profile (figure 5.4a). Measured soluble nitrate at the inflow of the column is higher than the nitrate concentration in the inflowing solution. We interpreted this finding as an indication of defective data because in the anoxic system there is no mechanism that could produce additional nitrate, and hence excluded the nitrate profile data from the calibration.

POC content Y. Liu et al. (2017) explain the observed increase of the nitrate concentration over time by decreasing denitrification rates due to a depletion of organic carbon. This explanation is consistent with our modeling results: The estimated initial bioavailable POC content is low (0.025 ‰), and the release of organic carbon into the

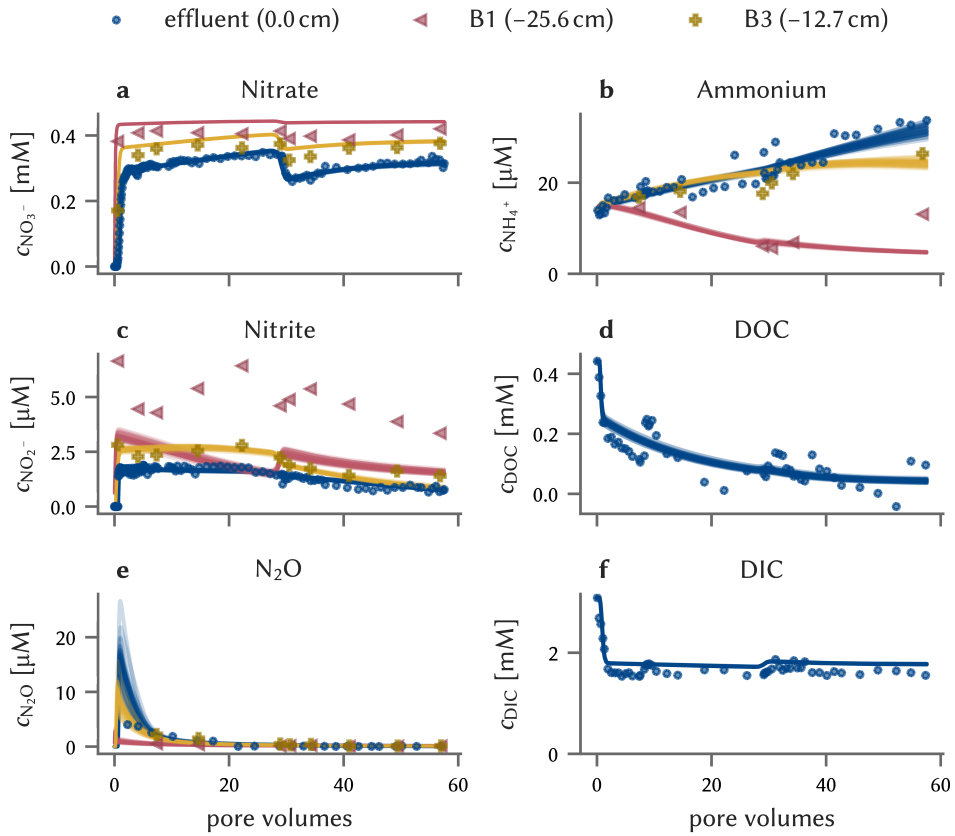


FIGURE 5.3: Simulated (lines) and measured (markers) aqueous concentrations at the column effluent and at two ports close to the inflow (B1) and in the middle of the column (B3) as a function of dimensionless time. Individual lines represent 40 draws from the posterior distribution of the model using gene data. Simulated concentrations for the model without gene data can be found in the appendix (figure D.3).

aqueous phase considerably depletes the POC content (figure D.1 in appendix D.4). In the second half of the experiment the decrease of the POC content slows down and reaches a steady state. This is when the production rate of bioavailable POC balances the release of carbon into the aqueous phase. Without this POC production term, the content of bioavailable POC in the sediment would constantly decrease, leading to an overestimation of the increase in nitrate towards the end of the simulation.

Nitrite time series Nitrite data are reproduced accurately in the effluent and at the two rear ports (17 cm and 24 cm). At the first two ports (figure 5.3c and figure D.2b), the model underestimates the overall nitrite level while reproducing some of the patterns present in the observations: After reducing the flow rate the nitrite concentration continuously decreases. The underestimation of nitrite concentrations in combination with overestimated nitrate concentrations at the first two ports suggests that the model underestimates the nitrate reduction rate in this part of the column. The reduction of nitrate requires an electron donor which is not supplied in the inflow solution. Nitrate reduction is therefore limited by the release kinetics of DOC (the only electron donor available). We hypothesize that the actual release of organic carbon near the inlet of the column is faster than what the model predicts. However, this is difficult to confirm because DOC concentrations have not been measured at the ports.

Nitrite profiles The soluble nitrite concentration measured in the spatial profiles at the end of the experiment is systematically higher than simulated concentrations throughout the profile (figure 5.4b). However, assuming that nitrite does not sorb onto the sediment, the soluble nitrite data is also inconsistent with pore water nitrite concentrations from the ports. Therefore, we considered the soluble nitrite measurements to be untrustworthy, and excluded the profile data from the calibration.

Ammonium profiles Extractable ammonium concentrations at the end of the experiment are two orders of magnitude higher than solute concentrations measured at the ports (figure 5.4c and figure 5.3b). The model can reproduce this behavior by explaining it with considerable sorption of ammonium. The equilibrium retardation factor obtained with the model ranges between 72.7 and 80.2 (based on the 5th and 95th percentile of the posterior distribution). Assuming equilibrium sorption, it would take 72.7 to 80.2 pore volumes for the ammonium-free artificial groundwater to break through at the end of the column – this is longer than the duration of the experiment.

Ammonium time series Only at the first port we see a drop in the ammonium concentrations corresponding to the flushing with ammonium-free water. Because sorption is not at equilibrium but kinetic, the front is smooth. Ammonium concentrations at the ports further away from the column inlet and in the effluent rise over time. At

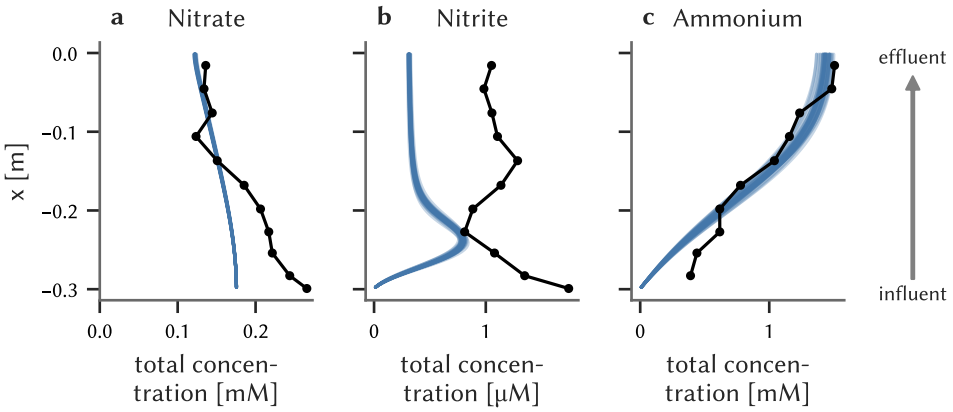


FIGURE 5.4: Spatial profiles of extractable ammonium, and soluble nitrate and nitrite at the end of the experiment. Only ammonium profile data were used for the parameter estimation.

the end of the experiment, there is a clear increasing spatial trend of the ammonium levels, both in the pore water (as measured in the ports, figure 5.3b and figure D.2d) and in the extractable ammonium measured in the profiles (figure 5.4c).

N_2O time series Measured N_2O concentrations at the effluent and at ports B2, B3 and B4 initially show elevated levels that gradually drop until 30 pore volumes, whereas N_2O at port B1 are low (<1 µM) all the time (figure 5.3e). The model qualitatively reproduces this behavior but exaggerates the height of the N_2O peak.

DIC and DOC time series Both DIC and DOC are present at relatively high levels after the pre-saturation phase but drop abruptly within the first pore volume as the water from the pre-incubation phase is flushed out (figure 5.3d and f). The DIC afterwards reaches a constant concentration. In contrast, DOC levels gradually decrease over time at a progressively slower rate. The model is able to capture this decrease, relating it to the decreasing POC content of the sediment that approaches a steady state towards the end of the simulation.

Spatial profiles of functional genes Figure 5.5 shows the simulated and measured fractions of nitrogen-cycling genes in the column at the end of the simulation. Overall, the gene fractions hardly change in comparison to the initial gene fractions (indicated in figure 5.5 by a gray line). This could either mean that there is little microbial growth or that all genes grow by the same factor. Only the shares of narG and *nirS* decrease and increase slightly, respectively, in the back two-thirds of the column. The model explains this with a growth of the *nirS* genes whereas the concentration of

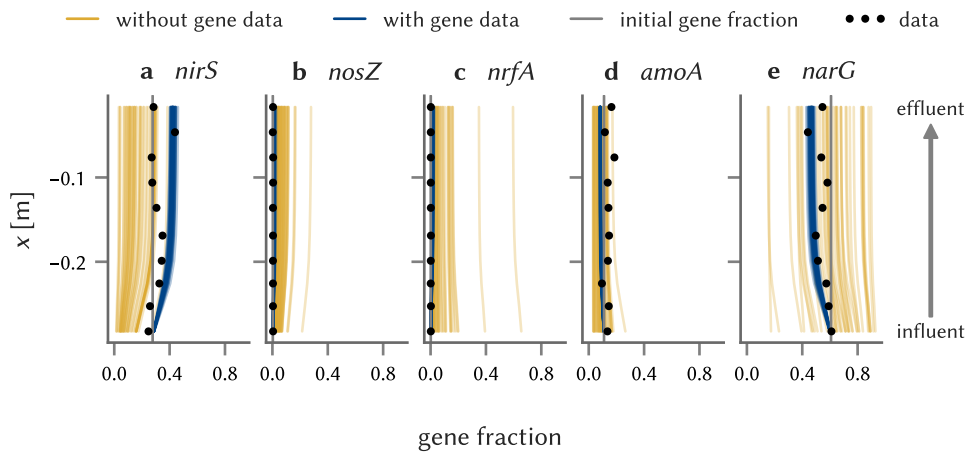


FIGURE 5.5: Simulated and measured profiles of relative abundances of functional genes. Dark blue lines show the model results when measured relative gene concentrations inform the initial conditions and the likelihood; yellow lines show the calibration results when only chemical data are used. In each case, 40 samples from the posterior distribution are drawn.

narG genes remains constant. While the data cannot distinguish whether *narG* genes increase or *nirS* genes decrease, the simulated absolute gene concentrations contain this information.

5.3.2 Nitrogen-cycling rates

Reaction rates of nitrogen-cycling processes were computed as outlined in section 5.2.9 to assess their importance for nitrogen turnover and uncertainty. Figure 5.6 shows how reaction rates are spatially distributed in the column (figure 5.6a), and how integrated reaction rates change over time (figure 5.6b).

Removal rates by denitrification processes become smaller over time. This decrease is caused by the depletion of bioavailable organic carbon. The relatively slow release of organic carbon from the sediment results in low, decreasing DOC concentrations that are in the range of the half-saturation constants, indicating that reaction rates are limited by the carbon substrate. The removal rate by DNRA (linked to the *nrfA* gene) is much lower than the denitrification rates, however, it slightly increases over time. Even though organic carbon availability limits the DNRA rates as well, the growth of the *nrfA*-carrying community leads to the increase of the reaction rate. Both denitrification and DNRA rates are low near the inflow of the column (figure 5.6a). Both reactions rely on the supply of organic carbon from the matrix as an electron donor, which requires a certain residence time within the column.

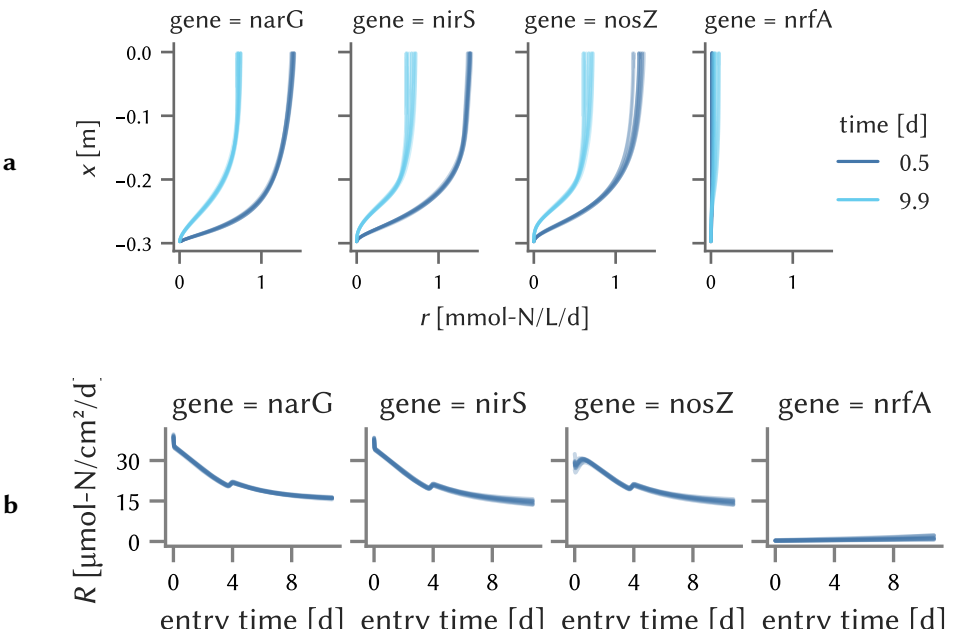
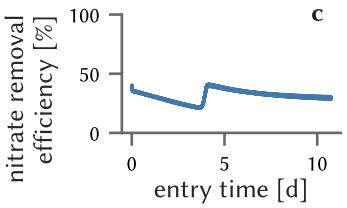


FIGURE 5.6: Simulated nitrogen-cycling rates. **a** Spatial profiles of the reaction rates. The genes *narG*, *nirS* and *nosZ* correspond to denitrification, and *nrfA* corresponds to DNRA. **b** Reaction rates integrated over the advective flow path of a water parcel. **c** Nitrate removal efficiency through *narG* genes.



The reduction in the flow velocity has only a small effect on the removal rates. However, the lower velocity also leads to a smaller mass flux of nitrate into the domain. We therefore calculated the nitrate removal efficiency, which we define as the ratio of the nitrate removal and the nitrate mass flux into the domain. The nitrate removal efficiency increases when the flow velocity is reduced (figure 5.6c). However, on the long term, it also declines by the decreasing organic carbon availability.

Figure 5.6 shows that the randomly chosen posterior samples yield almost identical predictions, indicating that the estimated uncertainty of reaction rates in the anoxic column is rather small. We expected that the interaction of several nitrogen-cycling processes would make it difficult to estimate reaction rates with confidence. However, the available data apparently provide enough information to constrain the reaction rates. It should be considered, though, that the small uncertainty of reaction rates could also arise from the fact that we neglect model structural uncertainty, and could be overconfident. We further discuss this issue in section 5.4.2.

5.3.3 Posterior parameter uncertainty and identifiability

The marginal parameter distributions (figure 5.7) show that, for some parameters, the posterior uncertainty is much smaller than the prior uncertainty, indicating that the data is informative about these parameters. For example, the maximum specific rate constants $v_{\max}^{narG}$ and $v_{\max}^{nosZ}$ have a very narrow posterior distribution (figure 5.7a and b). But also the equilibrium sorption coefficient of ammonium, $K_d^{\mathrm{NH_4^+}}$, is well determined, whereas the kinetic sorption coefficient k_{sorption} is less certain. Other parameters have broader posterior distributions that range over up to one order of magnitude, for example the maximum specific rate and the DOC half-saturation constant for DNRA, $v_{\max}^{nrfA}$ and K_{DOC}^{nrfA} (figure 5.7c and i). While the posterior distribution of $v_{\max}^{nrfA}$ is still much narrower than the prior distribution, the uncertainty of K_{DOC}^{nrfA} is hardly reduced in the posterior.

The marginal distribution of $v_{\max}^{nosZ}$ (figure 5.7b) is considerably shifted to larger values in comparison to the prior. This can be explained by the low abundance of *nosZ* genes. Even though *nosZ* genes are much less abundant than *narG* and *nirS* genes (corresponding to the first two steps of denitrification) we do not see N_2O accumulation. This indicates that the N_2O consumption is nearly as fast as its production. This can only be achieved if a high maximum specific rate compensates for the low abundance of *nosZ*. The half-saturation constant for nitrate $K_{\mathrm{NO_3^-}}^{narG}$ (figure 5.7g) is shifted to values that are much smaller than measured concentrations, effectively eliminating the Monod term from equation (5.1) as it becomes unity. That is, the rate law for nitrate reduction is effectively of zeroth order with respect to nitrate.

Some parameters are strongly correlated (figure 5.8), meaning that they are not uniquely identifiable. Particularly strong correlations can be observed between the

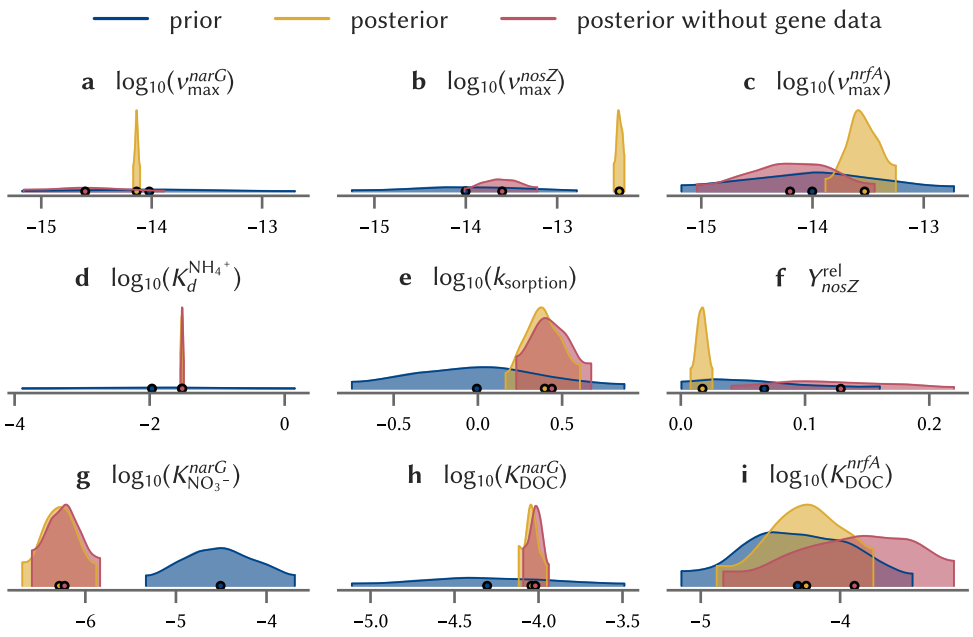


FIGURE 5.7: Kernel density estimates of the marginal posterior and prior distributions of selected parameters. Plots are cut at the 94 % highest density interval. Marginal distributions of all simulation parameters are plotted in figure D.4 in the appendix.

maximum specific rates v_{max}^{j} of the *narG*, *nirS* and *nrfA* gene, and the corresponding DOC half-saturation constants K_{DOC}^{j}. Additionally, the maximum specific rates of *narG* and *nirS* are correlated with each other. This correlation can be explained based on the observed data. The concentrations of NO_2^- are low, which is directly related to the ratio of the reaction rates to each other. If the second step was much slower than the first one, the intermediate product NO_2^- would accumulate. Since this is not the case, either both rate parameters must be large, or both must be small.

Pair-wise correlation coefficients and scatter plots can only show correlations between two parameters. However, combinations of more than two parameters can be correlated in the posterior. To explore the high-dimensional posterior parameter space, we computed the eigenvalues and eigenvectors of the covariance matrix. That is, we conducted a principal component analysis of the posterior parameter samples. Before computing the covariance matrix, all parameters were transformed to log-space (for positive parameters) or to logit-space (for parameters between 0 and 1). Each eigenvector is a linear combination of the original parameters representing a principal direction in the transformed parameter space. The set of eigenvectors span

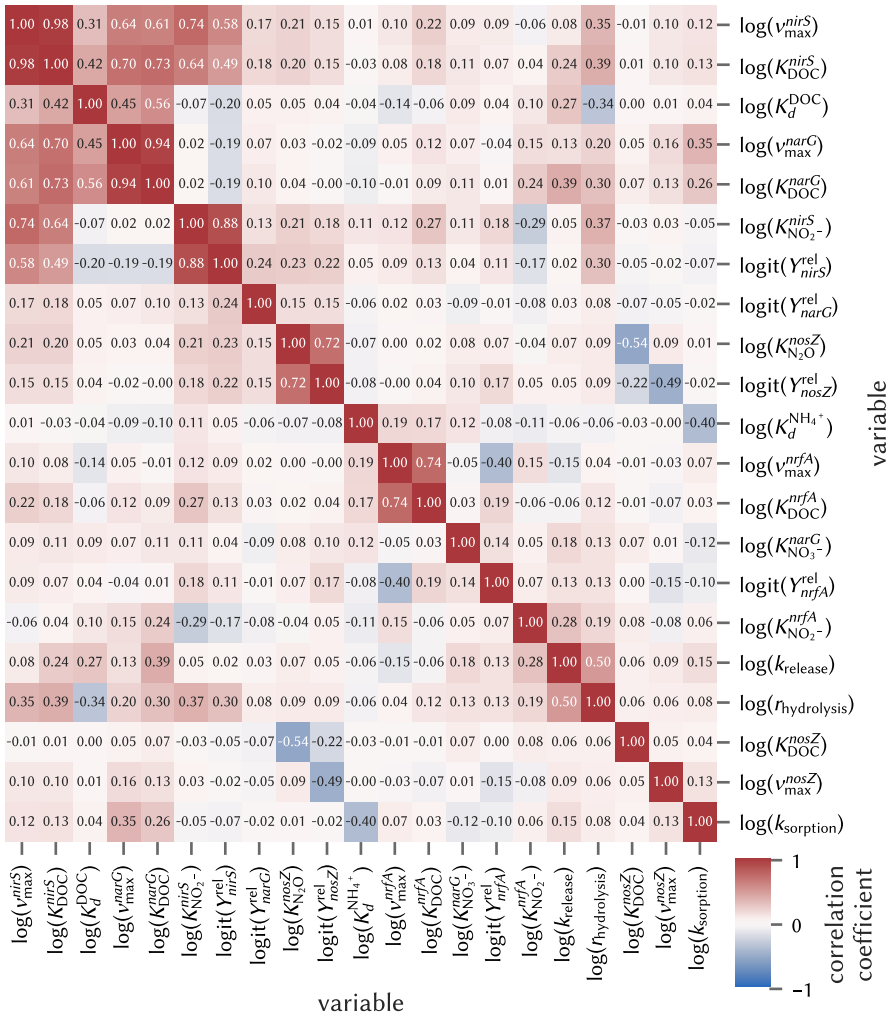


FIGURE 5.8: Heatmap showing the pairwise correlation coefficients of parameters in the posterior.

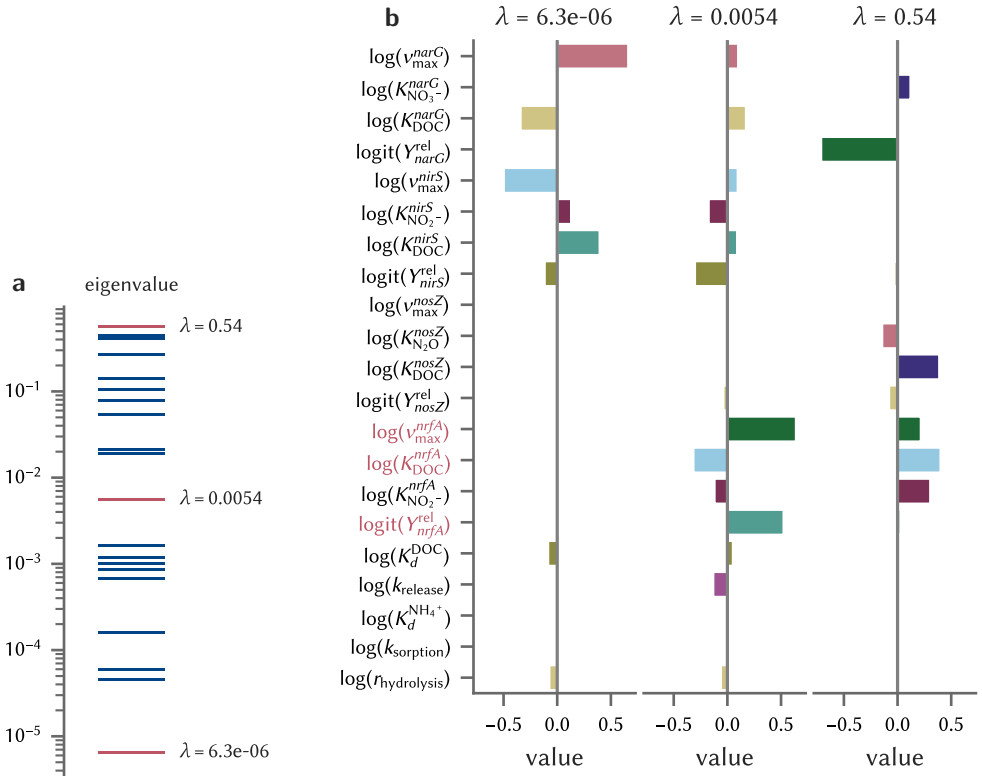


FIGURE 5.9: Eigenvalues of the posterior covariance matrix (**a**) and selected eigenvectors (**b**). The shown eigenvectors on the left and right correspond to smallest and the largest eigenvalue, respectively. The eigenvector in the middle corresponds to a relatively small eigenvalue but features contributions from parameters with large marginal variance (highlighted in red). Eigenvectors for all eigenvalues are shown in figures D.5 to D.7 in the supporting information.

a new orthogonal system of coordinates system and can be thought of as alternative parameters that are linearly uncorrelated. The eigenvalues λ indicate the posterior variance of these new parameters.

The eigenvalues cover a wide range of values from 6.3×10^{-6} to 0.54 (figure 5.9a). We computed the eigenvalue decomposition with log-transformed parameters, but it is more intuitive to look at the uncertainty of non-log parameters: An increase of the parameter value on the log scale by one standard deviation corresponds to an increase by a *factor* of $\exp(\sqrt{\lambda})$ on the non-log scale. Thus, $\lambda = 6.3 \times 10^{-6}$ corresponds to a factor of 1.003, and $\lambda = 0.54$ corresponds to a factor of 2.1. Eigenvectors of the small eigenvalues are a combination of the $v_{\max}$ of *narG* and *nirS* as well as the related half-saturation constants for DOC (figure 5.9b), indicating multiple dependencies between

the half-saturation constants and the $v_{\max}$, and between the Monod parameters of the two consecutive denitrification steps.

Although these parameters already have a relatively narrow marginal posterior distributions, a linear combination of these parameters can be inferred with even lower uncertainty. When the eigenvector of a small eigenvalue is not aligned with the original parameter axes, one can easily leave the typical set of parameters that significantly contribute to the posterior density by changing the original parameter by a small increment. That is, even though the marginal distributions cover a small parameter range, independent samples from the marginal distributions would likely produce simulation results that do not fit the data well.

The eigenvector for $\lambda = 0.0054$ is particularly interesting. It has large contributions from the Monod and growth parameters of the *nrfA* gene (K_{DOC}^{nrfA}, $v_{\max}^{nrfA}$, Y_{nrfA}^{rel}). The marginal posterior distributions of these parameters are broad, but the eigenvalue is still small, indicating that the individual parameters are poorly identifiable, whereas a linear combination of them can be well constrained by the data.

Non-identifiable parameters are a major challenge for model calibration, interpretation and predictions. When deterministic calibration methods (e. g., maximum likelihood estimation) are applied, non-identifiable parameters lead to an ill-posed optimization problem because the solution is not unique. Bayesian calibration methods can deal with non-uniqueness because they allow for joint distributions of parameter combinations (rather than a single optimum). Nonetheless, non-identifiability can still limit the usefulness of a model for system understanding and predictions. For example, the model parameters themselves may be of interest for interpreting the system behavior. Furthermore, non-identifiable parameters can cause ambiguous predictions for variables that have not been previously measured (*unobservability*), although this is not a necessary consequence (Villaverde, 2019).

One way of addressing non-identifiability is to find a reduced model that can still describe the system behavior by effective rate laws that eliminate the non-identifiable parameter combinations. For example, Marschmann et al. (2019) applied the manifold boundary approximation method (MBAM; Transtrum and Qiu, 2014) to soil carbon cycling models, and showed that the number of parameters can be dramatically reduced, replacing most of the non-linear Monod rate laws by linear equations. As discussed, the Monod parameters of our gene-based nitrogen-cycling model were not uniquely identifiable either. It would, thus, be interesting to apply MBAM in order to identify a reduced set of model equations that describe the effective model behavior. One draw-back of this data-driven model reduction method is that the reduced model obtained with it is specific to the used data set. When environmental conditions differ from the ones during the experiment (e. g., by a higher or lower substrate availability) the effective rate laws may change.

5.3.4 Information gain from functional gene data

When we do not use the gene data as the initial conditions and in the likelihood, the simulated concentrations of the solute look very similar to the case with gene data. However, the profiles of model-predicted gene fractions differ considerably between the two models (figure 5.5). Without the gene data, there is much more uncertainty about the share of each gene. For example, *nrfA* can be as abundant as *nirS* in some realizations whereas this is not possible when gene data are used. Without informing the model about the gene data, the posterior distribution fo the relative gene abundance remains close to the initial uniform distribution.

Adding the gene data does not affect some of the posterior parameter distributions at all, e. g., for sorption related parameters (figure 5.7). However, the maximum specific reaction rates and the growth yields have broader posterior distributions without the gene data. This indicates that the additional functional gene data help to better constrain the parameters, even though they only provide relative-quantitative information.

For the parameters $v_{\max}^{nosZ}$ and Y_{nosZ}^{rel} the additional gene data leads to a clear shift of the posterior distribution compared to the model without gene data instead of just narrowing it. This is related to the different predictions of the two models for the abundance of the *nosZ* gene. When the gene data are used, it forces the *nosZ* gene to be very low-abundant. As discussed above, this requires a compensation effect for the corresponding maximum specific rate parameter. Without the gene data, *nosZ* genes are much more abundant based on the prior information of equal shares of the genes. In this case, no compensation effect is necessary, and the posterior distribution of $v_{\max}^{nosZ}$ stays closer to its prior.

Although the gene data helps to determine the reaction parameters, it does not seem to be necessary in order to quantify reaction rates in the given set-up. Surprisingly, nitrogen-cycling rates are well constrained both with and without the gene data. This potentially indicates that the solute data is highly informative for inferring reaction rates. The set of possible nitrogen-cycling reactions in the anoxic system is still limited (denitrification, DNRA, and ammonification). The mass balance of nitrogen compounds needs to be closed, which sets an additional constraint on possible reaction rates. Together with the ample solute concentration data (including nitrate, nitrite, ammonium and N_2O concentrations), this leaves only little ambiguity with respect to the relevant nitrogen-cycling processes. For example, nitrate is only consumed, not produced, in an anoxic system. This means that the total removal of nitrate in the column can be easily calculated through a mass balance when the effluent nitrate concentration is known.

In a system where both oxic and anoxic nitrogen-cycling processes can act at the same time, however, the uncertainty about the magnitude of nitrogen-cycling process

rates might be considerably larger. In such a system, all nitrogen species can be either produced or consumed, such that solute data might not be sufficient to close the mass balance. We thus hypothesize that gene data become informative for constraining reaction rates when both oxic and anoxic nitrogen-cycling processes take place.

5.3.5 Reparametrization of the Monod rate law

Even though in theory the Bayesian sampler used in this study (sequential Monte Carlo) should converge towards the posterior distribution, the efficiency can be very low. That is, the sampler fails to reach the posterior distribution in a reasonable computational time – a common problem with many Markov chain Monte Carlo (MCMC) sampling algorithms. Correlations among parameters, particularly in multiple dimensions, can pose a major problem for the sampler. Even though the proposal distribution of the SMC sampler already takes into account the covariance structure of the parameters, strong parameter correlations lead to a poor performance of the sampler. Furthermore, relationships between parameters can also be non-linear, so that linear transformations of the parameters do not always improve the convergence behavior. How well sampling works strongly depends on the parametrization of the model.

As discussed, the half-saturation constants and the maximum specific rates appearing in the Monod law are strongly correlated, a problem well known from previous studies (C. Liu and Zachara, 2001; Robinson and Tiedje, 1983; Sierra et al., 2015). In our case, these multi-dimensional parameter correlations prevented effective sampling, so the sampler did not converge.

To solve this issue we introduced a reparametrization of the Monod parameters $v^j_{\max}$ and K^j_i similar to the one in Störiko et al. (2021a). $v^j_{\max}$ represents the rate at the limit of $c_i \to \infty$ for all substrates i, and K^j_i is the concentration of substrate i where $v_j = v^j_{\max}/2$. However, data will mainly provide information about the rate at the concentration levels found during the experiment. The same rates can be obtained with several Monod curves that can have very different limiting values. Therefore, instead of trying to determine the rate at the limit of infinite concentrations, we picked a point in the concentration space $\boldsymbol{c}^*$ that roughly corresponds to the observed concentrations. (Of course there is no unique or best choice because concentrations are not uniform in space and time.) We then chose the value and the slope of the Monod curve at this point as the parameters of the reparameterized kinetic function (figure 5.10).

More specifically, the first parameter v^*_j is the specific rate at fixed concentrations, that is

$$v^*_j = v_j(\boldsymbol{c}^*) = v^j_{\max} \cdot \prod_{i \in \text{substrates}} \frac{c^*_i}{c^*_i + K^j_i} \,. \tag{5.20}$$

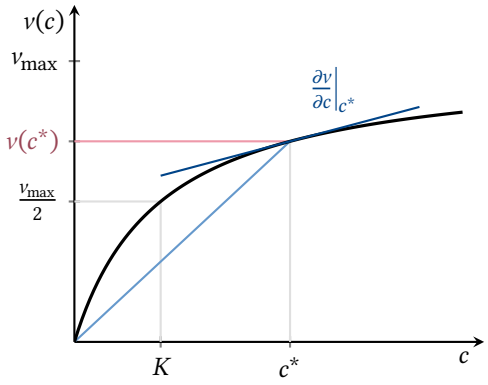


FIGURE 5.10: Reparametrization of the Monod parameters illustrated for the case of a single substrate. Instead of K and v_{max}, the value of the Monod curve at a fixed concentration, $v(c^*)$ (indicated in red), and the relative slope at that point (slope of the dark blue line divided by the slope of the light blue line) rescaled to $\mathbb{R}$ are used as parameters.

We denote the remaining parameters by z_{ij}. They represent the partial derivatives of the Monod curve with respect to concentration c_i evaluated at $\boldsymbol{c}^*$, $\frac{\partial v}{\partial c_i}\Big|_{\boldsymbol{c}^*}$, normalized by the maximum slope possible for a Monod curve running through the point $\boldsymbol{c}_i^*$. This slope is given by $v_j(\boldsymbol{c}^*)/c_i^*$ because the steepest possible Monod curve is a linear one, with all half-saturation constants approaching infinity. We rescaled this ratio to all of $\mathbb{R}$ by applying the sigmoid logit function which is given by $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ for $x \in (0, 1)$. This yields

$$z_{ij} = \text{logit}\left(\frac{\partial v}{\partial c_i}\Big|_{\boldsymbol{c}^*} \frac{c_i^*}{v_j(\boldsymbol{c}^*)}\right) = \text{logit}\left(\frac{K_i^j}{c_i^* + K_i^j}\right) = \log\left(\frac{K_i^j}{c_i^*}\right). \tag{5.21}$$

It turns out that this parameter equals the logarithm of the ratio of the half-saturation constant K_i^j and the fixed concentration c_i^*.

Instead of $\log(v_{max}^j)$ and $\log(K_i^j)$, the parameters that we estimate are $\log(v_j^*)$ and z_{ij}. PyMC can transform between the two parametrizations internally when the determinant of the transformation's Jacobian with respect to the parameters is provided. This allowed us to define prior distributions in terms of the original parameters where it is easier to find reference values in the literature. The new parameters were substantially less correlated, improving the sampler performance and enabling convergence.

5.4 DISCUSSION

5.4.1 *Improving sampler efficiency through reparametrization*

The fact that a reparametrization of the Monod parameters was necessary to achieve convergence illustrates one of the practical problems of Bayesian inference: that the sampler performance depends on the parametrization. While reparametrization can

alleviate the sampler problems, a good reparametrization may not always be evident. Analyzing the eigenvectors of the parameter covariance matrix can help identifying parameters that are strongly related, although it only provides a measure of linear dependence.

However, manual reparametrization becomes challenging and tedious for models with many parameters. Sampling algorithms that do not depend on parametrizations or that can automatically identify and apply the necessary transformations would, thus, be highly desirable, but their development is currently still an active field of research.

Transformations that convert a simple base distribution (e. g., a standard normal distribution) into a complex (e. g., asymmetric or multimodal) target distribution by a sequence of invertible and differentiable mappings are known as *normalizing flows* in the field of machine learning (Kobyzev et al., 2021). The individual transformations can be simple (e. g., linear) or more complex; common families of transformations are neural network and autoregressive models (Kobyzev et al., 2021). Normalizing flows can be combined with MCMC sampling to transform a posterior distribution with complex geometry into a simpler distribution that is easy to sample (Papamakarios et al., 2021). For example, M. Hoffman et al. (2019) used variational inference to fit a normalizing flow based on neural networks, and then sampled the well-behaving posterior density of the transformed parameter space with Hamiltonian Monte Carlo (HMC). Their study is a pioneering example showing how normalizing flows can be used to automate reparametrization and improve sampler performance.

5.4.2 Impact of model structure on uncertainty and parameter estimates

Parameter and uncertainty estimates obtained with Bayesian inference depend on the assumption that the underlying model is correct. Structural model errors can lead to biased parameter estimates and overconfident predictions (Sargsyan et al., 2019). When structural uncertainty or model error is neglected, a rich data set can lead to low predictive uncertainty, even if the predictions are wrong. These problems can be illustrated with our model.

1 As we have discussed above (section 5.3.3) the posterior distribution of $v_{\max}^{nosZ}$ is strongly shifted towards larger values compared to both the prior and the posterior distributions of other genes. This would indicate that N_2O reduction by organisms carrying the *nosZ* gene is one to two orders of magnitude more efficient than other nitrogen-cycling processes. It is doubtful that this is really the case. Another explanation could be that *nosZ* genes are underrepresented in the data, for example if the chosen PCR-primers cover only a fraction of the diverse *nosZ* genes (Sanford et al., 2012; B. Zhang et al., 2021).

This would mean that the assumption of our model that the relative abundance of functional genes can be compared between different functional genes is violated. In theory, we could try to correct different measurement efficiencies by introducing a new parameter for each gene that accounts for the gene-specific extraction or primer efficiency. However, in practice, it would be impossible to uniquely identify this parameter because it would always only appear as a product with the respective functional gene concentration. The additional parameter would thus lead to a more complex model without conferring a benefit. Instead, we should interpret the estimated maximum specific rates as a lumped parameter of the actual maximum specific rates and the efficiency with which a gene can be quantified. As long as the same analytical procedure (extraction method, choice of primers, etc.) is performed in another experiment, the determined parameters have predictive skills.

2 The second example shows how structural model error leads to overconfident estimates. In an earlier model variant of the model we had fixed the growth yield parameter in order to reduce the number of parameters, and because we had considered it to have a low impact on calibration results if only relative gene data is available. In the model with fixed growth yields, removing the gene data from the likelihood did not increase the uncertainty of gene fractions (figure D.8). On the contrary, the predicted distribution of gene fractions was narrow but wrong. Without the gene data, it would have been difficult to notice this erroneous result from a misbehaving model, because the predictions for the solute data hardly differed among the models. Along the same lines, the surprisingly low uncertainties of reaction rates estimated with the final model formulation could also indicate overconfidence as a result of model structural error.

These examples show the importance of evaluating the effects of model assumptions, and testing different model variants. More rigorous approaches exist for dealing with structural model errors. Some of them focus on diagnosing structural errors (e. g., Hsueh et al., 2022) in order to find better model formulations. Alternatively, structural model errors can be parameterized explicitly based on statistical modeling approaches (e. g., O'Hagan, 2013; Sargsyan et al., 2019; Xu et al., 2017).

5.4.3 Linking organic carbon dynamics and nitrogen cycling

Our results have shown that nitrate removal in the anoxic hyporheic-zone sediments is mainly limited by organic-carbon availability. This outcome is in line with previous findings from field experiments (Zarnetske et al., 2011) and sediment incubations (Lansdown et al., 2012). An appropriate description of organic-carbon dynamics is

therefore essential to capture reaction rates of nitrogen-cycling processes. While "organic carbon" in reality is a complex mixture of very different chemical compounds with highly variable properties, accounting for this in detail in a model is usually not feasible. Instead, one or several pools of generic organic carbon are used to describe the sum of many compounds.

For the system that we modeled, a single pool of POC in the sediment was not sufficient to describe the nitrate concentrations well at all time points. Adding a second, constant pool of organic carbon improved the description of the observed solute dynamics on the time scale of the experiment. However, over longer time scales, the assumption of a constant production of bioavailable organic carbon might be too simple when no additional organic carbon is supplied to the system. Over time, more difficult-to-access carbon pools would also get depleted, reducing the production rate. Our model might therefore not be appropriate for predictions on time scales of several months.

For the simulation of carbon dynamics on large time scales (e. g., in marine sediments), models have employed multiple carbon pools of different reactivity (*multi-G model*; e. g., Westrich and Berner, 1984). However, it is difficult to constrain the dynamics of these carbon pools because they do not correspond to actual chemical constituents. In the past years, metabolomics methods such as Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) have enabled better resolving the organic-matter composition of environmental samples (Bahureksa et al., 2021). The modeling approach presented by Song et al. (2020) allows to incorporate this kind of information into biogeochemical modeling. However, so far the data do not provide quantitative information, limiting their use for calibrating models with more complex representations of organic matter.

5.5 CONCLUSIONS

We have established a Bayesian workflow for gene-based modeling of nitrogen cycling. The developed likelihood function in this work is based on a Dirichlet distribution and represents a generally applicable approach to account for the dependence between measurements of different functional genes in model calibration when gene measurements are expressed as proportions of all considered genes. Sampling the posterior distribution of the non-linear biogeochemical model is challenging but model reparametrization simplifies the posterior geometry and enables efficient sampling.

Estimated uncertainties of nitrogen-cycling rates are low because tightly monitored solute concentrations contain sufficiently unique information on rates, at least in the relatively simple anoxic system under consideration. Specifically, modeling revealed that denitrification, not DNRA, is the dominant nitrate removal process.

The comparison of a model variant that uses functional-gene information against a

variant without these data shows that the additional gene data reduces the uncertainty of microbial kinetic parameters, particularly of maximum specific rate constants and growth yields. Even though the models with and without using gene data hardly differ in their estimates of solute concentrations and reaction rates for the given set-up, neither in terms of magnitude nor in terms of uncertainty, differences in parameter estimates and parametric uncertainties become important when the calibrated model is used for predictions under different conditions. For example, at lower flow rates than used for calibration, microbial growth becomes more important, and then it is essential to know the growth parameters.

The limiting factor for nitrogen cycling in the analyzed system is the availability of organic carbon because it serves as the electron donor in both denitrification and DNRA. An accurate description of the processes that supply organic carbon is, thus, important for the estimation of nitrogen-cycling rates. Reducing the flow rate and thereby increasing the residence time in the column has only a minor influence on the absolute reaction rates but it significantly increases the efficiency of nitrate removal from the water.

6 Conclusions & Outlook

6.1 SYNTHESIS OF MAJOR FINDINGS

The goal of my thesis was to advance the integration of biogeochemical models of microbial nitrogen cycling with molecular-biological data. To this end, I (1) developed model formulations the results of which can be compared to quantitative data of functional genes and transcripts, (2) explored the relationship between reaction rates and molecular-biological variables based on the models, and (3) quantified model-related uncertainties. In the following, I outline the major findings of my work, connecting them to the objectives presented in section 1.4.

6.1.1 Do enzyme-based models outperform Monod-type models?

In chapter 3, I have developed an enzyme-based model for denitrification that features a mechanistic process description for transcriptional regulation with transcription factors based on current knowledge from the model organism *Paracoccus denitrificans*. This offers an alternative approach to the enzyme-based model formulations of M. Li et al. (2017b) and Song et al. (2017) that do not depict the actual regulating mechanisms but describe the regulation based on energetic and cybernetic principles, respectively.

Including the regulation of reaction rates by enzyme levels into rate laws of biogeochemical models only improves the model if enzyme levels are actually a limiting factor for the reaction rates. For the test case in chapter 3, I showed that a Monod-type model that assumes constant enzyme levels describes the system behavior well enough, even though transcription was upregulated during the experiment. Explicitly simulating enzyme concentrations increased model complexity without improving model performance, that is the fit of the model output to the measured data. Thus, the enzyme-based model contradicts the principle of parsimony when only reaction rates are of interest. Nonetheless, it offers a description of transcript and enzyme concentrations that is useful to analyze the relationship between these quantities and reaction rates. The conclusion that an accurate representation of reaction rate regulation does not require enzyme levels cannot be generalized to all environmental conditions, or does not apply to other reactions. Considering the additional costs of enzyme-based modeling, the necessity of incorporating the regulation of reaction rates by enzymes into the model should be carefully evaluated. The additional com-

plexity makes is more difficult to analyze and calibrate the model, and leads to a higher computational effort.

Formal Bayesian model selection or cross validation help identifying the "best" model for a given purpose (e. g., predictions or system understanding) and data set (Höge et al., 2018). Formal model selection has been rarely applied to biogeochemical models, though. A notable examples is the study of Brunetti et al. (2020) who found that a Monod-type model with microbial growth can be justified over a first-order model for their nitrification experiment. Applying such model selection methods to assess the benefit of enzyme-based models over Monod-type models, or to compare different approaches for enzyme-based modeling can further advance the integration of biogeochemical modeling with molecular-biological data.

6.1.2 Can transcript measurements be used as a proxy for reaction rates?

The model-based analyses in chapters 3 and 4 highlight that transcript and enzyme concentrations are a poor proxy for reaction rates for at least two reasons:

1 Transcript concentrations react quickly to changes in the environmental conditions, but the response times of enzymes are slower. Therefore, the dynamics of transcript concentrations are decoupled from potential rates in systems with temporally changing substrate availability.

2 Reaction rates are not solely regulated by enzyme levels, but also controlled by substrate availability and inhibition effects. This implies that enzyme concentrations are not a good proxy for reaction rates either, even in a system at steady state.

While these issues have been raised before based on qualitative arguments (Moran et al., 2013), the model-based analysis in this work provides quantitative evidence and points out implications for experiments and measurement strategies.

The first point implies that a single transcript measurement can be completely unrelated to reaction rates in systems with high temporal variability. A potential strategy to deal with the short-lived nature of transcripts is to average over several transcript measurements in time. In a system with diurnal fluctuations, for example, calculating daily means is a promising strategy; more generally a weighted moving average could be used.

In chapter 4, I demonstrated that the issue of substrate limitation cannot be easily circumvented, particularly when several limiting or inhibiting compounds play a role. It requires defining a model that describes how substrate limitation and inhibiting compounds affect reaction rates, as well as knowing model parameters and concentrations of substrates and inhibitors. In that case we can no longer speak of a direct

prediction of reaction rates from transcript data, and it would be a more consistent approach to directly define a full biogeochemical model that can also predict chemical concentrations and reaction rates.

6.1.3 What kind of data do we need to better quantify reaction rates?

Molecular-biological and omics data comprise a variety of different data types, and not all of them are equally suitable for the integration with biogeochemical models and the quantification of reaction rates. Quantitative transcript and enzyme data are not only difficult to obtain but also difficult to integrate with biogeochemical models, because they require making models more complex, accounting for the production and decay processes of transcripts and enzymes. In addition, they can usually not provide a direct measure of reaction rates. In contrast, DNA-based measurements provide measures of functional biomass, which are easier to ingrate into conventional biogeochemical models without adding many parameters.

Based on my findings in chapter 4, I suggest that functional-gene data are more useful than transcript or enzyme data to constrain biogeochemical models, unless transcriptional regulation is known to be an influential factor for reaction rates, and extensive data is available. However, transcript and enzyme data can be useful to indicate qualitatively which reaction pathways are potentially active and should be integrated in the model. In any case, chemical data of substrates, intermediates, reaction products and inhibiting compounds are essential to constrain reaction rates and should be monitored with the appropriate spatial and temporal resolution.

6.1.4 Do functional-gene and transcript data reduce parameter uncertainty?

One potential benefit of using microbiological data for biogeochemical models is to better constrain parameter values. However, modeling studies that use functional gene and transcript data have not addressed this question so far. In chapter 3 I have shown that the posterior uncertainty of reaction parameters (more specifically, maximum-rate and inhibition constants for denitrification) is not reduced in the enzyme-based model that uses transcript data compared to the Monod-type model.

In contrast, including relative-quantitative measurements of functional genes for nitrogen cycling reduced the uncertainty of several parameters in the gene-based model in chapter 5. In particular, the gene data helped to constrain the posterior distributions of maximum specific reaction rates, half-saturation constants and growth yields.

Both transcript and functional-gene data have had little effect on the uncertainty of reaction rates. This can be explained by the strong constraints that solute data put on the reaction rates in the considered systems. As a result, the uncertainty of

reaction rates is already low without using the transcript and gene data. However, the small uncertainty estimates of reaction rates are likely overconfident due to structural model errors. The "true" uncertainty (accounting for structural model uncertainty) might be much higher. In that case the benefit of functional-gene data for reducing the uncertainty of reaction rates might be higher than it seems from my analysis that neglects structural model errors. More work is therefore required to obtain reliable uncertainty estimates for biogeochemical models in view of structural model errors.

6.1.5 Addressing the challenges of calibrating gene- and enzyme-based models

My simulations have shown that the multi-dimensional posterior distributions of gene- and enzyme-based models can be very complex, with strong correlations or non-linear dependencies between reaction parameters. This is a phenomenon known for biogeochemical models in general (Sierra et al., 2015) and can be related to the "sloppy" characteristics of the models (Marschmann et al., 2019). Complicated posterior geometries can pose a practical problem for effectively sampling the distribution. While MCMC algorithms theoretically converge to the posterior distribution with an infinite number of samples, the approximation of "infinite" can be prohibitively expensive when the sampler struggles to explore the posterior. One step towards addressing this problem for me was to use a state-of-the-art Hamiltonian Monte Carlo sampler for the reaction model in chapter 3 that can explore the posterior distribution more efficiently, and provides better diagnostics for pathological behavior (Betancourt, 2018; Monnahan et al., 2017). But even advanced sampling algorithms have difficulties sampling the complicated posterior landscape. Fortunately, reparametrization can help achieving convergence, as the simulations in chapters 3 and 5 have shown. In particular, I have introduced a reparameterization for the commonly used Monod or Michaelis-Menten rate law that can remove or reduce correlations between the maximum rate parameter and half-saturation constants.

6.2 RESEARCH PERSPECTIVES

Given the amount of molecular-biological and omics data retrieved from the environment in the past decade, biogeochemical models (including the ones in this thesis) still can use merely a fraction of the information that these data contain. In my view, the main reasons for this are (1) the complex nature of microbial reaction processes, (2) the challenges that are inherent to models describing these processes (e. g., non-identifiability, computational cost), and (3) the fact that molecular-biological and omics data are often not (yet) quantitative. In the following, I briefly discuss some fields of research and potential challenges that need to be addressed in order to close the gap. It requires further development of both models and omics methods.

6.2.1 Better quantification of microbes through advances in omics techniques

The fast development of omics methods in the past decades has revolutionized our understanding of the diversity of microorganisms, their metabolism and microbial ecology. Still, an accurate quantification of genes, transcripts, enzymes and metabolites remains a challenge. To date, most omics methods provide only qualitative or semi-quantitative information about the microorganisms in the sample. However, further advances in omics techniques promise to provide quantitative data that are more suitable for the integration with biogeochemical models. For example, recent studies have shown that spiking samples with an internal DNA standard enables the absolute quantification of microbial taxa, functional genes and transcripts via 16S rRNA sequencing (Jiang et al., 2019; Zemb et al., 2020), metagenomics (Crossette et al., 2021) and metatranscriptomics (Delogu et al., 2020; Gifford et al., 2011). The advantage of these new methods is that they allow quantifying many genes at the same time – in contrast to qPCR, which is currently a standard method for the absolute quantification of genes or transcripts but which is limited to a few selected genes.

Another recently developed method, bioorthogonal non-canonical amino acid tagging combined with fluorescence-activated cell sorting (BONCAT-FACS), allows to distinguish translationally active cells from inactive cells (Couradeau et al., 2019; Hatzenpichler et al., 2016) prior to genomic analyses such as metagenomics or 16S rRNA gene sequencing. This allows to exclude DNA from dormant microorganisms and relic DNA, enhancing the usability of genomic data as an estimate of active microbial biomass. The method has so far been tested on a range of environmental samples including soils (Couradeau et al., 2019; Marlow et al., 2021), deep-sea sediments (Hatzenpichler et al., 2016) and a coal seam (Schweitzer et al., 2022).

Overall, these developments will help to make functional-gene measurements a more reliable estimate for functional biomass that can be compared with the outputs from biogeochemical models.

6.2.2 Trait-based and genome-scale metabolic models

Most biogeochemical models describe the microbial community as a single entity that represents some kind of average behavior of all organisms. Even if gene-centric models account for dynamics of biomass with different metabolic capabilities they do not resolve ecophysiological differences between organisms that share the same function. Therefore, they cannot describe the evolution of a microbial community through, for example, competition or symbiotic relationships. The bulk description of microbial processes also has the consequence that the parameters are effective parameters and do not necessarily correspond to measurable biological properties. As such they must be calibrated with experimental data of the respective site. Several modeling

approaches try to better resolve microbial processes and, thus, eliminate the need for calibration of reaction parameters. Here, I want to briefly discuss the potential of trait-based modeling using genomic information and genome-scale metabolic models for microbial communities.

Trait-based models explicitly represent multiple groups of organisms with different ecophysiological traits and, as a consequence, different reaction parameters such as maximum growth rates or substrate affinities (see, for example, Bouskill et al., 2012). The community composition and, hence, the average reaction parameters are an emergent property of trait-based models. A major challenge of trait-based modeling is to determine the reaction parameters for each guild, particularly considering that most environmental microorganisms cannot be cultivated. However, recent modeling approaches try to directly translate genomic information like the frequency of certain amino acids in the genome into physiological parameters such as the optimum growth temperature or maximum growth rates (Cheng et al., 2018; Sokol et al., 2022). The obtained parameter values can then be used in trait-based biogeochemical models (Cheng et al., 2018).

Genome-scale metabolic models have mostly been developed for single organisms so far. This is not suitable to simulate nitrogen cycling which relies on a community of organisms with different functions. However, modern bioinformatics tools allow assembling genomes from metagenomic data, providing information about the potential metabolisms of the organisms in an environmental sample. In addition, the construction of genome-scale metabolic models can now be partly automated, even though manual curation is currently still necessary, presenting one of the major bottlenecks. Together, these developments allow to construct genome-scale metabolic models for whole communities from metagenomic data (Frioux et al., 2020). For example, Rubinstein et al. (2022) have used a genome-scale metabolic model derived from metagenome data to obtain fixed reaction stoichiometries of nitrification and denitrification that they subsequently used in a conventional reactive transport model. However, a tighter coupling between community-scale metabolic models and reactive transport is also thinkable: reaction rates could be directly derived from dynamic predictions of the genome-scale metabolic model as in the early work by Scheibe et al. (2009). Even though genome-scale metabolic models for whole communities are still in their infancy, they promise to be a valuable tool for quantitatively studying the interactions of nitrogen cycling organisms.

Trait-based and genome-scale metabolic models allow for a more fine-grained representation of the microbial metabolisms and communities that is directly based on genomic information. They might have the potential to eliminate or reduce the need of calibrating reaction parameters. However, they are still simplified representations of the actual processes, and it needs to be tested whether they can indeed reproduce observations without tuning reaction parameters. Model results from the studies

applying these new modeling concepts have not been validated against data (Cheng et al., 2018; Rubinstein et al., 2022). Validation requires developing approaches for the comparison of a model-predicted community composition with (semi-)quantitative data, for instance from 16S rRNA sequencing. This is not a trivial task, because the selected group of organisms represented in the model comprises only a fraction of the actual microbial community, and it is not evident how simulated and measured abundances should be matched.

6.2.3 *Improving uncertainty quantification of biogeochemical models*

Mechanistic biogeochemical models of microbial reactions are simplistic, compared to natural biogeochemical complexity, and "sloppy" (Marschmann et al., 2019). This means that the effective variability of outputs is much lower than could be expected from the number of parameters, because there are only few parameter *combinations* that produce a change in the model behavior whereas other parameter combinations are essentially unconstrained (Gutenkunst et al., 2007). As a result, it is often both difficult to uniquely identify parameters (due to the sloppiness), *and* difficult to match all patterns present in the data because the model is missing yet unidentified processes.

Quantifying the uncertainty of a model is particularly important when parameters cannot be uniquely identified and when there is uncertainty about the model structure. However, Bayesian uncertainty quantification relies on the assumption that the data were actually generated by the model used. Structural model errors lead to biased parameter estimates and an underestimation of uncertainty (O'Hagan, 2013; Sargsyan et al., 2019). Reliably estimating uncertainty therefore requires adopting modeling concepts that explicitly account for model errors.

A common approach is to add a bias term to model outputs that is described with a Gaussian process (M. C. Kennedy and O'Hagan, 2001; Xu et al., 2017). While this can improve predictions and uncertainty estimates for observed variables, it does not correct the uncertainty estimates of unobserved variables, nor does it help identifying the causes for model structural error (Reichert and Mieleitner, 2009). Other approaches try to integrate the representation of structural error more tightly with the model, for example by treating model parameters as time-variable, stochastic processes (Reichert and Mieleitner, 2009) or by defining model parameters as random variables and estimating their density instead of parameter values (Sargsyan et al., 2019, 2015).

Few examples (Pan et al., 2020; Sun et al., 2021) for explicit error modeling exist in the reactive transport modeling literature. Adopting the existing statistical approaches to biogeochemical modeling promises to improve parameter and uncertainty estimates. One of the major questions that need to be investigated is how the "black box" statistical error models can still be useful to enhance system understanding

and improve process-based models. In this regard, the second approach for model structural errors that places the model error representation *inside* the process-based model seems promising for biogeochemical models because it would allow to obtain consistent uncertainty estimates of the unobserved reaction rates, and to relate model structural error to specific reaction parameters.

6.2.4 Field-scale models integrating molecular-biological data

So far, models that integrate functional-gene, transcript or enzyme data have often been tested on a lab scale, for example with batch or column experiments (Chavez Rodriguez et al., 2020; M. Li et al., 2017b; Murray et al., 2019; Pagel et al., 2016; Song et al., 2017). The studies by Hui et al. (2021) addressing nitrogen cycling in a river channel and Louca et al. (2016) and Reed et al. (2014) in oceanic systems are an exception.

There is a need to extend combined modeling and field studies integrating molecular-biological data to subsurface environments such as groundwater and soils. These environments are linked to a number of additional challenges compared to lab-scale studies. For example, reactivity and microbial abundance can be highly heterogeneous in space ("hot spots", Kuzyakov and Blagodatskaya, 2015; Sawyer, 2015). Limited access to the subsurface also makes sampling difficult. Additionally, relic DNA may distort estimates of functional biomass (Carini et al., 2016), multi-dimensional transport processes interact with reactions, initial and boundary conditions are often poorly known, and the reactive system is much more complex than in a controlled laboratory experiment, because substrates are less well defined.

Even though molecular-biological tools have also been applied at the field scale, the collected data sets are often not optimal for the integration with reactive transport models considering these challenges. They cover only few points in space and time, often do not provide quantitative information about functional genes or transcripts, or lack complementary chemical or hydrological data that is necessary to set-up and calibrate a reactive transport model.

To address the challenges of understanding microbial reactions at the field scale, it takes an iterative approach of collecting data in the field, in the lab and modeling (Lui et al., 2021). Characterizing the microbial community and potentially active reaction pathways with qualitative omics methods and sparse chemical data would only be a first step. Subsequent lab-scale experiments combined with process-based models could then be used to better understand the reactive system and identify variables that need close monitoring in the field. Next, the hydrological functioning of the system needs to be studied in order to parameterize flow and boundary conditions. A preliminary model of the field system could then be used to determine optimal measurement locations and the required temporal resolution. Finally, spatially and

temporally resolved, quantitative measurements of functional genes and chemical data need to be collected in order to calibrate a field-scale reactive transport model integrating molecular-biological data.

Multi-dimensional field-scale reactive transport models are computationally much more intensive than spatially zero- or one-dimensional models for the lab scale. Moreover, they comprise transport parameters that need to be calibrated in addition to reaction parameters, increasing the dimensionality of the inverse problem further. This can make rigorous parameter estimation and uncertainty quantification impractical or even impossible. *Surrogate* or *proxy models* provide a fast approximation for the outputs of the complex (mechanistic) model, given a set of input parameters. They can be based on (1) data-driven approaches such as polynomial chaos expansion (PCE), artificial neural networks (ANNs) or Gaussian processes (GPs), to name a few, (2) projections of the model onto a lower-dimensional space, or (3) lower-fidelity approximations, for instance, by reducing the numerical resolution (Asher et al., 2015). Relatively few studies have used surrogates for reactive transport modeling so far (e. g., Dietzel and Reichert, 2014; Y. Li et al., 2021; Scheurer et al., 2021; Zhou et al., 2018). Developing surrogate modeling strategies for gene-based reactive transport models is necessary in order to more easily simulate the full microbial nitrogen-cycling network and assess the related uncertainties on a field scale.

A Supporting information for chapter 2

A.1 Solving the enzyme production equation

The equation describing the production and decay of enzymes (equation (2.9)) is a linear, inhomogeneous ODE. We can rearrange it as follows:

$$\frac{\mathrm{d}E}{\mathrm{d}t} + k^E_{\mathrm{dec}} E = k_{\mathrm{trnsl}} T = q(t)\,, \tag{A.1}$$

with the initial condition

$$E(0) = E_0\,. \tag{A.2}$$

Linear ODEs can be solved applying the Laplace transformation as outlined in Dyke (2001). We make use of the property that the Laplace transform $\mathscr{L}$ of the derivative of a function f is

$$\mathscr{L}\left\{\frac{\mathrm{d}f}{\mathrm{d}t}\right\} = s\tilde{f}(s) - f(0)\,, \tag{A.3}$$

where $\tilde{f}$ is the Laplace transform of f, and s is the complex frequency parameter of the Laplace transform.

We start by applying the Laplace transform to equation (A.1):

$$\mathscr{L}\left\{\frac{\mathrm{d}E}{\mathrm{d}t}\right\} + \mathscr{L}\left\{k^E_{\mathrm{dec}} E\right\} = \mathscr{L}\left\{q(t)\right\}\,, \tag{A.4}$$

which can be rewritten to

$$s\tilde{E}(s) - E(0) + k^E_{\mathrm{dec}}\tilde{E}(s) = \tilde{q}(s) \tag{A.5}$$

using equation (A.3). Equation (A.5) is an algebraic equation that can be solved for $\tilde{E}$, yielding

$$\tilde{E} = \frac{\tilde{q}(s) + E_0}{s + k^E_{\mathrm{dec}}}\,. \tag{A.6}$$

Next, we apply the back-transform and obtain

$$E = \mathscr{L}^{-1}\left\{\frac{\tilde{q}(s) + E_0}{s + k^E_{\mathrm{dec}}}\right\} \tag{A.7}$$

$$= \mathscr{L}^{-1}\left\{\frac{\tilde{q}(s)}{s + k^E_{\mathrm{dec}}}\right\} + \mathscr{L}^{-1}\left\{\frac{E_0}{s + k^E_{\mathrm{dec}}}\right\}\,, \tag{A.8}$$

where the second equality holds due to linearity of $\mathscr{L}^{-1}$.

The convolution of two functions $f(t)$ and $g(t)$ in time domain corresponds to a multiplication in Laplace domain:

$$f(t) * g(t) = \mathscr{L}^{-1}\left\{\tilde{f}\tilde{g}\right\}, \tag{A.9}$$

where

$$f * g = \int_0^t f(\tau)g(t-\tau)\,\mathrm{d}\tau \tag{A.10}$$

Furthermore, it is known that

$$\mathscr{L}^{-1}\left\{\frac{1}{s+k}\right\} = \exp(-kt)\,. \tag{A.11}$$

Applying equations (A.9) and (A.11), we can write equation (A.8) as follows:

$$E(t) = q(t) * \exp(-k_{\mathrm{dec}}^E t) + E_0 \exp(-k_{\mathrm{dec}}^E t)\,. \tag{A.12}$$

Inserting the definition of the convolution integral (equation (A.10)), this becomes

$$E(t) = \int_0^t q(\tau)\exp\left(-k_{\mathrm{dec}}^E(t-\tau)\right)\mathrm{d}\tau + E_0\exp(-k_{\mathrm{dec}}^E t)\,. \tag{A.13}$$

Substituting the definition of $q(t)$ yields

$$E(t) = k_{\mathrm{trnsl}}\int_0^t T(\tau)\exp\left(-k_{\mathrm{dec}}^E(t-\tau)\right)\mathrm{d}\tau + E_0\exp(-k_{\mathrm{dec}}^E t)\,, \tag{A.14}$$

which is the same as equation (2.10).

B Supporting information for chapter 3

B.1 INITIAL VALUES

Table B.1 lists the initial values of all solute and gas concentrations. Transcription factor concentrations are initialized based on their quasi-steady state concentrations and, thus, depend on the values of the regulation parameters.

TABLE B.1: Initial concentration values used for the simulation.

Substance	Value	Units	Reference
Nitrate	2×10^{-3}	M	known from experimental set-up
Nitrite	0	M	known from experimental set-up
N_2 in the gas phase	5.2×10^{-5}	M	measurement data
N_2 in water	8.7×10^{-7}	M	equilibrium with the gas phase
O_2 in the gas phase	2.9×10^{-3}	M	known from experimental set-up
O_2 in water	9.3×10^{-5}	M	equilibrium with the gas phase
Cells	variable	cells L^{-1}	estimation parameter (see table B.2)
active fraction of FnrP/NarR/NNR	variable	dimensionless	quasi-steady state concentration
NAR/NIR enzymes	0	M	assumption

B.2 PRIOR AND POSTERIOR PARAMETER DISTRIBUTIONS

Figure B.1 shows the prior and posterior distributions of all parameters related to the ODEs used in the Monod-type and enzyme based model. Table B.2 lists the values of fixed parameters and the prior distributions of the fitted parameters. All

parameters related to transcript and enzyme concentrations were estimated with the exception of the maximum enzyme concentration β_E. Figures B.2 and B.3 show pairwise correlation coefficients of all parameters in the Monod-type and enzyme-based models, respectively.

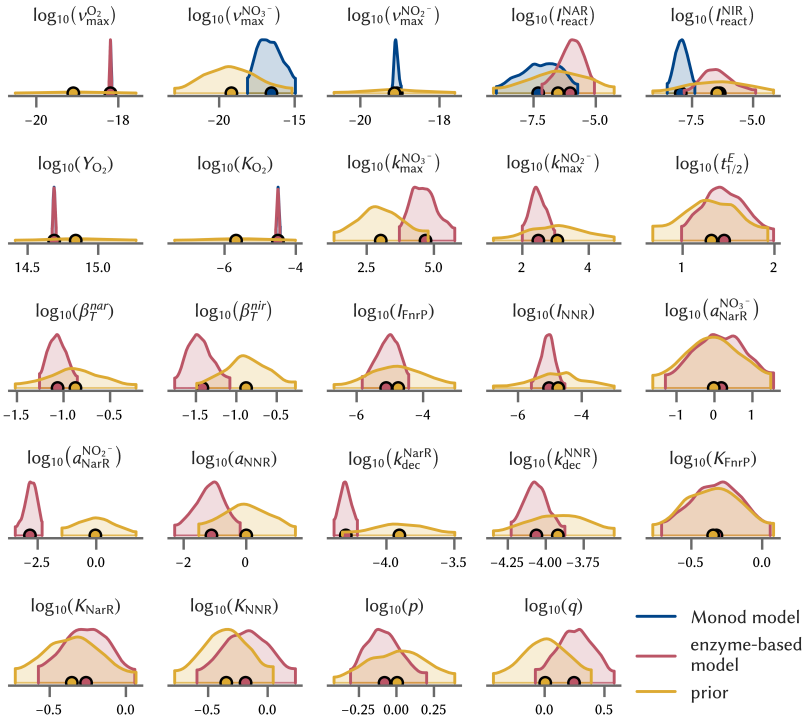


FIGURE B.1: Kernel density estimates of the marginal posterior and prior distributions for all model parameters. Densities are cut off at the 94 % highest density intervals, circular markers indicate the mean.

TABLE B.2: Simulation parameters, their prior distributions or fixed values, and their posterior medians and percentiles.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles			Model[b]
					10 %	Median	90 %	
$t_{1/2}^{E}$	enzyme half life	h	$N(3, 0.8)$	f, g	1.3×10^{1}	2.8×10^{1}	6.4×10^{1}	E
β_T^{nar}	maximum *narG* concentration	transcripts cell^{-1}	$T(-2, 0.7, 10)$	f, h	6.3×10^{-2}	8.5×10^{-2}	1.2×10^{-1}	E
β_T^{nir}	maximum *nirS* concentration	transcripts cell^{-1}	$T(-2, 0.7, 10)$	f, h	2.3×10^{-2}	3.5×10^{-2}	6.7×10^{-2}	E
$I_{\text{trs}}^{\text{FnrP}}$	O_2 inhibition constant of FnrP	M	$T(-11, 2, 10)$	c	2.3×10^{-6}	8.6×10^{-6}	2.2×10^{-5}	E
$I_{\text{trs}}^{\text{NNR}}$	O_2 inhibition constant of NNR	M	$T(-11, 2, 10)$	c	4.4×10^{-6}	8.9×10^{-6}	1.8×10^{-5}	E
p	Hill coefficient for FnrP	–	$N(0, 0.5)$		5.7×10^{-1}	8.2×10^{-1}	1.2	E
q	Hill coefficient for NNR	–	$N(0, 0.5)$		1.0	1.8	2.9	E
$a_{\text{NarR}}^{NO_3^-}$	NarR activation constant for NO_3^-	s^{-1} M^{-1}	$N(0, 2)$	c	1.6×10^{-1}	1.4	1.5×10^{1}	E
$a_{\text{NarR}}^{NO_2^-}$	NarR activation constant for NO_2^-	s^{-1} M^{-1}	$N(0, 2)$	c	5.6×10^{-4}	1.6×10^{-3}	3.4×10^{-3}	E
a_{NNR}	NNR activation constant for NO_2^-	s^{-1} M^{-1}	$N(0, 2)$	c	1.4×10^{-2}	7.7×10^{-2}	3.8×10^{-1}	E
$k_{\text{dec}}^{\text{NarR}}$	dissociation constant of NarR	s^{-1}	$N(-9, 0.5)$		4.4×10^{-5}	5.0×10^{-5}	5.8×10^{-5}	E
$k_{\text{dec}}^{\text{NNR}}$	dissociation constant of NNR	s^{-1}	$N(-9, 0.5)$	k	6.7×10^{-5}	8.6×10^{-5}	1.1×10^{-4}	E

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters except for σ_i. $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] E means *enzyme-based model* and M means *Monod-type model* [c] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [d] Reparametrization of v_{max}^{i}, see appendix B.4 for details. [e] Based on experimental data. [f] Maier et al. (2011) [g] Roberts et al. (2011) [h] Qu et al. (2015) [i] Suenaga et al. (2018) [j] Hassan et al. (2016) [k] Lee et al. (2006)

TABLE B.2: (continued) Simulation parameters, their prior distributions or fixed values, and their posterior medians and percentiles.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles			Model[b]
					10 %	Median	90 %	
K_{FnrP}	half-saturation constant for binding of FnrP	–	$N(-0.8, 0.5)$		2.5×10^{-1}	4.8×10^{-1}	8.7×10^{-1}	E
K_{NarR}	half-saturation constant for binding of NarR	–	$N(-0.8, 0.5)$		3.3×10^{-1}	5.5×10^{-1}	9.0×10^{-1}	E
K_{NNR}	half-saturation constant for binding of NNR	–	$N(-0.8, 0.5)$		3.4×10^{-1}	6.6×10^{-1}	1.3	E
$I_{\mathrm{reac}}^{\mathrm{NAR}}$	O_2 inhibition parameter for reaction of NO_3^-	M	$N(-15, 3)$		3.0×10^{-9} 1.6×10^{-7}	5.9×10^{-8} 1.0×10^{-6}	6.1×10^{-7} 4.3×10^{-6}	M E
$I_{\mathrm{reac}}^{\mathrm{NIR}}$	O_2 inhibition parameter for reaction of NO_2^-	M	$N(-15, 3)$		4.6×10^{-9} 6.1×10^{-8}	1.1×10^{-8} 3.4×10^{-7}	2.6×10^{-8} 4.4×10^{-6}	M E
$K_{\mathrm{O_2}}$	O_2 half-saturation constant	M	$T(-13, 2, 10)$	i	2.9×10^{-5} 2.9×10^{-5}	3.2×10^{-5} 3.1×10^{-5}	3.5×10^{-5} 3.4×10^{-5}	M E
$K_{\mathrm{NO_3^-}}$	NO_3^- half-saturation constant	M	5×10^{-6}	j		fixed value		E, M
$K_{\mathrm{NO_2^-}}$	NO_2^- half-saturation constant	M	5×10^{-6}	j		fixed value		E, M
$Y_{\mathrm{O_2}}$	O_2 growth yield per amount e^--donor	cells mol^{-1}	$T(32.2, 0.5, 10)$	j	4.8×10^{14} 4.8×10^{14}	4.9×10^{14} 4.9×10^{14}	5.0×10^{14} 5.0×10^{14}	M E

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters except for σ_i. $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] E means *enzyme-based model* and M means *Monod-type model* [c] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [d] Reparametrization of v_{max}^i, see appendix B.4 for details. [e] Based on experimental data. [f] Maier et al. (2011) [g] Roberts et al. (2011) [h] Qu et al. (2015) [i] Suenaga et al. (2018) [j] Hassan et al. (2016) [k] Lee et al. (2006)

TABLE B.2: (continued) Simulation parameters, their prior distributions or fixed values, and their posterior medians and percentiles.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles			Model[b]
					10 %	Median	90 %	
$v_{\text{fix}}^{O_2}$	O_2 reaction rate at fixed O_2 concentration	mol cell^{-1} s^{-1}	$T(-44.2, 1.6, 10)$	d	4.1×10^{-19} 4.1×10^{-19}	4.2×10^{-19} 4.2×10^{-19}	4.3×10^{-19} 4.3×10^{-19}	M E
$v_{\text{fix}}^{NO_2^-}$	NO_2^- reaction rate at fixed O_2 concentration	mol cell^{-1} s^{-1}	$T(-44.2, 1.7, 10)$	d	2.3×10^{-22}	4.8×10^{-22}	1.1×10^{-21}	M
$v_{\text{fix}}^{NO_3^-}$	NO_3^- reaction rate at fixed O_2 concentration	mol cell^{-1} s^{-1}	$T(-49.4, 3.3, 10)$	d	1.7×10^{-20}	7.1×10^{-20}	5.0×10^{-19}	M
$k_{\text{max}}^{NO_3^-}$	NAR turnover number	s^{-1}	$T(7, 2, 10)$		7.5×10^{4}	2.5×10^{5}	1.4×10^{6}	E
$k_{\text{max}}^{NO_2^-}$	NIR turnover number	s^{-1}	$T(7, 2, 10)$		1.3×10^{2}	2.0×10^{2}	3.8×10^{2}	E
β_E	maximum enzyme concentration	enzymes cell^{-1}	1125	f		fixed value		E
B_0	initial cell concentration	cells L^{-1}	$T(21, 0.5, 10)$	e	3.0×10^{9} 3.0×10^{9}	3.3×10^{9} 3.3×10^{9}	3.5×10^{9} 3.5×10^{9}	M E
b_{O_2}	background value of O_2	M	$N(-18, 3)$		8.1×10^{-8} 8.2×10^{-8}	9.5×10^{-8} 9.5×10^{-8}	1.1×10^{-7} 1.1×10^{-7}	M E
$b_{NO_2^-}$	background value of NO_2^-	M	$N(-18, 3)$		3.6×10^{-10} 3.0×10^{-10}	1.0×10^{-8} 1.2×10^{-8}	1.3×10^{-7} 1.2×10^{-7}	M E
σ_B	constant error of Box-Cox transformed cell densities		Half-Normal(0.1)		2.2×10^{-1} 2.2×10^{-1}	2.6×10^{-1} 2.6×10^{-1}	3.1×10^{-1} 3.2×10^{-1}	M E

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters except for σ_i. $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] E means *enzyme-based model* and M means *Monod-type model* [c] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [d] Reparametrization of v_{max}^i, see appendix B.4 for details. [e] Based on experimental data. [f] Maier et al. (2011) [g] Roberts et al. (2011) [h] Qu et al. (2015) [i] Suenaga et al. (2018) [j] Hassan et al. (2016) [k] Lee et al. (2006)

TABLE B.2: (continued) Simulation parameters, their prior distributions or fixed values, and their posterior medians and percentiles.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles			Model[b]
					10 %	Median	90 %	
σ_{O_2}	constant error of Box-Cox transformed O_2 gas data		$H(0, 0.1)$		2.5×10^{-2} 2.6×10^{-2}	2.9×10^{-2} 2.9×10^{-2}	3.3×10^{-2} 3.4×10^{-2}	M E
$\sigma_{NO_2^-}$	constant error of Box-Cox transformed NO_2^- data		$H(0.025, 0.1)$		2.5×10^{-2} 2.5×10^{-2}	2.5×10^{-2} 2.5×10^{-2}	2.6×10^{-2} 2.6×10^{-2}	M E
σ_{N_2}	constant error of Box-Cox transformed N_2 gas data		$H(0.1, 0.1)$		1.0×10^{-1} 1.0×10^{-1}	1.0×10^{-1} 1.0×10^{-1}	1.0×10^{-1} 1.0×10^{-1}	M E
σ_{nar}	constant error of Box-Cox transformed *nar* mRNA data		$H(0, 0.1)$		8.7×10^{-3}	4.0×10^{-2}	9.9×10^{-2}	E
σ_{nir}	constant error of Box-Cox transformed *nir* mRNA data		$H(0, 0.1)$		3.1×10^{-2}	4.2×10^{-2}	5.7×10^{-2}	E
λ_B	Box-Cox transformation parameter of the cell densities		5×10^{-2}			fixed value		E, M
$\lambda_{NO_2^-}$	Box-Cox transformation parameter of the NO_2^- data		4×10^{-1}			fixed value		E, M
λ_{N_2}	Box-Cox transformation parameter of the N_2 data		1.8×10^{-1}			fixed value		E, M
λ_{O_2}	Box-Cox transformation parameter of the O_2 data		1.8×10^{-1}			fixed value		E, M

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters except for σ_i. $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] E means *enzyme-based model* and M means *Monod-type model* [c] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [d] Reparametrization of v^i_{max}, see appendix B.4 for details. [e] Based on experimental data. [f] Maier et al. (2011) [g] Roberts et al. (2011) [h] Qu et al. (2015) [i] Suenaga et al. (2018) [j] Hassan et al. (2016) [k] Lee et al. (2006)

TABLE B.2: (continued) Simulation parameters, their prior distributions or fixed values, and their posterior medians and percentiles.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles			Model[b]
					10 %	Median	90 %	
λ_{nar}	Box-Cox transformation parameter of the *nar* mRNA data		10^{-1}			fixed value		E
λ_{nir}	Box-Cox transformation parameter of the *nir* mRNA data		4×10^{-1}			fixed value		E

[a] Prior distributions are defined for the log of all parameters except for σ_i. $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] E means *enzyme-based model* and M means *Monod-type model* [c] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [d] Reparametrization of $v^i_{\max}$, see appendix B.4 for details. [e] Based on experimental data. [f] Maier et al. (2011) [g] Roberts et al. (2011) [h] Qu et al. (2015) [i] Suenaga et al. (2018) [j] Hassan et al. (2016) [k] Lee et al. (2006)

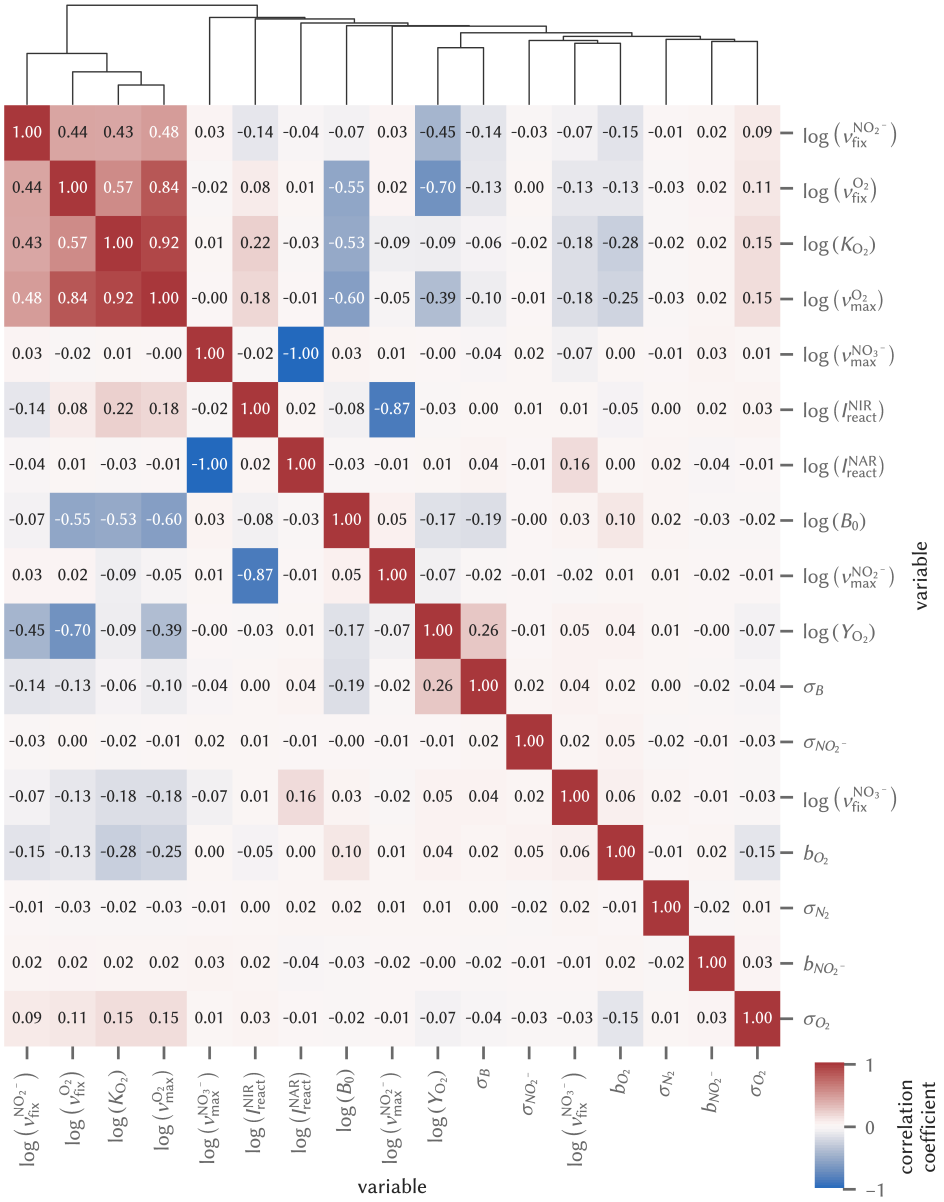


FIGURE B.2: Hierarchically-clustered heatmap showing the correlation coefficients of the estimated parameters in the posterior distribution of the Monod-type model.

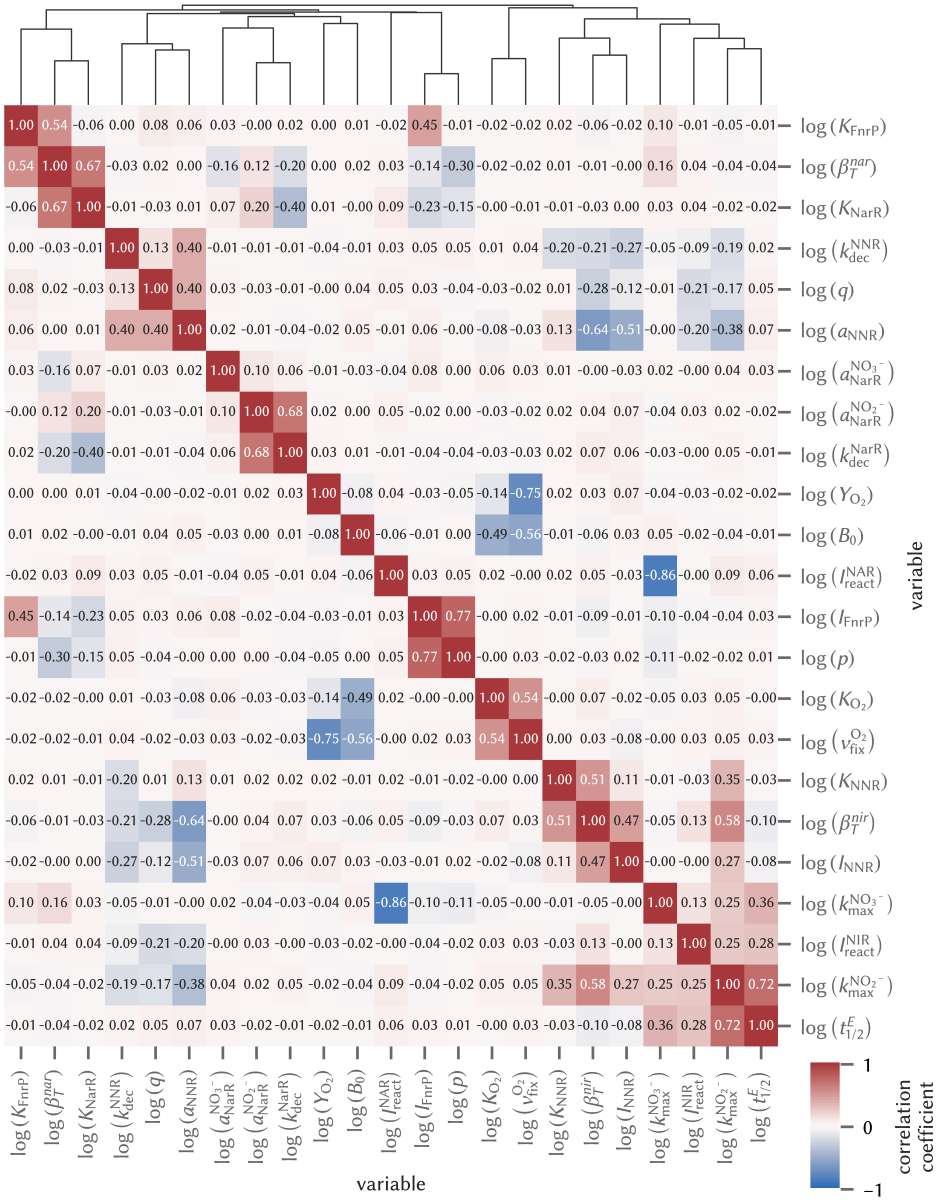


FIGURE B.3: Hierarchically-clustered heatmap showing the correlation coefficients of the estimated parameters in the posterior distribution of the enzyme-based model.

B.3 TRANSCRIPTION FACTOR CONCENTRATIONS

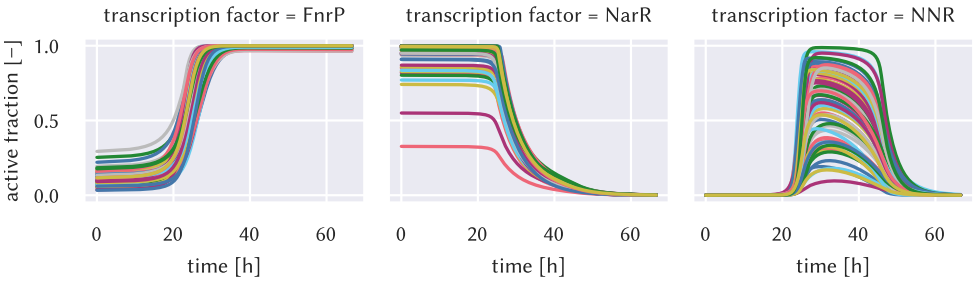


FIGURE B.4: Simulated fraction of active transcription factors over time. Several draws from the posterior are indicated by color.

B.4 REPARAMETRIZATION OF MONOD PARAMETERS

The Monod parameters of aerobic respiration, K_{O_2} and $\nu_{max}^{O_2}$, showed a strong correlation in the posterior. Heavy parameter correlations can hinder effective sampling from the posterior. We therefore decided to apply a reparametrization that describes the same rate law in terms of different parameters. $\nu_{max}^{O_2}$ can be interpreted as the cell-specific reaction rate at the limit of the oxygen concentration going to infinity. Instead of using the rate at the limit, we use the rate at a fixed finite oxygen concentration, $c_{O_2}^{fix}$:

$$\nu_{fix}^{O_2} = \nu_{max}^{O_2} \frac{c_{O_2}^{fix}}{c_{O_2}^{fix} + K_{O_2}} . \tag{B.1}$$

We can solve equation (B.1) for the maximum cell specific rate $\nu_{max}^{O_2}$, expressing it in terms of K_{O_2}, $c_{O_2}^{fix}$ and $\nu_{fix}^{O_2}$. We chose $c_{O_2}^{fix}$ manually such that the correlation between $\nu_{fix}^{O_2}$ and K_{O_2} is low.

We apply a similar reparametrization for the maximum cell-specific denitrification rates and oxygen inhibition constants in the Monod-type model. The maximum cell-specific rate ν_{max}^{i} of nitrogen substrate i is the rate at the limit of the substrate concentration reaching infinity and an oxygen concentration of zero. Instead, we use the cell-specific rate at a fixed, non-zero oxygen concentration $c_{O_2}^{fix,i}$ and without substrate limitation (i. e. $c_i \to \infty$) as parameter:

$$\nu_{fix}^{i} = \nu_{max}^{i} \frac{I_{reac}^{i}}{I_{reac}^{i} + c_{O_2}^{fix,i}} . \tag{B.2}$$

B.5 MASS TRANSFER TO THE GAS PHASE AND GAS SAMPLING

N_2 and O_2 are gases that partition into the headspace. The flasks used in the experiment had a headspace. Therefore, the model needs to account for the partitioning of molecular nitrogen and oxygen between the water and gas phases. Henry's law describes the relation between aqueous and gas phase concentrations of compound i (c^i_w and c^i_g) at equilibrium:

$$c^i_g = H_i c^i_w \,, \tag{B.3}$$

with the Henry's-law coefficient $H_i[\mathrm{mol\,L^{-1}_{gas}\,mol^{-1}\,L_{liq}}]$. Kinetic mass transfer of compound i between the two phases is described by a linear driving force model with rate coefficient $k^i_{tr}\left[\mathrm{s^{-1}}\right]$. The transfer rate r^i_{tr} $[\mathrm{mol\,L^{-1}_{liq}\,s^{-1}}]$, expressed as the mass-transfer related concentration change in the liquid phase, is then given by

$$r^i_{tr} = k^i_{tr}\left(\frac{c^i_g}{H_i} - c^i_w\right). \tag{B.4}$$

In the experiment, the head space was sampled in regular intervals. Helium was subsequently injected to compensate the pressure decrease resulting from the removal of gas. This led to a dilution of the gas-phase concentrations, which we describe by a first order rate law:

$$r^i_{sample} = f_{dil} k_{sample}(t)\, c^i_g \,, \tag{B.5}$$

in which f_{dil} is the fraction of gas exchanged at each sampling event. The function $k_{sample}(t)[\mathrm{s^{-1}}]$ is a scaled rate constant. It would most appropriately be described by several pulses, each integrating to one. To avoid discontinuities we replaced it by a linear function with approximately the same integral (figure B.5).

Following Qu et al. (2015) we also account for small leakage rates r^i_{leak} of N_2 and O_2 into the system, with a constant diffusion rate r_{diff} $[\mathrm{mol\,L^{-1}_{gas}\,s^{-1}}]$ and a time-dependent rate that accounts for higher leakage during sampling:

$$r^i_{leak} = r^i_{diff} + k_{sample}(t) c^i_{leak} \,, \tag{B.6}$$

in which c^i_{leak} $[\mathrm{mol\,L^{-1}_{gas}}]$ is the concentration increase of compound i during a single sampling event. The rate of change of gas-phase concentrations is then given by

$$\frac{\mathrm{d}c^i_g}{\mathrm{d}t} = r^i_{leak} - r^i_{sample} - \frac{V_w}{V_g} r^i_{tr} \,, \tag{B.7}$$

with V_g and V_w being the volume of the headspace and the liquid phase, respectively.

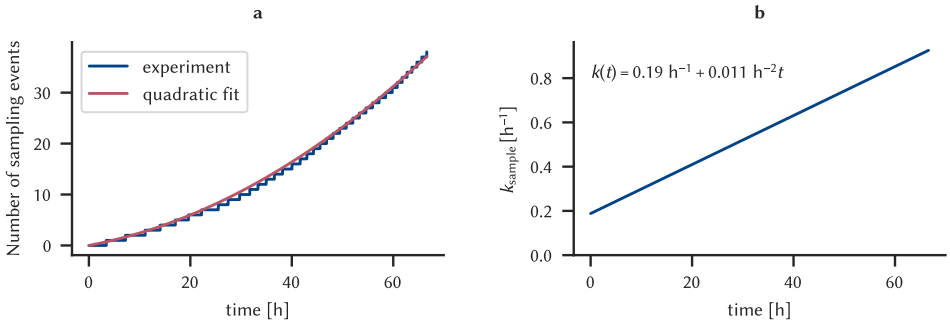


FIGURE B.5: Determination of the time-dependent gas sampling rate coefficient k_{sample}. **a** Quadratic fit to the number of sampling events in the experiment. **b** The derivative of the function gives the time-dependent coefficient for sampling related rates k_{sample}.

TABLE B.3: Parameters related to mass transfer and gas sampling with their values and units.

Symbol	Description	Value	Units	Note
k_{tr}	mass transfer coefficient	0.08	s^{-1}	a
H_{O_2}	Henry's coefficient of O_2 at 20 °C	31.2	$M\,M^{-1}$	b
H_{N_2}	Henry's coefficient of N_2 at 20 °C	59.5	$M\,M^{-1}$	b
V_g	volume of the gas phase	0.07	L	c
V_w	volume of the liquid phase	0.05	L	c
f_{dil}	volume fraction replaced per sampling	0.013	–	a
$r_{diff}^{O_2}$	diffusion rate of O_2 into the system	2.7×10^{-12}	$M\,s^{-1}$	a
$r_{diff}^{N_2}$	diffusion rate of N_2 into the system	3.5×10^{-12}	$M\,s^{-1}$	a
$c_{leak}^{O_2}$	O_2 concentration increase per sampling	4.2×10^{-8}	M	a
$c_{leak}^{N_2}$	N_2 concentration increase per sampling	3.5×10^{-7}	M	a

[a] Linda Bergaust (personal communication, 2018-11-05) [b] Sander (2015) [c] Qu et al. (2015)

B.6 SIMPLIFIED MODEL OF TRANSCRIPTIONAL REGULATION

B.6.1 Model equations

In addition to the model that represents the transcription factors FnrP, NarR and NNR explicitly we set up a simplified model variant where transcriptional regulation directly depends on nitrogen oxide and oxygen concentrations. As we will show, however, this formulation can also be interpreted mechanistically in terms of regulation by transcription factors if some simplifying assumptions are made. The necessary assumptions partly differ from those made in the model that explicitly simulates transcription factor concentrations and are discussed in appendix B.6.3.

We assume that the transcription rate scales with the fraction of operator sites where an activating, but no repressing transcription factor is bound. Transcription factors can only bind to the operator when they are activated by a signaling molecule: N-substrates in the case of activators and oxygen for inhibitors. In our formulation of transcription kinetics, we assume that concentrations of activated transcription factors are proportional to the concentrations of the respective signaling compounds. The transcription rate for gene i is

$$r^i_{\text{trnscr}} = \alpha f^i_{\text{act}} f^i_{\text{inh}} B, \tag{B.8}$$

in which α [transcripts cell^{-1} s^{-1}] is the maximum transcription rate, and B is the cell density [cells L^{-1}].

The dimensionless factors f_{act} and f_{inh} (ranging between 0 and 1) regulate the transcription rate as influenced by external factors. The subscript "act" denotes an activator compound. For the case of *nirG* transcription, NO acts as the single activator compound (Bergaust et al., 2012; Spiro, 2012). Analogously to our description of NNR activation, however, we used nitrite as a proxy activator compound because we describe denitrification as a two-step reaction and do not explicitly simulate NO.

Thus, the regulation factor for *nir* transcription is

$$f^{nir}_{\text{act}} = \frac{c_{NO_2^-}}{A^{NO_2^-}_{nir} + c_{NO_2^-}}, \tag{B.9}$$

in which $c_{NO_2^-}$ is the concentration of nitrite [M] and $A^{NO_2^-}_{nir}$ is the half-velocity constant of transcription, that is, the concentration of the activator compound nitrite at which the transcription reaches half of its maximum rate (in absence of repressors). In contrast, transcription of *narG* in *P. denitrificans* can be activated by either nitrate or nitrite (Wood et al., 2001):

$$f^{nar}_{\text{act}} = \frac{\frac{c_{NO_3^-}}{A^{NO_3^-}_{nar}} + \frac{c_{NO_2^-}}{A^{NO_2^-}_{nar}}}{1 + \frac{c_{NO_3^-}}{A^{NO_3^-}_{nar}} + \frac{c_{NO_2^-}}{A^{NO_2^-}_{nar}}}, \tag{B.10}$$

in which $c_{NO_3^-}$ is the concentration of nitrate [M] whereas $A_{nar}^{NO_3^-}$ and $A_{nar}^{NO_2^-}$ are the half-velocity constants of the activator compounds nitrate and nitrite, respectively.

Transcription of the *nirS* and *narG* genes is inhibited by the presence of oxygen, described by the oxygen transcriptional inhibition factor, f_{inh}^{i}:

$$f_{\text{inh}}^{i} = \frac{I_{\text{trs}}^{i}}{c_{O_2} + I_{\text{trs}}^{i}}, \tag{B.11}$$

in which c_{O_2} is the concentration of oxygen [M] and I_{trs}^{i} is the transcript inhibition constant of oxygen for gene *i*. In contrast to the model that represents transcription factors explicitly, we described transcript concentrations by the transient equation instead of a quasi-steady state.

B.6.2 *Simulation results*

Figure B.6 shows the simulation results obtained with the simplified model formulation. Table B.4 lists the prior distributions and statistics of posterior distributions for all estimated parameters used in the simplified model simulation. Fixed parameters have the same values as in table B.2 and are therefore omitted.

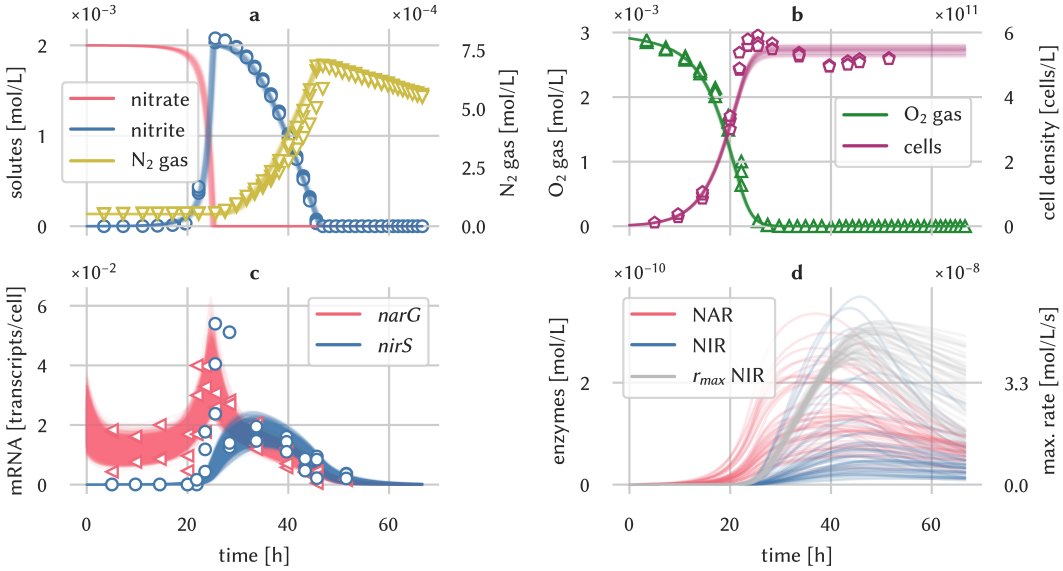


FIGURE B.6: Measurement data, 50 draws from the posterior of the enzyme-based model with a simplified description of transcriptional regulation as presented in appendix B.6.1: **a** Nitrogen compounds, **b** oxygen and cell densities, **c** transcripts, **d** enzymes and maximum rates.

TABLE B.4: Simulation parameters of the enzyme-based model with a simplified formulation for transcription, their prior distributions, and their posterior medians and percentiles. Fixed parameters have the same values as in table B.2 and are omitted.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles		
					10 %	Median	90 %
$t^T_{1/2}$	transcript half life	min	$N(1.8, 0.4)$	c, d	1.3×10^{2}	1.6×10^{2}	1.9×10^{2}
$A^{NO_3^-}_{nar}$	NO_3^- concentration inducing transcription of *nar*	M	$N(-12, 2)$	b	4.3×10^{-7}	5.9×10^{-6}	7.1×10^{-5}
$A^{NO_2^-}_{nar}$	NO_2^- concentration inducing transcription of *nar*	M	$N(-8, 2)$	b	4.8×10^{-3}	8.5×10^{-3}	1.6×10^{-2}
$A^{NO_2^-}_{nir}$	NO_2^- concentration inducing transcription of *nir*	M	$N(-11, 3)$	b	2.6×10^{-3}	7.8×10^{-3}	2.0×10^{-2}
I^{nar}_{trs}	O_2 inhibition parameter for *nar*-transcription	M	$T(-11, 2, 10)$	b	1.4×10^{-5}	4.2×10^{-5}	8.6×10^{-5}
I^{nir}_{trs}	O_2 inhibition parameter for *nir*-transcription	M	$T(-11, 2, 10)$	b	1.9×10^{-5}	4.8×10^{-5}	3.5×10^{-4}
$t^E_{1/2}$	enzyme half life	h	$N(3, 0.8)$	g, h	1.1×10^{1}	2.2×10^{1}	4.9×10^{1}
β^{nar}_T	maximum *narG* concentration	transcripts cell^{-1}	$T(-2, 0.7, 10)$	g, i	4.8×10^{-2}	6.7×10^{-2}	1.1×10^{-1}
β^{nir}_T	maximum *nirS* concentration	transcripts cell^{-1}	$T(-2, 0.7, 10)$	g, i	4.4×10^{-2}	1.0×10^{-1}	2.2×10^{-1}
I^{NAR}_{reac}	O_2 inhibition parameter for NAR	M	$N(-15, 3)$		1.3×10^{-7}	8.7×10^{-7}	3.6×10^{-6}
I^{NIR}_{reac}	O_2 inhibition parameter for NIR	M	$N(-15, 3)$		1.8×10^{-7}	1.5×10^{-6}	3.4×10^{-5}
K_{O_2}	O_2 half-saturation constant	M	$T(-13, 2, 10)$	j	2.9×10^{-5}	3.2×10^{-5}	3.5×10^{-5}

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters but σ_i . $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's *t*-distribution with location μ, scale σ and degrees of freedom ν. [b] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [c] Bernstein et al. (2002)
[d] Härtig and Zumft (1999) [e] Reparametrization of v^i_{max}, see appendix B.4 for details. [f] Based on experimental data. [g] Maier et al. (2011)
[h] Roberts et al. (2011) [i] Qu et al. (2015) [j] Suenaga et al. (2018) [k] Hassan et al. (2016)

TABLE B.4: (continued) Simulation parameters of the enzyme-based model with a simplified formulation for transcription, their prior distributions, and their posterior medians and percentiles. Fixed parameters have te same values as in table B.2 and are omitted.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles		
					10 %	Median	90 %
Y_{O_2}	O_2 growth yield per amount e^--donor	cells mol^{-1}	$T(32.2, 0.5, 10)$	k	4.8×10^{14}	4.9×10^{14}	5.0×10^{14}
$v_{fix}^{O_2}$	O_2 reaction rate at fixed O_2 concentration	mol cell^{-1} s^{-1}	$T(-44.2, 1.6, 10)$	e	4.1×10^{-19}	4.2×10^{-19}	4.3×10^{-19}
$k_{max}^{NO_3^-}$	NAR turnover number	s^{-1}	$T(7, 2, 10)$		7.8×10^{3}	3.8×10^{4}	2.7×10^{5}
$k_{max}^{NO_2^-}$	NIR turnover number	s^{-1}	$T(7, 2, 10)$		2.9×10^{2}	7.2×10^{2}	1.9×10^{3}
B_0	initial cell concentration	cells L^{-1}	$T(21, 0.5, 10)$	f	3.0×10^{9}	3.3×10^{9}	3.5×10^{9}
b_{O_2}	background value of O_2	M	$N(-18, 3)$		8.1×10^{-8}	9.5×10^{-8}	1.1×10^{-7}
$b_{NO_2^-}$	background value of NO_2^-	M	$N(-18, 3)$		3.4×10^{-10}	1.1×10^{-8}	1.3×10^{-7}
σ_B	constant error of Box-Cox transformed cell density data		Half-Normal(0.1)		2.1×10^{-1}	2.6×10^{-1}	3.1×10^{-1}
σ_{O_2}	constant error of Box-Cox transformed O_2 gas data		$H(0, 0.1)$		2.5×10^{-2}	2.9×10^{-2}	3.3×10^{-2}
$\sigma_{NO_2^-}$	constant error of Box-Cox transformed NO_2^- data		$H(0.025, 0.1)$		2.5×10^{-2}	2.5×10^{-2}	2.6×10^{-2}
σ_{N_2}	constant error of Box-Cox transformed N_2 gas data		$H(0.1, 0.1)$		1.0×10^{-1}	1.0×10^{-1}	1.0×10^{-1}
σ_{nar}	constant error of Box-Cox transformed *nar* mRNA data		$H(0, 0.1)$		1.1×10^{-1}	1.6×10^{-1}	2.2×10^{-1}

Continued on the next page.

[a] Prior distributions are defined for the log of all parameters but σ_i . $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's *t*-distribution with location μ, scale σ and degrees of freedom ν. [b] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [c] Bernstein et al. (2002)
[d] Härtig and Zumft (1999) [e] Reparametrization of v_{max}^{i}, see appendix B.4 for details. [f] Based on experimental data. [g] Maier et al. (2011)
[h] Roberts et al. (2011) [i] Qu et al. (2015) [j] Suenaga et al. (2018) [k] Hassan et al. (2016)

TABLE B.4: (continued) Simulation parameters of the enzyme-based model with a simplified formulation for transcription, their prior distributions, and their posterior medians and percentiles. Fixed parameters have te same values as in table B.2 and are omitted.

Symbol	Description	Units	Prior[a]/Value	Note	Posterior Percentiles		
					10 %	Median	90 %
σ_{nir}	constant error of Box-Cox transformed *nir* mRNA data		$H(0, 0.1)$		2.9×10^{-2}	4.1×10^{-2}	5.7×10^{-2}

[a] Prior distributions are defined for the log of all parameters but σ_i . $N(\mu, \sigma)$ is a normal distribution, $H(\mu, \sigma)$ is a half-normal distribution and $T(\mu, \sigma, \nu)$ a Student's t-distribution with location μ, scale σ and degrees of freedom ν. [b] Parameters for transcriptional regulation are not well known. We chose very broad distributions spanning several orders of magnitude. [c] Bernstein et al. (2002)
[d] Härtig and Zumft (1999) [e] Reparametrization of $v^i_{\max}$, see appendix B.4 for details. [f] Based on experimental data. [g] Maier et al. (2011)
[h] Roberts et al. (2011) [i] Qu et al. (2015) [j] Suenaga et al. (2018) [k] Hassan et al. (2016)

B.6.3 *Deriving the transcriptional regulation from mechanistic principles*

The formulation of transcriptional regulation presented in appendix B.6.1, even though simplified compared to the actual underlying mechanisms, can be interpreted in terms of mechanistic principles. The derivations follow the analysis of Ingalls (2013, chapter 7.1.2).

One activator and one repressor We consider a gene that is regulated by two transcription factors, an activator A and a repressor R. They can bind at two distinct operator sites of the regulated gene's promoter region. Transcription will only occur if the activator is bound but the repressor is not bound to the operator O. We can describe the binding of the transcription factors at the operator by the following reactions:

$$\mathrm{O} + \mathrm{A} \underset{d_1}{\overset{a_1}{\rightleftharpoons}} \mathrm{OA}\,, \qquad \{1\}$$

$$\mathrm{O} + \mathrm{R} \underset{d_2}{\overset{a_2}{\rightleftharpoons}} \mathrm{OR}\,, \qquad \{2\}$$

$$\mathrm{OR} + \mathrm{A} \underset{d_1}{\overset{a_1}{\rightleftharpoons}} \mathrm{OAR}\,, \qquad \{3\}$$

$$\mathrm{OA} + \mathrm{R} \underset{d_2}{\overset{a_2}{\rightleftharpoons}} \mathrm{OAR}\,, \qquad \{4\}$$

in which a_1 and a_2 [$\mathrm{M}^{-1}\,\mathrm{s}^{-1}$] are reaction constants for the forward reaction and d_1, d_2 [s^{-1}] are the backwards reaction constants. Note that the binding of the two transcriptions factors is assumed to be independent. That is, the binding kinetics of R to O and to OA are the same. Likewise A binds to O and OR with the same rate constants.

The transcription rate can be considered to be proportional to the fraction of operator sites where only the activator is bound, f_{OA}:

$$r_{\mathrm{trnscr}} = \alpha \cdot f_{\mathrm{OA}}\,, \qquad \text{(B.12)}$$

$$f_{\mathrm{OA}} = \frac{c_{\mathrm{OA}}}{c_{\mathrm{O}} + c_{\mathrm{OA}} + c_{\mathrm{OR}} + c_{\mathrm{OAR}}}\,. \qquad \text{(B.13)}$$

Assuming that the binding reactions are at steady state and setting

$$K_A = \frac{d_1}{a_1}, \qquad K_R = \frac{d_2}{a_2}\,, \qquad \text{(B.14)}$$

we can write:

$$f_{\mathrm{OA}} = \frac{c_{\mathrm{A}}}{K_A + c_{\mathrm{A}} + \frac{c_{\mathrm{R}} K_A}{K_R} + \frac{c_{\mathrm{A}} c_{\mathrm{R}}}{K_R}} = \frac{c_{\mathrm{A}}}{c_{\mathrm{A}} + K_A} \frac{K_R}{K_R + c_{\mathrm{R}}} \,. \tag{B.15}$$

The transcription factor concentrations react to changes in environmental conditions. In a simplified model, this interaction could be described as the reaction of an inactive form of the transcription factor with an activating molecule. In the case of the denitrification genes, we assume that a nitrogen substrate N reacts with the activator and oxygen with the repressor:

$$\mathrm{N} + \mathrm{A}^{*} \underset{}{\overset{K_1}{\rightleftharpoons}} \mathrm{A} \tag{5}$$

$$\mathrm{O_2} + \mathrm{R}^{*} \underset{}{\overset{K_2}{\rightleftharpoons}} \mathrm{R} \cdot \tag{6}$$

Here, the inactive form of the transcription factors are indicated by an asterisk. We now assume: (i) The activation and deactivation reactions follow first order kinetics with respect to N or $\mathrm{O_2}$ and A or R, respectively. (ii) Activation and deactivation of transcription factors are fast compared to gene transcription, i. e. A and R are in quasi-steady state. The concentrations of active transcription factors, c_{A} and c_{R} [M], then are proportional to the nitrogen substrate and oxygen concentrations:

$$c_{\mathrm{A}} = m_{\mathrm{A}} \cdot c_{\mathrm{N}} \,, \tag{B.16}$$

$$c_{\mathrm{R}} = m_{\mathrm{R}} \cdot c_{\mathrm{O_2}} \,, \tag{B.17}$$

with the proportionality factors m_{A} and m_{R} [mol mol^{-1}]. Substituting these expressions into equation (B.15) we obtain

$$f_{\mathrm{OA}} = \frac{c_{\mathrm{N}}}{c_{\mathrm{N}} + \frac{K_A}{m_{\mathrm{A}}}} \cdot \frac{\frac{K_R}{m_{\mathrm{R}}}}{\frac{K_R}{m_{\mathrm{R}}} + c_{\mathrm{O_2}}} = \frac{c_{\mathrm{N}}}{c_{\mathrm{N}} + A_{\mathrm{N}}} \cdot \frac{I_{\mathrm{trs}}^{\mathrm{O_2}}}{I_{\mathrm{trs}}^{\mathrm{O_2}} + c_{\mathrm{O_2}}} \,, \tag{B.18}$$

with the transcription half-saturation constant A_{N} for substrate N and the oxygen inhibition constant for transcription $I_{\mathrm{trs}}^{\mathrm{O_2}}$. Combining equations (B.12) and (B.18) leads to our formulation of the *nir* transcription rate given in equation (B.9).

Two activators and one repressor The transcription of some genes might be activated by several different compounds. We can describe this with two different mechanisms which eventually lead to the same rate law for the transcription rate: (i) There is a single activating transcription factor A binding to the operator site. The concentration of the activator is determined by the concentrations of both activating compounds.

(ii) There are two different activating transcription factors A_1 and A_2 that can both bind to the same operating site, i. e. either A_1 *or* A_2 can be bound. The concentration of each transcription factor is determined by the concentration of one of the activating compounds. In the latter case the transcription rate is proportional to the fraction of operator sites in the states OA_1 and OA_2. This fraction is given by

$$f_{OA_1,OA_2} = \frac{c_{OA_1} + c_{OA_2}}{c_O + c_{OA_1} + c_{OA_2} + c_{OA_1R} + c_{OA_2R} + c_{OR}} \,. \tag{B.19}$$

Assuming equilibrium conditions like in equation (B.14) with K_{A_1}, K_{A_2} and K_R being the equilibrium constants for the binding of A_1, A_2, and R we obtain

$$f_{OA_1,OA_2} = \frac{K_R}{K_R + c_R} \cdot \frac{\frac{c_{A_1}}{K_{A_1}} + \frac{c_{A_2}}{K_{A_2}}}{1 + \frac{c_{A_1}}{K_{A_1}} + \frac{c_{A_2}}{K_{A_2}}} \,. \tag{B.20}$$

Assuming that the transcription factor concentrations follow

$$c_{A_1} = m_{A_1} \cdot c_{N_1} \,, \tag{B.21}$$

$$c_{A_2} = m_{A_2} \cdot c_{N_2} \,, \tag{B.22}$$

$$c_R = m_R \cdot c_{O_2} \,, \tag{B.23}$$

where N_1 and N_2 denote the two inducing nitrogen substrates, and setting

$$A_{N_1} = \frac{K_{A_1}}{m_{A_1}} \,, \qquad A_{N_2} = \frac{K_{A_2}}{m_{A_2}} \,, \qquad I_{trs}^{O_2} = \frac{K_R}{m_R} \,, \tag{B.24}$$

we can write:

$$f_{OA_1,OA_2} = \frac{I_{trs}^{O_2}}{I_{trs}^{O_2} + c_{O_2}} \cdot \frac{\frac{c_{N_1}}{A_{N_1}} + \frac{c_{N_2}}{A_{N_2}}}{1 + \frac{c_{N_1}}{A_{N_1}} + \frac{c_{N_2}}{A_{N_2}}} \,. \tag{B.25}$$

This is the formulation that we used for the transcription of *nar*, which is triggered by nitrate or nitrite given in equation (B.10).

Discussion of the underlying assumptions We argue that a linear relation between signalling compounds and transcription factor concentrations is a reasonable simplifying assumption. It enables the description of the regulation mechanism with a single parameter per transcription factor. However, when interpreting results based on this assumption one should keep its implications in mind:

1 First-order kinetics for the activation reaction of the transcription factor as given in equations (B.21) to (B.23) require that the concentration of inactive transcription factors is not limiting for the binding reaction, that is, c_N and c_{O_2} must be small compared to the equilibrium constant of the activation reaction K. Otherwise, assuming a constant total amount of transcription factors (active and inactive) c_A^{tot}, the relationship would be of the form

$$c_A = c_A^{tot} \frac{c_N}{c_N + K} \tag{B.26}$$

instead of a linear relationship.

2 Assuming a quasi-steady state of the active transcription factors is only valid if the activation or deactivation are much faster than substrate dynamics and transcription itself. Crack et al. (2016) showed that the deactivation of the oxygen-sensitive transcription factor FnrP by exposure to oxygen takes several minutes in *P. denitrificans* (mean lifetime of 3.5 min), which is similar to literature mRNA half-lives. Our simulation results reproduce the continued presence of transcripts after all substrate is used up. However, it does not resolve the dynamics of transcription factors and transcripts separately due to the quasi-steady state assumption. As a consequence, the mRNA half-life acts as a lumped parameter that accounts for both effects (non-immediate inactivation of transcription factors and decay of transcripts) and is much longer than expected from the literature.

C Supporting information for chapter 4

C.1 ADDITIONAL FIGURES

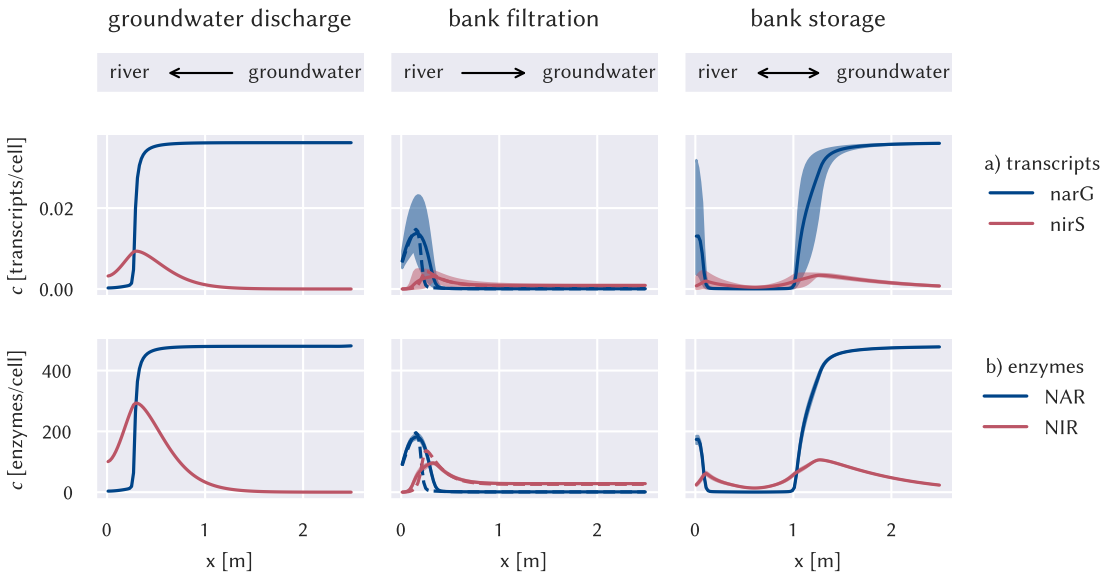


FIGURE C.1: Spatial distributions of transcript and enzyme concentrations normalized by biomass. The steady-state solution in scenario BSC is indicated by a dashed line. For the periodic solution in scenarios BFP and BS, the minimum and maximum value over time are indicated by the shaded area, the mean value is plotted as a solid line. Concentrations between 2.5 m and the groundwater-side domain boundary at 4 m are omitted because they are almost constant.

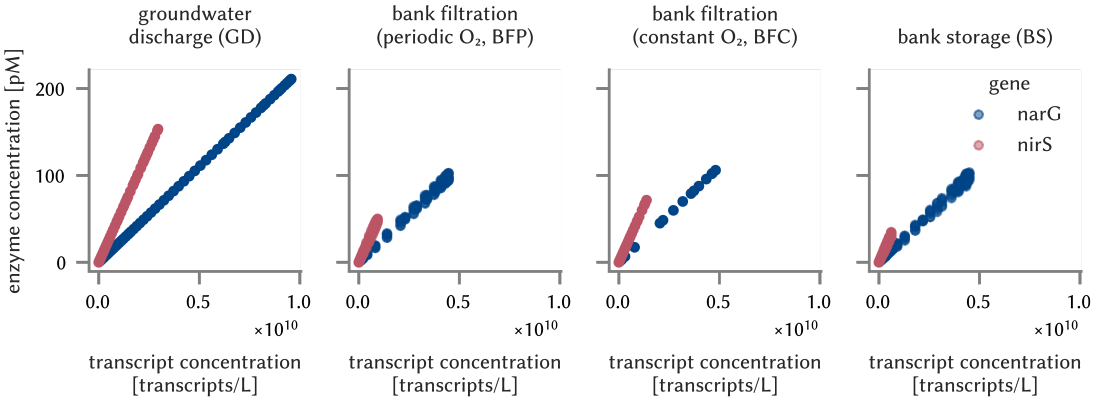


FIGURE C.2: Relationships between transcript concentrations and enzyme concentrations. In scenarios GD and BSC, concentrations are at steady state. In scenarios BFP and BS, transcript concentrations, but not enzyme concentrations, are averaged over time.

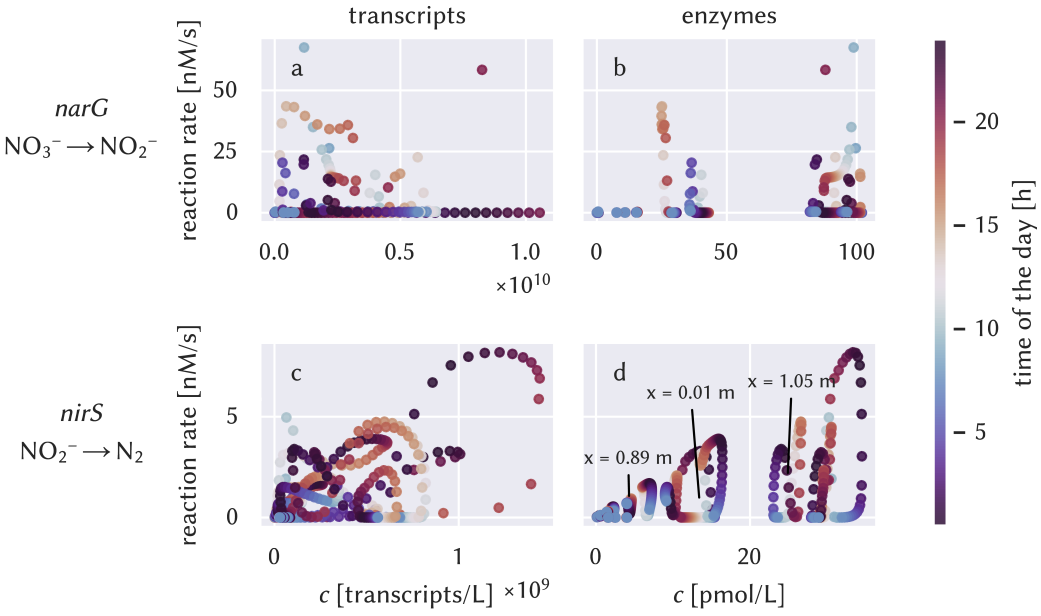


FIGURE C.3: Relationships between transcript (left column) respectively enzyme (right column) concentrations and denitrification rates for the bank storage (BS) scenario. Colors indicate the time point within the diurnal cycle. Every location shows a distinct pattern (with one "loop" corresponding to one location), and many of them are non-linear and hysteretic in time.

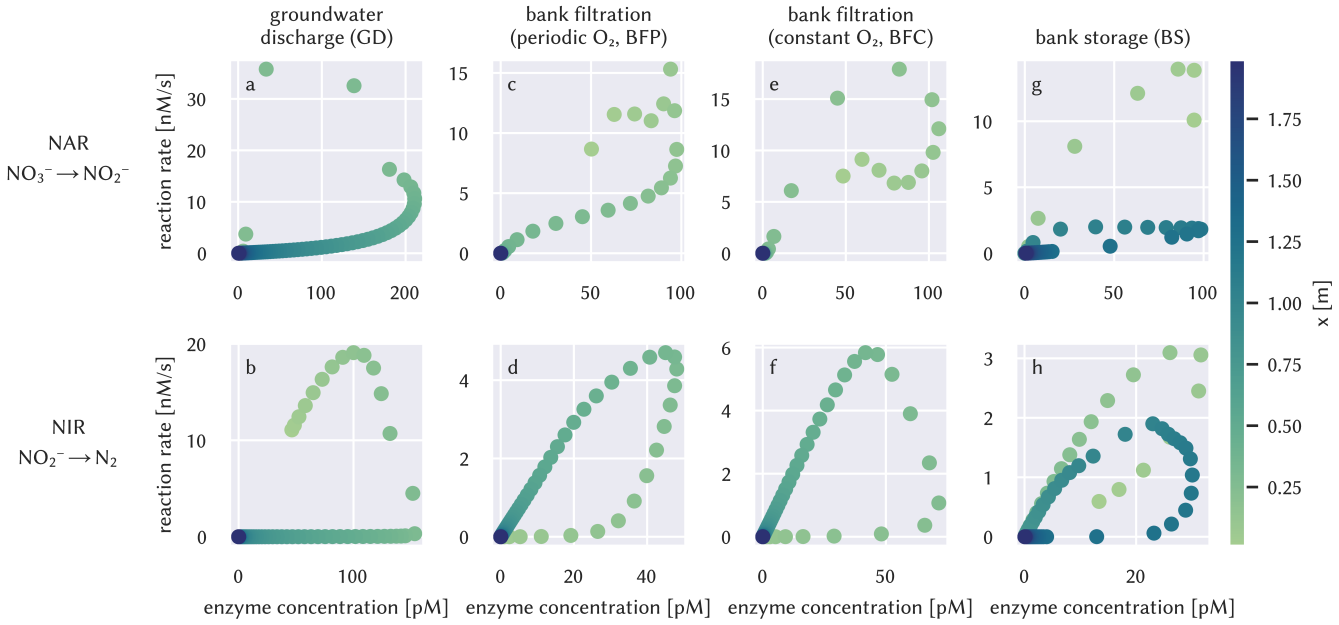


FIGURE C.4: Relationships between the concentrations of enzymes NAR (upper row) and NIR (lower row) with the denitrification rates in the different scenarios. In the scenarios where concentrations do not reach a steady state but stable diurnal cycles, daily averages of rates and concentrations are shown. The color indicates the spatial coordinate with dark blue corresponding to the groundwater inflow boundary and light green corresponding to the river boundary.

D Supporting information for chapter 5

D.1 Calculation of the reaction stoichiometry

The stoichiometry of the overall metabolic reaction depends on the stoichiometry of the anabolic and the catabolic reaction, and on the growth yield. Let $\alpha_{ji} \in \mathbb{R}$ be the stoichiometric coefficients of the catabolic reaction j (with i running from 1 to the number of reactants N) and $\beta_{ji} \in \mathbb{R}$ the coefficients for the corresponding anabolic reaction. Stoichiometric coefficients of educts are negative and coefficients for products are positive. The coefficients γ_{ji} for the overall metabolic reaction are then given by

$$\gamma_{ji} = \lambda_j \alpha_{ji} + \beta_{ji}\,, \tag{D.1}$$

where the factor $\lambda_j \in \mathbb{R}^+$ describes the number of times the catabolic reaction needs to run to complete the stoichiometric reaction of the anabolic reaction once. The factor λ_j depends on the molar growth yield Y_j^* [$\mathrm{mol_{bio}\,mol_{ED}^{-1}}$] that is defined as

$$Y_j^* = \frac{|\gamma_{j,\mathrm{bio}}|}{|\gamma_{j,\mathrm{ED}}|} = \frac{\beta_{j,\mathrm{bio}}}{\lambda_j(-\alpha_{j,\mathrm{ED}}) + (-\beta_{j,\mathrm{ED}})}\,, \tag{D.2}$$

where $\beta_{j,\mathrm{bio}}$ is the stoichiometric coefficient of biomass in the anabolic reaction, and $\alpha_{j,\mathrm{ED}}$, $\beta_{j,\mathrm{ED}}$, and $\gamma_{j,\mathrm{ED}}$ are the stoichiometric coefficients of the electron donor in the catabolic, anabolic and overall metabolic reactions, respectively. The second equality holds because we know that (1) $\alpha_{j,\mathrm{bio}} = 0$, and (2) in the overall metabolic reaction electron donor must be consumed, that is, $\gamma_{j,\mathrm{ED}} < 0$, implying that $|\lambda_j \alpha_{j,\mathrm{ED}} + \beta_{j,\mathrm{ED}}| = \lambda_j(-\alpha_{j,\mathrm{ED}}) + (-\beta_{j,\mathrm{ED}})$. It follows that

$$\lambda_j = \frac{\frac{\beta_{j,\mathrm{bio}}}{Y_j^*} - (-\beta_{j,\mathrm{ED}})}{(-\alpha_{j,\mathrm{ED}})}\,. \tag{D.3}$$

The molar growth yield Y_j^* relates to the more commonly used cellular growth yield Y_j by the molar weight of the biomass M_{bio} and the average cell weight w_{bio} by

$$Y_j^* = Y_j \frac{M_{\mathrm{bio}}}{w_{\mathrm{bio}}}\,. \tag{D.4}$$

When both the catabolic and the anabolic reaction consume the electron donor ($\alpha_{j,\mathrm{ED}}, \beta_{j,\mathrm{ED}} < 0$) as is the case for denitrification and DNRA, there is a theoretical maximum for the molar growth yield. The theoretical maximum is reached for $\lambda_j = 0$, that is, if the anabolic reaction could take place without requiring energy from the catabolic reaction. The maximum value is therefore given by

$$Y_j^{\max} = \frac{\beta_{j,\mathrm{bio}}}{\beta_{j,\mathrm{ED}}}\,. \tag{D.5}$$

In order to prevent nonphysical values for Y_j^* we defined the prior distribution in terms of a relative growth yield $Y_j^{\mathrm{rel}} \in [0, 1]$ and obtained the absolute growth yields by

$$Y_j^* = Y_j^{\mathrm{rel}}\, Y_j^{\max}\,. \tag{D.6}$$

D.2 DESCRIPTION OF THE CUSTOM SMC KERNEL

Sequential Monte Carlo (SMC) combines the ideas of importance sampling and tempering. That is, the algorithm starts sampling from the prior distribution and then moves towards the posterior distribution in several stages. The tempered posterior can be written as

$$p(\boldsymbol{\theta} \mid \boldsymbol{y})_\beta \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})^\beta\, p(\boldsymbol{\theta})\,. \tag{D.7}$$

When temperature parameter β is zero the tempered posterior distribution corresponds to the prior distribution, whereas a β of one corresponds to the true posterior distribution. Initially, β equals zero. It is then incrementally increased until it becomes unity.

In each stage an ensemble of independent MCMC chains is run for several steps. At the end of each stage, samples from the several chains are reweighted by importance weights to determine the starting points for the next stage.

By default, PyMC's SMC sampler uses an independent Metropolis Hastings kernel with a normal distribution as a kernel. In order to successfully sample the reactive transport model, the kernel was modified in two ways.

Thanks to Adrian Seyboldt for help with the modified MCMC kernel.

1 Instead of a normal distribution, we used a multivariate Gaussian mixture model with up to two components for the proposal distribution. The parameters of the Gaussian mixture model were fitted to the samples resulting from the reweighting step with the `GaussianMixture` class from the Scikit-learn library (Pedregosa et al., 2011). Using a Gaussian mixture model instead of a normal distribution allows for longer tails in the proposal distribution by superimposing normal distributions with different standard deviations.

2 We adapted the tempering schedule such that stages are repeated with the same β if too many MCMC chains failed to move from their starting position (that happens when all proposed points in a given chain are rejected).

D.3 REFERENCE PARAMETER VALUES

TABLE D.1: Nitrate half-saturation constants for denitrification. To convert between mass concentrations and molar concentrations, the molecular weight of nitrate $M_{NO_3^-} = 62.004\,\mathrm{g\,mol^{-1}}$ was used.

Reference	Note	Value	Units	Value [M]
Gu et al. (2007)		2	$\mathrm{mg\,L^{-1}}$	3.2×10^{-5}
Hassan et al. (2016)		5×10^{-6}	$\mathrm{mol\,L^{-1}}$	5×10^{-6}
Yan et al. (2016)	oxidized sediment	0.0018	mM	1.8×10^{-6}
Yan et al. (2016)	reduced sediment	0.1	mM	0.0001
Knights et al. (2017)		1.6	$\mathrm{mg\,L^{-1}}$	2.7×10^{-5}

TABLE D.2: Nitrite half-saturation constants for denitrification and DNRA.

Reference	Note	Value	Units	Value [M]
Hassan et al. (2016)	*Paracoccus denitrificans*	4.1×10^{-6}	$\mathrm{mol\,L^{-1}}$	4.1×10^{-6}
Yan et al. (2016)	oxidized and reduced sediment	0.0041	mM	4.1×10^{-6}
Zumft (1997)	*Pseudomonas aeruginosa*	53	µM	5.3×10^{-5}
Zumft (1997)	*Paracoccus denitrificans*	6	µM	6×10^{-6}

TABLE D.3: Doc half-saturation constants. To convert between mass concentrations and molar concentrations, the molecular weight of carbon $M_c = 12\,g\,mol^{-1}$ was used.

Reference	Note	Value	Units	Value [M]
Loschko et al. (2018)		20	$\mu mol\,L^{-1}$	2×10^{-5}
M. Li et al. (2017b)		0.000 25	$mol\,L^{-1}$	0.000 25
Song et al. (2017)		0.25	mM	0.000 25
Kinzelbach et al. (1991)		6	$mg\,L^{-1}$	0.0005
Sanz-Prat et al. (2016)		10	µM	10^{-5}
Knights et al. (2017)		8.7	$mg\,L^{-1}$	0.000 72
Gu et al. (2007)		1	$mg\,L^{-1}$	8.3×10^{-5}
Yan et al. (2016)		0.1	mM	0.0001
Yan et al. (2016)		0.001	mM	10^{-6}

TABLE D.4: Microbial decay constants.

Reference	Note	Value	Units	Value [d^{-1}]
Ding (2010)	reactive-tranpsort model for *Desulfovibrio vulgaris*	10^{-9}	s^{-1}	8.6×10^{-5}
Bælum et al. (2013)	*Dehalococcoides* spp.	0.1	d^{-1}	0.1
Song et al. (2017)		0.11	d^{-1}	0.11
M. Li et al. (2017b)		0.01	h^{-1}	0.24
Loschko et al. (2018)	modelling study	0.05	d^{-1}	0.05
Reed et al. (2014)	reactive-transport model in a marine system	0.001	d^{-1}	0.001
Kinzelbach et al. (1991)	reactive-transport model for a sandy aquifer	0.15	d^{-1}	0.15
Pagel et al. (2016)	carbon cycling model for soil	0.11	d^{-1}	0.11
Yan et al. (2016)	denitrification model for low permeable sediments	0.04	d^{-1}	0.04
Mellage et al. (2018)		0.0022	h^{-1}	0.053

TABLE D.5: NH_4^+ equilibrium sorption constants.

Reference	Note	Value	Units	Value [L g^{-1}]
Triska et al. (1994)	lower limit, river channel and riparian sediments	10	$g_w\,g_s^{-1}$	0.01
Triska et al. (1994)	upper limit, river channel and riparian sediments	2.8×10^3	$g_w\,g_s^{-1}$	2.8
Böhlke et al. (2006)	extraction experiments, aquifer sediments	0.46	$g_w\,g_s^{-1}$	0.000 46
Ceazan et al. (1989)	extraction, Cape Cod, near-bed site	0.87	$mL\,g_s^{-1}$	0.000 87
Ceazan et al. (1989)	extraction, Cape Cod, down-gradient site	0.59	$mL\,g_s^{-1}$	0.000 59
Ceazan et al. (1989)	linear isotherm, Cape Cod	0.34	$mL\,g_s^{-1}$	0.000 34

TABLE D.6: Doc release rate constants.

Reference	Note	Value	Units	Value [d^{-1}]
Sawyer (2015)	groundwater simulation study	48	d^{-1}	48
Gu et al. (2007)	streambed sediments	0.41	d^{-1}	0.41
Vavilin et al. (2008)	hydrolysis in forest soil	0.2	d^{-1}	0.2
Catalán et al. (2016)	mean value in freshwaters	0.28	y^{-1}	0.000 77
Catalán et al. (2016)	largest value in freshwaters	54	y^{-1}	0.15
Catalán et al. (2016)	smallest value in freshwaters	0.01	y^{-1}	2.7×10^{-5}
Sanz-Prat et al. (2016)	modelling study	2	d^{-1}	2

TABLE D.7: Maximum cell-specific substrate consumption rates v_{max}.

Reference	Reaction	Note	v_{max} [fmol cell^{-1} d^{-1}]
Hassan et al. (2016)	$NO_3^- \longrightarrow NO_2^-$	for *P. denitrificans* with butyrate as electron donor	120
Mellage et al. (2018)[a]	$NO_3^- \longrightarrow NO_2^-$	for *S. oneidensis*	21
Zarnetske et al. (2012)	$NO_3^- \longrightarrow N_2$	modeling study in the hyporheic zone	76
Hassan et al. (2016)	$NO_2^- \longrightarrow NO$	for *P. denitrificans* with butyrate as electron donor	64
Mellage et al. (2018)	$NO_2^- \longrightarrow NH_4^+$	for *S. oneidensis*	6
Hassan et al. (2016)	$N_2O \longrightarrow N_2$	for *P. denitrificans* with butyrate as electron donor	66

[a] Values were converted from maximum growth rate constants μ_{max} and growth yields.

D.4 ADDITIONAL FIGURES

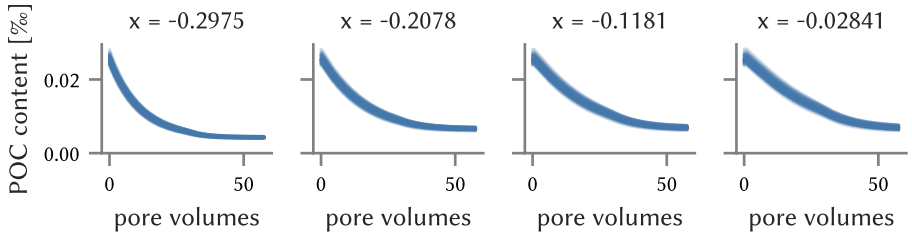


FIGURE D.1: Simulated content of bioavailable POC in the sediment, plotted over dimensionless time. Individual lines represent 40 random samples from the posterior distribution.

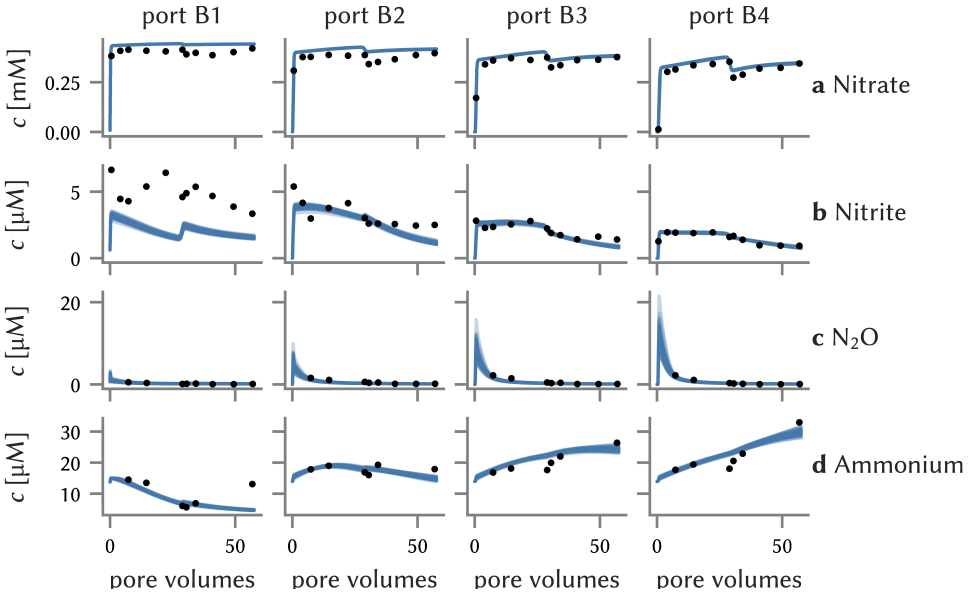


FIGURE D.2: Simulated (lines) and measured (markers) aqueous concentrations at the ports as a function of dimensionless time. Individual lines represent 40 draws from the posterior distribution of the model that uses functional gene data.

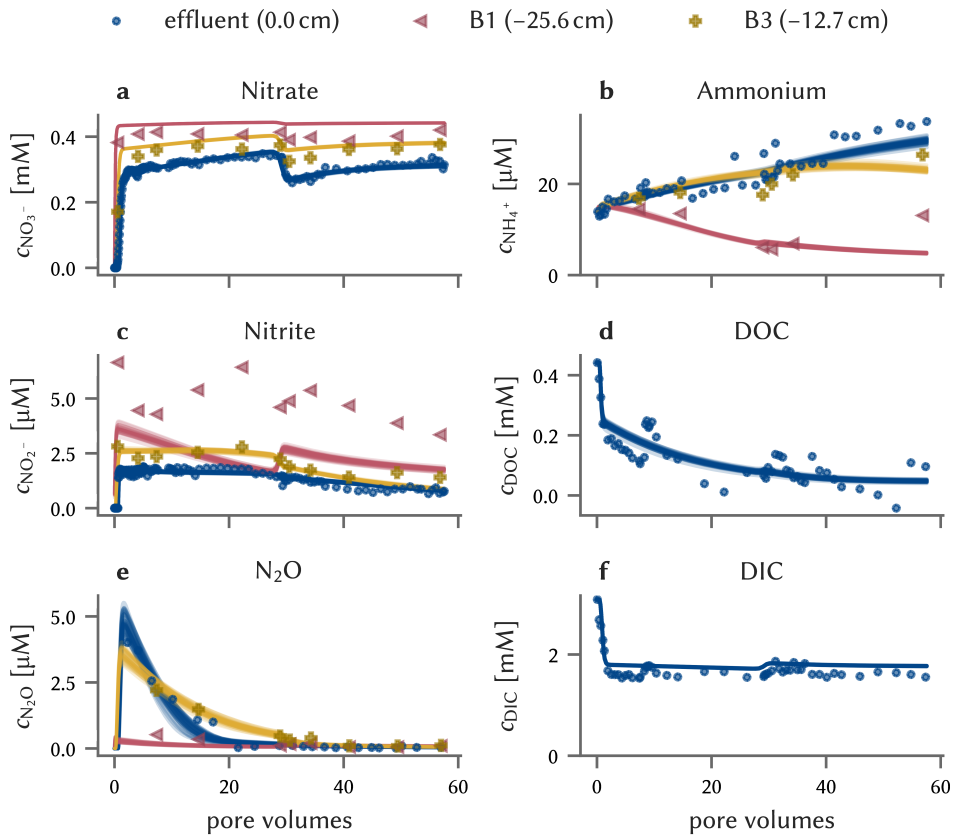


FIGURE D.3: Simulated (lines) and measured (markers) aqueous concentrations at the column effluent and at two ports close to the inflow (B1) and in the middle of the column (B3) as a function of dimensionless time. Individual lines represent 50 draws from the posterior distribution of the model that does not use functional gene data.

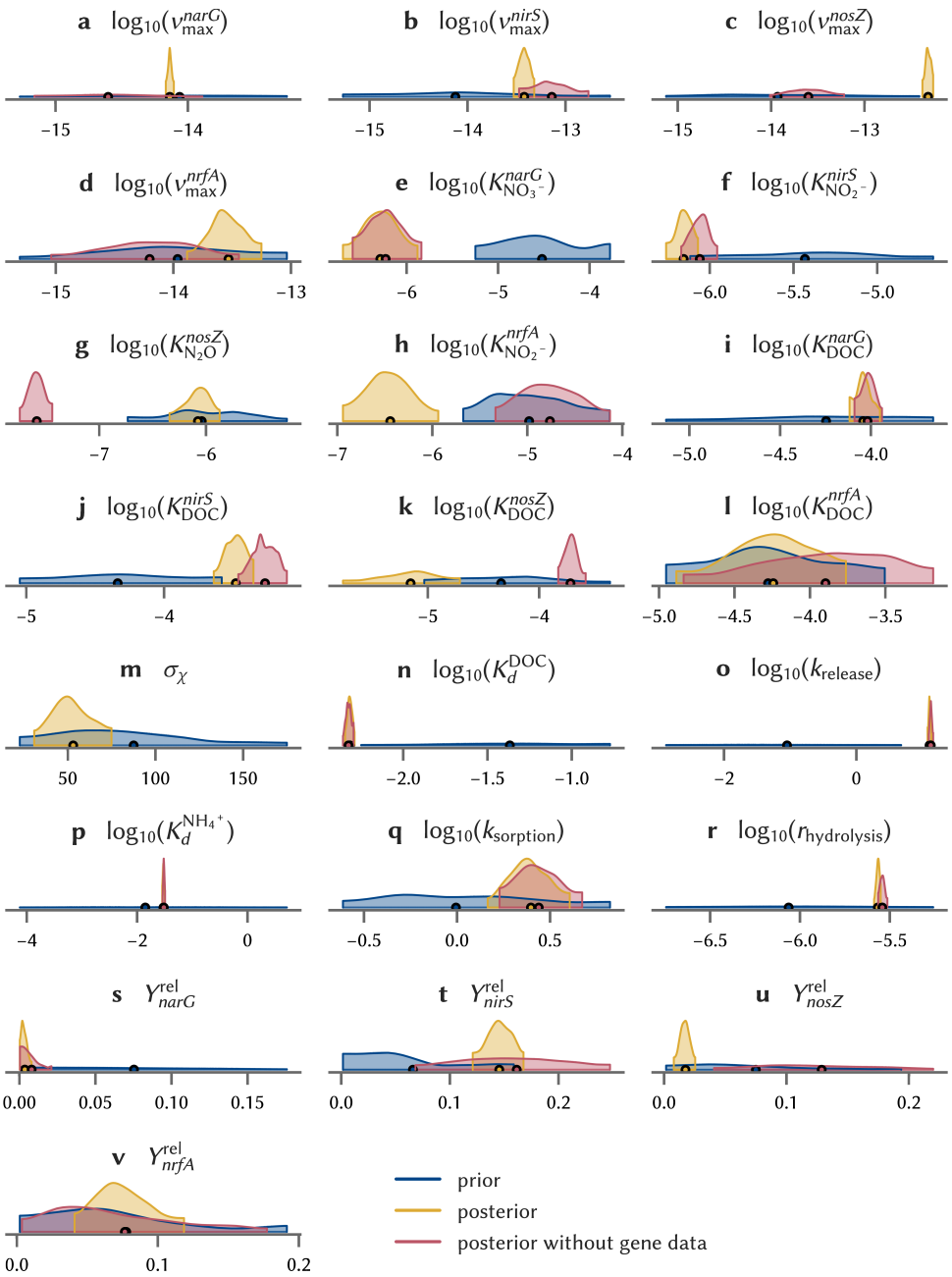


FIGURE D.4: Kernel density estimates of the marginal posterior and prior distributions for all parameters. Plots are cut at the 94 % highest density interval.

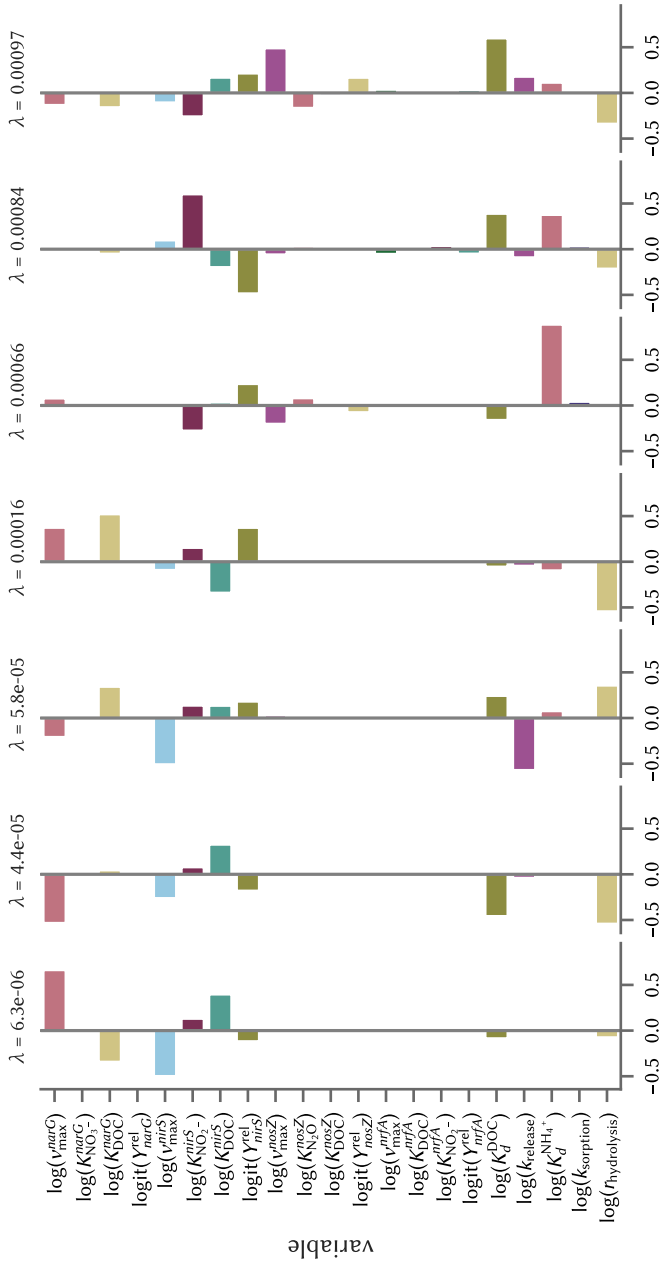


FIGURE D.5: Eigenvectors of the posterior covariance matrix corresponding to the smallest 7 eigenvalues.

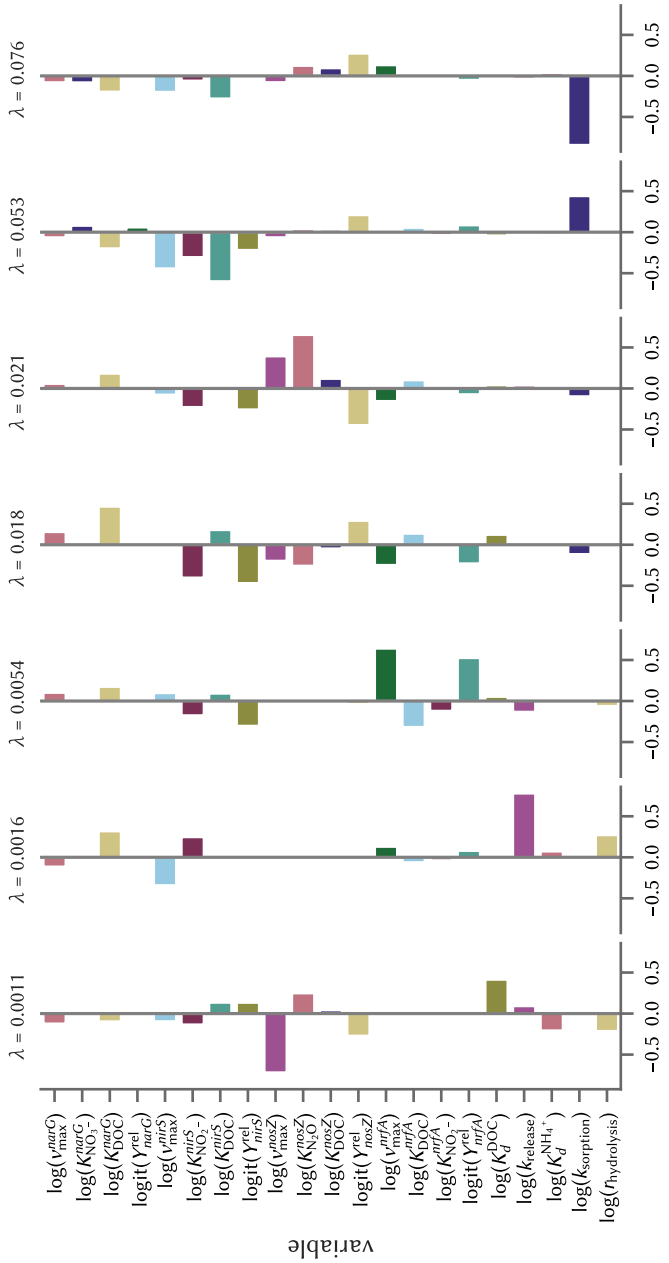


FIGURE D.6: Eigenvectors of the posterior covariance matrix corresponding to the 7th to 14th eigenvalues.

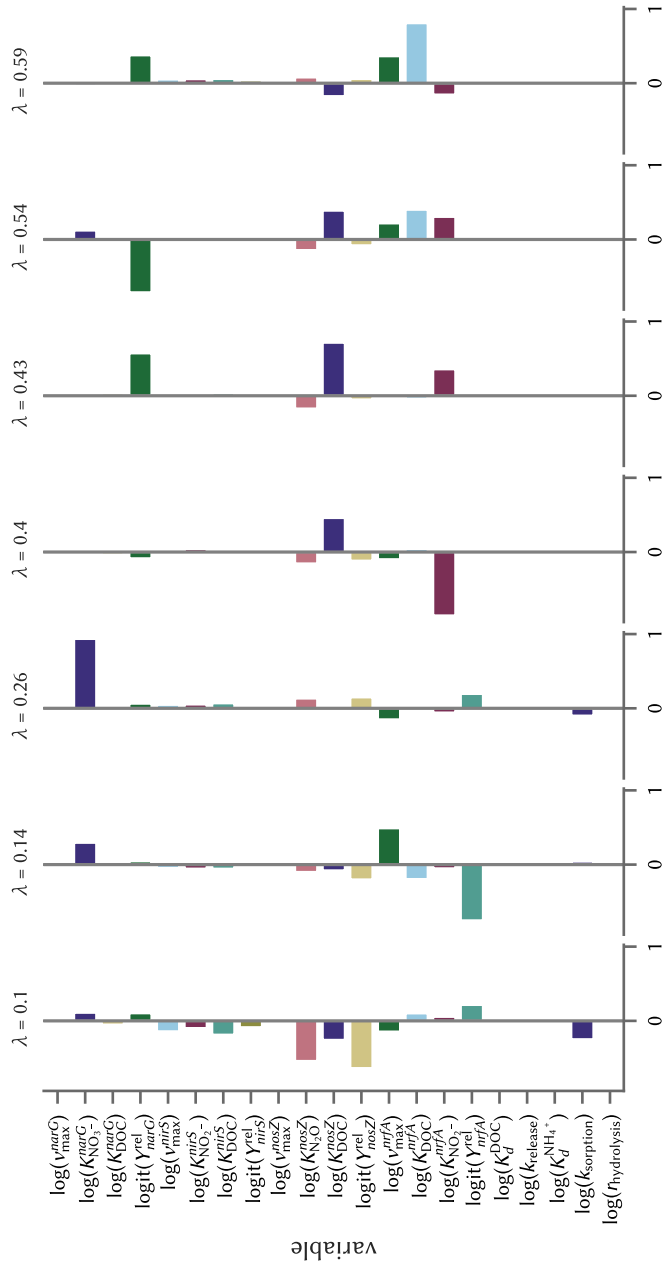


FIGURE D.7: Eigenvectors of the posterior covariance matrix corresponding to the largest 7 eigenvalues.

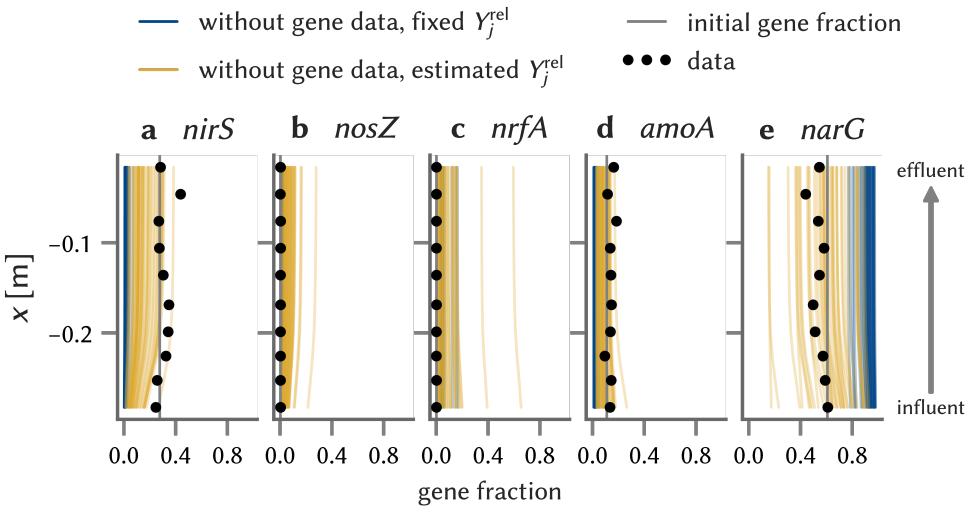


FIGURE D.8: Comparison of a model variant with fixed growth yield parameters Y_j^{rel} to the model where growth yields are estimated from the data. The plot shows the simulated and measured profiles of relative abundances of functional genes. In each case, 40 samples from the posterior distribution are drawn. For the simulation with fixed growth yields, the following parameters were used: $Y_{narG}^{rel} = Y_{nirS}^{rel} = Y_{nosZ}^{rel} = Y_{nrfA}^{rel} = 0.08$.

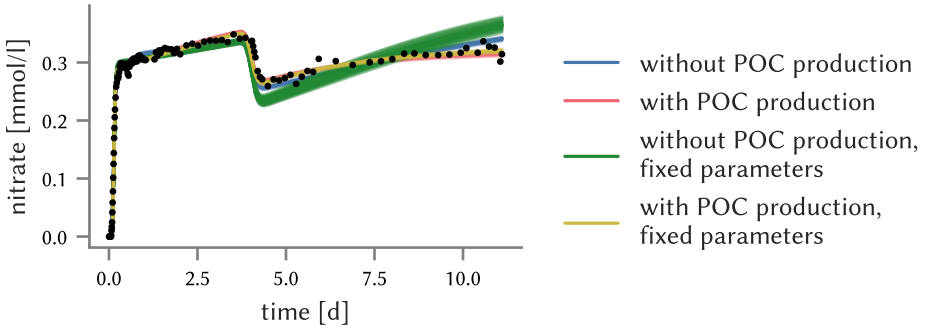


FIGURE D.9: Comparison of the model variants with and without a production term for bioavailable POC. The plot shows simulated nitrate concentrations from four model runs. We initially conducted the comparison of the model variants with a model where several reaction parameters (all growth yields and $K^{narG}_{NO_3^-}$) were set to fixed values (green and yellow lines). Model variants without a POC production term overestimate the increase of the nitrate concentration in the last three days of the experiment. In contrast, model variants with a POC production term capture the leveling-off well. The difference between the model variants with and without POC production was larger when a subset of the reaction parameters was fixed, indicating that other model parameters could partially compensate for the missing POC production term. Note that the sampler did not fully converge with the model variants without the POC production term.

List of Abbreviations

ANN · artificial neural network 103

ATP · adenosine triphosphate 2

BDF · backwards differentiation formula 14, 47, 76

BME · Bayesian model evidence 15

BONCAT-FACS · bioorthogonal non-canonical amino acid tagging combined with fluorescence-activated cell sorting 59, 99

DIC · dissolved inorganic carbon 66, 68, 71, 75, 80

DNRA · dissimilatory nitrate reduction to ammonium xiv, 1, 18, 63, 64, 67, 68, 81–83, 88, 93, 94, 132, 133

DOC · dissolved organic carbon xi, xiv, 40–50, 52, 53, 55–58, 61, 66, 67, 69, 71, 73, 79–81, 83, 84, 86, 134, 135

FTICR-MS · Fourier transform ion cyclotron resonance mass spectrometry 93

FVM · finite volume method 14, 15

GP · Gaussian process 101, 103

HMC · Hamiltonian Monte Carlo 16, 17, 26, 91, 98

MBAM · manifold boundary approximation method 87

MCMC · Markov chain Monte Carlo 16, 89, 91, 98, 132

NAR · nitrate reductase xii, 24, 33, 42, 43, 46, 49, 106, 110, 121, 122, 130

NIR · nitrite reductase xii, 24, 29, 31, 42, 43, 46, 49, 106, 110, 121, 122, 130

NUTS · No-U-Turn sampler 16, 17, 26

ODE · ordinary differential equation 13, 14, 26, 47, 76, 77, 104, 106

OD · optical density 19

PCE · polynomial chaos expansion 103

PCR · polymerase chain reaction 59, 65, 91

PDE · partial differential equation 5, 14

POC · particulate organic carbon xii, xiii, 40, 43, 44, 46, 47, 50, 52, 69, 71, 73, 77, 79, 80, 93, 136, 144

RNA · ribonucleic acid 11

RT-qPCR · reverse-transcription quantitative PCR 5, 19

SMC · sequential Monte Carlo x, 16, 76, 89, 132

anammox · anaerobic ammonium oxidation 1, 63, 67

mRNA · messenger RNA 4, 5, 12, 13, 19, 22, 23, 39, 50, 111, 112, 122, 123, 127

P. denitrificans · Paracoccus denitrificans 18–21, 24, 28, 35, 42, 95, 136

qPCR · quantitative PCR 3, 66, 71, 74, 99

rRNA · ribosomal RNA 71, 74, 99, 101

S. oneidensis · Shewanella oneidensis 136

Bibliography

Achermann, S., C. B. Mansfeldt, M. Müller, D. R. Johnson, and K. Fenner (2020). "Relating metatranscriptomic profiles to the micropollutant biotransformation potential of complex microbial communities." In: *Environmental Science & Technology* 54.1, pp. 235–244. DOI: 10.1021/acs.est.9b05421.

Akbarzadeh, Z., A. M. Laverman, F. Rezanezhad, M. Raimonet, E. Viollier, B. Shafei, and P. Van Cappellen (2018). "Benthic nitrite exchanges in the Seine River (France): an early diagenetic modeling analysis." In: *Science of The Total Environment* 628–629, pp. 580–593. DOI: 10.1016/j.scitotenv.2018.01.319.

Anantharaman, K., C. T. Brown, L. A. Hug, I. Sharon, C. J. Castelle, A. J. Probst, B. C. Thomas, A. Singh, M. J. Wilkins, U. Karaoz, E. L. Brodie, K. H. Williams, S. S. Hubbard, and J. F. Banfield (2016). "Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system." In: *Nature Communications* 7, p. 13219. DOI: 10.1038/ncomms13219.

Arhonditsis, G., D.-K. Kim, N. Kelly, A. Neumann, and A. Javed (2018). "Uncertainty analysis by Bayesian inference." In: *Ecological Informatics: Data Management and Knowledge Discovery*. Ed. by F. Recknagel and W. K. Michener. Cham: Springer International Publishing, pp. 215–249. DOI: 10.1007/978-3-319-59928-1_11.

Arora, B., D. Dwivedi, S. S. Hubbard, C. I. Steefel, and K. H. Williams (2016). "Identifying geochemical hot moments and their controls on a contaminated river floodplain system using wavelet and entropy approaches." In: *Environmental Modelling & Software* 85, pp. 27–41. DOI: 10.1016/j.envsoft.2016.08.005.

Arora, B., B. P. Mohanty, J. T. McGuire, and I. M. Cozzarelli (2013). "Temporal dynamics of biogeochemical processes at the Norman Landfill site." In: *Water Resources Research* 49.10, pp. 6909–6926. DOI: 10.1002/wrcr.20484.

Asher, M. J., B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters (2015). "A review of surrogate models and their application to groundwater modeling." In: *Water Resources Research* 51.8, pp. 5957–5973. DOI: 10.1002/2015WR016967.

Bælum, J., J. C. Chambon, C. Scheutz, P. J. Binning, T. Laier, P. L. Bjerg, and C. S. Jacobsen (2013). "A conceptual model linking functional gene expression and reductive dechlorination rates of chlorinated ethenes in clay rich groundwater sediment." In: *Water Research* 47.7, pp. 2467–2478. DOI: 10.1016/j.watres.2013.02.016. PMID: 23490098.

Bælum, J., M. H. Nicolaisen, W. E. Holben, B. W. Strobel, J. Sørensen, and C. S. Jacobsen (2008). "Direct analysis of *tfdA* gene expression by indigenous bacteria in phenoxy acid amended agricultural soil." In: *The ISME Journal* 2.6, pp. 677–687. DOI: 10.1038/ismej.2008.21.

Bahureksa, W., M. M. Tfaily, R. M. Boiteau, R. B. Young, M. N. Logan, A. M. McKenna, and T. Borch (2021). "Soil organic matter characterization by Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS): A critical review of sample preparation, analysis,

and data interpretation." In: *Environmental Science & Technology* 55.14, pp. 9637–9656. DOI: 10.1021/acs.est.1c01135.

Bergaust, L., R. J. M. van Spanning, Å. Frostegård, and L. R. Bakken (2012). "Expression of nitrous oxide reductase in *Paracoccus denitrificans* is regulated by oxygen and nitric oxide through FnrP and NNR." In: *Microbiology (Reading, England)* 158 (Pt 3), pp. 826–834. DOI: 10.1099/mic.0.054148-0. PMID: 22174385.

Bernstein, J. A., A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen (2002). "Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.15, pp. 9697–9702. DOI: 10.1073/pnas.112318199. PMID: 12119387.

Bertin, C. and A. C. M. Bourg (1994). "Radon-222 and chloride as natural tracers of the infiltration of river water into an alluvial aquifer in which there is significant river/groundwater mixing." In: *Environmental Science & Technology* 28.5, pp. 794–798. DOI: 10.1021/es00054a008.

Betancourt, M. (2018). *A Conceptual Introduction to Hamiltonian Monte Carlo*. DOI: 10.48550/arXiv.1701.02434. arXiv: 1701.02434 [stat.ME].

Blaszczyk, M. (1993). "Effect of medium composition on the denitrification of nitrate by *Paracoccus denitrificans*." In: *Applied and Environmental Microbiology* 59.11, pp. 3951–3953. PMID: 16349097.

Böhlke, J. K., R. L. Smith, and D. N. Miller (2006). "Ammonium transport and reaction in contaminated groundwater: Application of isotope tracers and isotope fractionation studies." In: *Water Resources Research* 42.5. DOI: 10.1029/2005WR004349.

Bol, R., A. Lücke, W. Tappe, S. Kummer, M. Krause, S. Weigand, T. Pütz, and H. Vereecken (2015). "Spatio-temporal variations of dissolved organic matter in a German forested mountainous headwater catchment." In: *Vadose Zone Journal* 14.4, vzj2015.01.0005. DOI: 10.2136/vzj2015.01.0005.

Boogerd, F. C., H. W. van Verseveld, D. Torenvliet, M. Braster, and A. H. Stouthamer (1984). "Reconsideration of the efficiency of energy transduction in *Paracoccus denitrificans* during growth under a variety of culture conditions." In: *Archives of Microbiology* 139.4, pp. 344–350. DOI: 10.1007/BF00408377.

Bouchez, T. et al. (2016). "Molecular microbiology methods for environmental diagnosis." In: *Environmental Chemistry Letters* 14.4 (4), pp. 423–441. DOI: 10.1007/s10311-016-0581-3.

Bouskill, N., J. Tang, W. J. Riley, and E. L. Brodie (2012). "Trait-based representation of biological nitrification: model development, testing, and predicted community composition." In: *Frontiers in Microbiology* 3. DOI: 10.3389/fmicb.2012.00364.

Bowen, J. L., A. R. Babbin, P. J. Kearns, and B. B. Ward (2014). "Connecting the dots: linking nitrogen cycle gene expression to nitrogen fluxes in marine sediment mesocosms." In: *Frontiers in Microbiology* 5. DOI: 10.3389/fmicb.2014.00429.

Box, G. E. P. and D. R. Cox (1964). "An analysis of transformations." In: *Journal of the Royal Statistical Society, Series B (Methodological)* 26.2, pp. 211–252. JSTOR: 2984418.

Brow, C. N., R. O'Brien Johnson, R. L. Johnson, and H. M. Simon (2013). "Assessment of anaerobic toluene biodegradation activity by *bssA* transcript/gene ratios." In: *Applied and Environmental Microbiology* 79.17, pp. 5338–5344. DOI: 10.1128/AEM.01031-13.

Brunetti, G., J. Šimůnek, D. Glöckler, and C. Stumpp (2020). "Handling model complexity with parsimony: numerical analysis of the nitrogen turnover in a controlled aquifer model setup." In: *Journal of Hydrology* 584, p. 124681. DOI: 10.1016/j.jhydrol.2020.124681.

Burri, N. M., R. Weatherl, C. Moeck, and M. Schirmer (2019). "A review of threats to groundwater quality in the anthropocene." In: *Science of The Total Environment* 684, pp. 136–154. DOI: 10.1016/j.scitotenv.2019.05.236.

Byrne, G. D. and A. C. Hindmarsh (1975). "A polyalgorithm for the numerical solution of ordinary differential equations." In: *ACM Transactions on Mathematical Software* 1.1, pp. 71–96. DOI: 10.1145/355626.355636.

Carini, P., P. J. Marsden, J. W. Leff, E. E. Morgan, M. S. Strickland, and N. Fierer (2016). "Relic DNA is abundant in soil and obscures estimates of soil microbial diversity." In: *Nature Microbiology* 2.3 (3), pp. 1–6. DOI: 10.1038/nmicrobiol.2016.242.

Catalán, N., R. Marcé, D. N. Kothawala, and L. J. Tranvik (2016). "Organic carbon decomposition rates controlled by water retention time across inland waters." In: *Nature Geoscience* 9.7 (7), pp. 501–504. DOI: 10.1038/ngeo2720.

Ceazan, M. L., E. M. Thurman, and R. L. Smith (1989). "Retardation of ammonium and potassium transport through a contaminated sand and gravel aquifer: the role of cation exchange." In: *Environmental Science & Technology* 23.11, pp. 1402–1408. DOI: 10.1021/es00069a012.

Chakrawal, A., A. M. Herrmann, J. Koestel, J. Jarsjö, N. Nunan, T. Kätterer, and S. Manzoni (2020). "Dynamic upscaling of decomposition kinetics for carbon cycling models." In: *Geoscientific Model Development* 13.3, pp. 1399–1429. DOI: 10.5194/gmd-13-1399-2020.

Chavez Rodriguez, L., B. Ingalls, E. Schwarz, T. Streck, M. Uksa, and H. Pagel (2020). "Gene-centric model approaches for accurate prediction of pesticide biodegradation in soils." In: *Environmental Science & Technology* 54.21, pp. 13638–13650. DOI: 10.1021/acs.est.0c03315.

Cheng, Y., C. G. Hubbard, L. Zheng, B. Arora, L. Li, U. Karaoz, J. Ajo-Franklin, and N. J. Bouskill (2018). "Next generation modeling of microbial souring – parameterization through genomic information." In: *International Biodeterioration & Biodegradation* 126, pp. 189–203. DOI: 10.1016/j.ibiod.2017.06.014.

Couradeau, E., J. Sasse, D. Goudeau, N. Nath, T. C. Hazen, B. P. Bowen, R. Chakraborty, R. R. Malmstrom, and T. R. Northen (2019). "Probing the active fraction of soil microbiomes using BONCAT-FACS." In: *Nature Communications* 10.1 (1), p. 2770. DOI: 10.1038/s41467-019-10542-0.

Crack, J. C., M. I. Hutchings, A. J. Thomson, and N. E. Le Brun (2016). "Biochemical properties of *Paracoccus denitrificans* FnrP: reactions with molecular oxygen and nitric oxide." In: *JBIC Journal of Biological Inorganic Chemistry* 21.1, pp. 71–82. DOI: 10.1007/s00775-015-1326-7.

Crossette, E., J. Gumm, K. Langenfeld, L. Raskin, M. Duhaime, and K. Wigginton (2021). "Metagenomic quantification of genes with internal standards." In: *mBio* 12.1, e03173–20. DOI: 10.1128/mBio.03173-20.

Danczak, R. E., A. H. Sawyer, K. H. Williams, J. C. Stegen, C. Hobson, and M. J. Wilkins (2016). "Seasonal hyporheic dynamics control coupled microbiology and geochemistry in Colorado River sediments." In: *Journal of Geophysical Research: Biogeosciences* 121.12, pp. 2976–2987. DOI: 10.1002/2016JG003527.

Delogu, F., B. J. Kunath, P. N. Evans, M. Ø. Arntzen, T. R. Hvidsten, and P. B. Pope (2020). "Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes." In: *Nature Communications* 11.1 (1), p. 4708. DOI: 10.1038/s41467-020-18543-0.

Dick, G. J. (2016). "Embracing the mantra of modellers and synthesizing omics, experiments and models." In: *Environmental Microbiology Reports* 9.1, pp. 18–20. DOI: 10.1111/1758-2229.12491.

Dietzel, A. and P. Reichert (2014). "Bayesian inference of a lake water quality model by emulating its posterior density." In: *Water Resources Research* 50.10, pp. 7626–7647. DOI: 10.1002/2012WR013086.

Ding, D. (2010). "Transport of bacteria in aquifer sediment: experiments and modeling." In: *Hydrogeology Journal* 18.3, pp. 669–679. DOI: 10.1007/s10040-009-0559-3.

Dyke, P. P. G. (2001). "Convolution and the solution of ordinary differential equations." In: *An Introduction to Laplace Transforms and Fourier Series*. Springer Undergraduate Mathematics Series. London: Springer, pp. 37–77. DOI: 10.1007/978-1-4471-0505-3_3.

Erisman, J. W., J. N. Galloway, S. Seitzinger, A. Bleeker, N. B. Dise, A. M. R. Petrescu, A. M. Leach, and W. de Vries (2013). "Consequences of human modification of the global nitrogen cycle." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1621, p. 20130116. DOI: 10.1098/rstb.2013.0116.

Fischer, H., F. Kloep, S. Wilzcek, and M. T. Pusch (2005). "A river's liver – microbial processes within the hyporheic zone of a large lowland river." In: *Biogeochemistry* 76.2, pp. 349–371. DOI: 10.1007/s10533-005-6896-y.

Fittipaldi, M., A. Nocker, and F. Codony (2012). "Progress in understanding preferential detection of live cells using viability dyes in combination with DNA amplification." In: *Journal of Microbiological Methods* 91.2, pp. 276–289. DOI: 10.1016/j.mimet.2012.08.007.

Foreman-Mackey, D. (2020). *Exoplanet-Dev/Pymc3-Ext*. Version 0.0.2.

Freitag, T. E. and J. I. Prosser (2009). "Correlation of methane production and functional gene transcriptional activity in a peat soil." In: *Applied and Environmental Microbiology* 75.21, pp. 6679–6687. DOI: 10.1128/AEM.01021-09.

Frioux, C., D. Singh, T. Korcsmaros, and F. Hildebrand (2020). "From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes." In: *Computational and Structural Biotechnology Journal* 18, pp. 1722–1734. DOI: 10.1016/j.csbj.2020.06.028.

Gaimster, H., M. Alston, D. J. Richardson, A. J. Gates, and G. Rowley (2018). "Transcriptional and environmental control of bacterial denitrification and N_2O emissions." In: *FEMS Microbiology Letters* 365.5. DOI: 10.1093/femsle/fnx277.

García-Ruiz, R., S. N. Pattinson, and B. A. Whitton (1998). "Kinetic parameters of denitrification in a river continuum." In: *Applied and Environmental Microbiology* 64.7, pp. 2533–2538. DOI: 10.1128/AEM.64.7.2533-2538.1998. PMID: 9647826.

Gelhar, L. W., C. Welty, and K. R. Rehfeldt (1992). "A critical review of data on field-scale dispersion in aquifers." In: *Water Resources Research* 28.7, pp. 1955–1974. DOI: 10.1029/92WR00607.

Gerecht, K. E., M. B. Cardenas, A. J. Guswa, A. H. Sawyer, J. D. Nowinski, and T. E. Swanson (2011). "Dynamics of hyporheic flow and heat transport across a bed-to-bank continuum in a large regulated river." In: *Water Resources Research* 47.3. DOI: 10.1029/2010WR009794.

Gesztelyi, R., J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz, and A. Tosaki (2012). "The Hill equation and the origin of quantitative pharmacology." In: *Archive for History of Exact Sciences* 66.4, pp. 427–438. DOI: 10.1007/s00407-012-0098-5.

Giannopoulos, G., M. J. Sullivan, K. R. Hartop, G. Rowley, A. J. Gates, N. J. Watmough, and D. J. Richardson (2017). "Tuning the modular *Paracoccus denitrificans* respirome to adapt from aerobic respiration to anaerobic denitrification." In: *Environmental Microbiology* 19.12, pp. 4953–4964. DOI: 10.1111/1462-2920.13974.

Gifford, S. M., S. Sharma, J. M. Rinta-Kanto, and M. A. Moran (2011). "Quantitative analysis of a deeply sequenced marine microbial metatranscriptome." In: *The ISME Journal* 5.3 (3), pp. 461–472. DOI: 10.1038/ismej.2010.141.

Gonzalez-Quiñones, V., E. A. Stockdale, N. C. Banning, F. C. Hoyle, Y. Sawada, A. D. Wherrett, D. L. Jones, and D. V. Murphy (2011). "Soil microbial biomass—interpretation and consideration for soil monitoring." In: *Soil Research* 49.4, p. 287. DOI: 10.1071/SR10203.

Griebler, C., B. Mindl, D. Slezak, and M. Geiger-Kaiser (2002). "Distribution patterns of attached and suspended bacteria in pristine and contaminated shallow aquifers studied with an in situ sediment exposure microcosm." In: *Aquatic Microbial Ecology* 28.2, pp. 117–129. DOI: 10.3354/ame028117.

Grösbacher, M., D. Eckert, O. A. Cirpka, and C. Griebler (2018). "Contaminant concentration versus flow velocity: drivers of biodegradation and microbial growth in groundwater model systems." In: *Biodegradation* 29.3, pp. 211–232. DOI: 10.1007/s10532-018-9824-2.

Grossart, H.-P., R. Massana, K. D. McMahon, and D. A. Walsh (2020). "Linking metagenomics to aquatic microbial ecology and biogeochemical cycles." In: *Limnology and Oceanography* 65.S1, S2–S20. DOI: 10.1002/lno.11382.

Gu, C., G. M. Hornberger, A. L. Mills, J. S. Herman, and S. A. Flewelling (2007). "Nitrate reduction in streambed sediments: effects of flow and biogeochemical kinetics." In: *Water Resources Research* 43.12. DOI: 10.1029/2007WR006027.

Guo, X. et al. (2020). "Gene-informed decomposition model predicts lower soil carbon loss due to persistent microbial adaptation to warming." In: *Nature Communications* 11.1 (1), p. 4897. DOI: 10.1038/s41467-020-18706-z.

Gutenkunst, R. N., J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna (2007). "Universally sloppy parameter sensitivities in systems biology." In: *PLoS Computational Biology* 3.10, e189. DOI: 10.1371/journal.pcbi.0030189. arXiv: q-bio/0701039.

Gutiérrez, M., R. N. Biagioni, M. Teresa Alarcón-Herrera, and B. A. Rivas-Lucero (2018). "An overview of nitrate sources and operating processes in arid and semiarid aquifer systems." In: *Science of the Total Environment* 624, pp. 1513–1522. DOI: 10.1016/j.scitotenv.2017.12.252.

Hallin, S., L. Philippot, F. E. Löffler, R. A. Sanford, and C. M. Jones (2018). "Genomics and ecology of novel N_2O-reducing microorganisms." In: *Trends in Microbiology* 26.1, pp. 43–55. DOI: 10.1016/j.tim.2017.07.003.

Harper, M. (2021). *python-ternary: Ternary Plots in Python*. In collab. with B. Weinstein, Tgwoodcock, C. Simon, Chebee7i, W. Morgan, V. Knight, N. Swanson-Hysell, M. Evans, Jl-Bernal, ZGainsforth, The Gitter Badger, SaxonAnglo, M. Greco, and G. Zuidhof. Version 1.0.8. Zenodo. DOI: 10.5281/zenodo.594435.

Härtig, E. and W. G. Zumft (1999). "Kinetics of *nirS* expression (cytochrome cd_1 nitrite reductase) in *Pseudomonas stutzeri* during the transition from aerobic respiration to denitrifi-

cation: evidence for a denitrification-specific nitrate- and nitrite-responsive regulatory system." In: *Journal of Bacteriology* 181.1, pp. 161–166. DOI: 10.1128/JB.181.1.161-166.1999. PMID: 9864326.

Harvey, J. W., J. K. Böhlke, M. A. Voytek, D. Scott, and C. R. Tobias (2013). "Hyporheic zone denitrification: controls on effective reaction depth and contribution to whole-stream mass balance." In: *Water Resources Research* 49.10, pp. 6298–6316. DOI: 10.1002/wrcr.20492.

Hassan, J., L. L. Bergaust, I. D. Wheat, and L. R. Bakken (2014). "Low probability of initiating *nirS* transcription explains observed gas kinetics and growth of bacteria switching from aerobic respiration to denitrification." In: *PLoS Computational Biology* 10.11. Ed. by R. Fulweiler, e1003933. DOI: 10.1371/journal.pcbi.1003933.

Hassan, J., Z. Qu, L. L. Bergaust, and L. R. Bakken (2016). "Transient accumulation of NO_2^- and N_2O during denitrification explained by assuming cell diversification by stochastic transcription of denitrification genes." In: *PLoS Computational Biology* 12.1. DOI: 10.1371/journal.pcbi.1004621. PMID: 26731685.

Hatzenpichler, R., S. A. Connon, D. Goudeau, R. R. Malmstrom, T. Woyke, and V. J. Orphan (2016). "Visualizing in situ translational activity for identifying and sorting slow-growing archaeal–bacterial consortia." In: *Proceedings of the National Academy of Sciences* 113.28. DOI: 10.1073/pnas.1603757113.

Hayashi, M., T. Vogt, L. Mächler, and M. Schirmer (2012). "Diurnal fluctuations of electrical conductivity in a pre-alpine river: effects of photosynthesis and groundwater exchange." In: *Journal of Hydrology* 450–451, pp. 93–104. DOI: 10.1016/j.jhydrol.2012.05.020.

Hazard, C., J. I. Prosser, and G. W. Nicol (2021). "Use and abuse of potential rates in soil microbiology." In: *Soil Biology and Biochemistry* 157, p. 108242. DOI: 10.1016/j.soilbio.2021.108242.

Hindmarsh, A. C., P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward (2005). "SUNDIALS: suite of nonlinear and differential/algebraic equation solvers." In: *ACM Transactions on Mathematical Software (TOMS)* 31.3, pp. 363–396. DOI: 10.1145/1089014.1089020.

Hoffman, M., P. Sountsov, J. V. Dillon, I. Langmore, D. Tran, and S. Vasudevan (2019). *NeuTralizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport*. arXiv: 1903.03704.

Hoffman, M. D. and A. Gelman (2014). "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *The Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Höge, M., T. Wöhling, and W. Nowak (2018). "A primer for model selection: the decisive role of model complexity." In: *Water Resources Research* 54.3, pp. 1688–1715. DOI: 10.1002/2017WR021902.

Howarth, R. W. (2008). "Coastal nitrogen pollution: A review of sources and trends globally and regionally." In: *Harmful Algae*. HABs and Eutrophication 8.1, pp. 14–20. DOI: 10.1016/j.hal.2008.08.015.

Hsueh, H.-F., A. Guthke, T. Wöhling, and W. Nowak (2022). "Diagnosis of model-structural errors with a sliding time-window Bayesian analysis." In: *Water Resources Research* 58.2, e2021WR030590. DOI: 10.1029/2021WR030590. arXiv: 2107.09399.

Hui, C., Y. Li, W. Zhang, G. Yang, H. Wang, Y. Gao, L. Niu, L. Wang, and H. Zhang (2021). "Coupling genomics and hydraulic information to predict the nitrogen dynamics in a

channel confluence." In: *Environmental Science & Technology* 55.8, pp. 4616–4628. DOI: 10.1021/acs.est.0c04018.

Ingalls, B. P. (2013). *Mathematical Modeling in Systems Biology: An Introduction*. Cambridge, Mass.: MIT Press. 408 pp.

Jackson, K. R. and R. Sacks-Davis (1980). "An alternative implementation of variable step-size multistep formulas for stiff ODEs." In: *ACM Transactions on Mathematical Software* 6.3, pp. 295–318. DOI: 10.1145/355900.355903.

Jiang, S.-Q., Y.-N. Yu, R.-W. Gao, H. Wang, J. Zhang, R. Li, X.-H. Long, Q.-R. Shen, W. Chen, and F. Cai (2019). "High-throughput absolute quantification sequencing reveals the effect of different fertilizer applications on bacterial community in a tomato cultivated coastal saline soil." In: *Science of The Total Environment* 687, pp. 601–609. DOI: 10.1016/j.scitotenv.2019.06.105.

Jin, Q. and C. M. Bethke (2005). "Predicting the rate of microbial respiration in geochemical environments." In: *Geochimica et Cosmochimica Acta* 69.5, pp. 1133–1143. DOI: 10.1016/j.gca.2004.08.010.

Jones, C., B. Stres, M. Rosenquist, and S. Hallin (2008). "Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification." In: *Molecular Biology and Evolution* 25, pp. 1955–66. DOI: 10.1093/molbev/msn146.

Kennedy, C. D., D. P. Genereux, D. R. Corbett, and H. Mitasova (2009). "Spatial and temporal dynamics of coupled groundwater and nitrogen fluxes through a streambed in an agricultural watershed." In: *Water Resources Research* 45.9. DOI: 10.1029/2008WR007397.

Kennedy, M. C. and A. O'Hagan (2001). "Bayesian calibration of computer models." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 425–464. DOI: 10.1111/1467-9868.00294.

Kinzelbach, W., W. Schäfer, and J. Herzer (1991). "Numerical modeling of natural and enhanced denitrification processes in aquifers." In: *Water Resources Research* 27.6, pp. 1123–1135. DOI: 10.1029/91WR00474.

Knights, D., A. H. Sawyer, R. T. Barnes, C. T. Musial, and S. Bray (2017). "Tidal controls on riverbed denitrification along a tidal freshwater zone." In: *Water Resources Research* 53.1, pp. 799–816. DOI: 10.1002/2016WR019405.

Kobyzev, I., S. J. D. Prince, and M. A. Brubaker (2021). "Normalizing flows: an introduction and review of current methods." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11, pp. 3964–3979. DOI: 10.1109/TPAMI.2020.2992934. arXiv: 1908.09257.

Koutinas, M., A. Kiparissides, R. Silva-Rocha, M.-C. Lam, V. A. P. Martins dos Santos, V. de Lorenzo, E. N. Pistikopoulos, and A. Mantalaris (2011). "Linking genes to microbial growth kinetics—An integrated biochemical systems engineering approach." In: *Metabolic Engineering* 13.4, pp. 401–413. DOI: 10.1016/j.ymben.2011.02.001.

Krapu, C., M. Borsuk, and M. Kumar (2019). "Gradient-based inverse estimation for a rainfall-runoff model." In: *Water Resources Research* 55.8, pp. 6625–6639. DOI: 10.1029/2018WR024461.

Krause, S., D. M. Hannah, J. H. Fleckenstein, C. M. Heppell, D. Kaeser, R. Pickup, G. Pinay, A. L. Robertson, and P. J. Wood (2011). "Inter-disciplinary perspectives on processes in the hyporheic zone." In: *Ecohydrology* 4.4, pp. 481–499. DOI: 10.1002/eco.176.

Krause, S., J. Lewandowski, N. B. Grimm, D. M. Hannah, G. Pinay, K. McDonald, E. Martí, A. Argerich, L. Pfister, J. Klaus, T. Battin, S. T. Larned, J. Schelker, J. Fleckenstein, C. Schmidt, M. O. Rivett, G. Watts, F. Sabater, A. Sorolla, and V. Turk (2017). "Ecohydrological interfaces as hot spots of ecosystem processes." In: *Water Resources Research* 53.8, pp. 6359–6376. DOI: 10.1002/2016WR019516.

Kumar, R., C. Carroll, A. Hartikainen, and O. Martin (2019). "ArviZ a unified library for exploratory analysis of Bayesian models in Python." In: *Journal of Open Source Software* 4.33, p. 1143. DOI: 10.21105/joss.01143.

Kunz, J. V., R. Hensley, L. Brase, D. Borchardt, and M. Rode (2017). "High frequency measurements of reach scale nitrogen uptake in a fourth order river with contrasting hydromorphology and variable water chemistry (Weiße Elster, Germany)." In: *Water Resources Research* 53.1, pp. 328–343. DOI: 10.1002/2016WR019355.

Kuypers, M. M. M., H. K. Marchant, and B. Kartal (2018). "The microbial nitrogen-cycling network." In: *Nature Reviews Microbiology* 16.5, pp. 263–276. DOI: 10.1038/nrmicro.2018.9.

Kuzyakov, Y. and E. Blagodatskaya (2015). "Microbial hotspots and hot moments in soil: concept & review." In: *Soil Biology and Biochemistry* 83, pp. 184–199. DOI: 10.1016/j.soilbio.2015.01.025.

Lam, S. K., A. Pitrou, and S. Seibert (2015). "Numba: a LLVM-based Python JIT compiler." In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (Austin, Texas). LLVM '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. DOI: 10.1145/2833157.2833162.

Lansdown, K., M. Trimmer, C. M. Heppell, F. Sgouridis, S. Ullah, A. L. Heathwaite, A. Binley, and H. Zhang (2012). "Characterization of the key pathways of dissimilatory nitrate reduction and their response to complex organic substrates in hyporheic sediments." In: *Limnology and Oceanography* 57.2, pp. 387–400. DOI: 10.4319/lo.2012.57.2.0387.

Lee, Y.-Y., N. Shearer, and S. Spiro (2006). "Transcription factor NNR from *Paracoccus denitrificans* is a sensor of both nitric oxide and oxygen: isolation of *nnr** alleles encoding effector-independent proteins and evidence for a haem-based sensing mechanism." In: *Microbiology* 152.5, pp. 1461–1470. DOI: 10.1099/mic.0.28796-0.

Lennon, J. T., M. E. Muscarella, S. A. Placella, and B. K. Lehmkuhl (2018). "How, when, and where relic DNA affects microbial diversity." In: *mBio* 9.3. Ed. by J. Zhou. DOI: 10.1128/mBio.00637-18.

Lewandowski, J. et al. (2019). "Is the hyporheic zone relevant beyond the scientific community?" In: *Water* 11.11 (11), p. 2230. DOI: 10.3390/w11112230.

Li, L. et al. (2017). "Expanding the role of reactive transport models in critical zone processes." In: *Earth-Science Reviews* 165, pp. 280–301. DOI: 10.1016/j.earscirev.2016.09.001.

Li, M., Y. Gao, W.-J. Qian, L. Shi, Y. Liu, W. C. Nelson, C. D. Nicora, C. T. Resch, C. Thompson, S. Yan, J. K. Fredrickson, J. M. Zachara, and C. Liu (2017a). "Targeted quantification of functional enzyme dynamics in environmental samples for microbially mediated biogeochemical processes." In: *Environmental Microbiology Reports* 9.5, pp. 512–521. DOI: 10.1111/1758-2229.12558.

Li, M., W.-J. Qian, Y. Gao, L. Shi, and C. Liu (2017b). "Functional enzyme-based approach for linking microbial community functions with biogeochemical process kinetics." In:

Environmental Science & Technology 51.20, pp. 11848–11857. DOI: 10.1021/acs.est.7b03158.

Li, Y., P. Lu, and G. Zhang (2021). "An artificial-neural-network-based surrogate modeling workflow for reactive transport modeling." In: *Petroleum Research*. DOI: 10.1016/j.ptlrs.2021.06.002.

Linde, N., D. Ginsbourger, J. Irving, F. Nobile, and A. Doucet (2017). "On uncertainty quantification in hydrogeology and hydrogeophysics." In: *Advances in Water Resources* 110, pp. 166–181. DOI: 10.1016/j.advwatres.2017.10.014.

Liu, C. and J. M. Zachara (2001). "Uncertainties of Monod kinetic parameters nonlinearly estimated from batch experiments." In: *Environmental Science & Technology* 35.1, pp. 133–141. DOI: 10.1021/es001261b.

Liu, L. and T. L. Greaver (2009). "A review of nitrogen enrichment effects on three biogenic GHGs: the CO_2 sink may be largely offset by stimulated N_2O and CH_4 emission." In: *Ecology Letters* 12.10, pp. 1103–1117. DOI: 10.1111/j.1461-0248.2009.01351.x.

Liu, Y., C. Liu, W. C. Nelson, L. Shi, F. Xu, Y. Liu, A. Yan, L. Zhong, C. Thompson, J. K. Fredrickson, and J. M. Zachara (2017). "Effect of water chemistry and hydrodynamics on nitrogen transformation activity and microbial community functional potential in hyporheic zone sediment columns." In: *Environmental Science & Technology* 51.9, pp. 4877–4886. DOI: 10.1021/acs.est.6b05018.

Loferer-Krößbacher, M., J. Klima, and R. Psenner (1998). "Determination of bacterial cell dry mass by transmission electron microscopy and densitometric image analysis." In: *Applied and Environmental Microbiology* 64.2, pp. 688–694. DOI: 10.1128/AEM.64.2.688-694.1998.

Loschko, M., T. Wöhling, D. L. Rudolph, and O. A. Cirpka (2018). "Accounting for the decreasing reaction potential of heterogeneous aquifers in a stochastic framework of aquifer-scale reactive transport." In: *Water Resources Research* 54.1, pp. 442–463. DOI: 10.1002/2017WR021645.

Louca, S., A. K. Hawley, S. Katsev, M. Torres-Beltran, M. P. Bhatia, S. Kheirandish, C. C. Michiels, D. Capelle, G. Lavik, M. Doebeli, S. A. Crowe, and S. J. Hallam (2016). "Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone." In: *Proceedings of the National Academy of Sciences* 113.40, E5925–E5933. DOI: 10.1073/pnas.1602897113.

Lui, L. M., E. L.-W. Majumder, H. J. Smith, H. K. Carlson, F. von Netzer, M. W. Fields, D. A. Stahl, J. Zhou, T. C. Hazen, N. S. Baliga, P. D. Adams, and A. P. Arkin (2021). "Mechanism across scales: A holistic modeling framework integrating laboratory and field studies for microbial ecology." In: *Frontiers in Microbiology* 12. DOI: 10.3389/fmicb.2021.642422.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK ; New York: Cambridge University Press. 628 pp.

Maier, T., A. Schmidt, M. Güell, S. Kühner, A.-C. Gavin, R. Aebersold, and L. Serrano (2011). "Quantification of mRNA and protein and integration with protein turnover in a bacterium." In: *Molecular Systems Biology* 7, p. 511. DOI: 10.1038/msb.2011.38. PMID: 21772259.

Mailloux, B. J. and M. E. Fuller (2003). "Determination of in situ bacterial growth rates in aquifers and aquifer sediments." In: *Applied and Environmental Microbiology* 69.7, pp. 3798–3808. DOI: 10.1128/AEM.69.7.3798-3808.2003.

Marlow, J., R. Spietz, K.-Y. Kim, M. Ellisman, P. Girguis, and R. Hatzenpichler (2021). "Spatially resolved correlative microscopy and microbial identification reveal dynamic depth- and mineral-dependent anabolic activity in salt marsh sediment." In: *Environmental Microbiology* 23.8, pp. 4756–4777. DOI: 10.1111/1462-2920.15667.

Marmonier, P., D. Fontvieille, J. Gibert, and V. Vanek (1995). "Distribution of dissolved organic carbon and bacteria at the interface between the Rhône river and its alluvial aquifer." In: *Journal of the North American Benthological Society* 14.3, pp. 382–392. DOI: 10.2307/1467204.

Marschmann, G. L., H. Pagel, P. Kügler, and T. Streck (2019). "Equifinality, sloppiness, and emergent structures of mechanistic soil biogeochemical models." In: *Environmental Modelling & Software* 122, p. 104518. DOI: 10.1016/j.envsoft.2019.104518.

Meile, C. and T. D. Scheibe (2019). "Reactive transport modeling of microbial dynamics." In: *Elements* 15.2, pp. 111–116. DOI: 10.2138/gselements.15.2.111.

Mellage, A., D. Eckert, M. Grösbacher, A. Z. Inan, O. A. Cirpka, and C. Griebler (2015). "Dynamics of suspended and attached aerobic toluene degraders in small-scale flow-through sediment systems under growth and starvation conditions." In: *Environmental Science & Technology* 49.12, pp. 7161–7169. DOI: 10.1021/es5058538.

Mellage, A., C. M. Smeaton, A. Furman, E. A. Atekwana, F. Rezanezhad, and P. V. Cappellen (2019). "Bacterial Stern layer diffusion: experimental determination with spectral induced polarization and sensitivity to nitrite toxicity." In: *Near Surface Geophysics* 17.6, pp. 623–635. DOI: 10.1002/nsg.12058.

Mellage, A., C. M. Smeaton, A. Furman, E. A. Atekwana, F. Rezanezhad, and P. Van Cappellen (2018). "Linking spectral induced polarization (SIP) and subsurface microbial processes: results from sand column incubation experiments." In: *Environmental Science & Technology* 52.4, pp. 2081–2090. DOI: 10.1021/acs.est.7b04420.

Michaelis, L. and M. L. Menten (2011). "The kinetics of invertase action." Trans. by K. A. Johnson and R. S. Goody. In: *Biochemistry* 50.39, pp. 8264–8269. DOI: 10.1021/bi201284u. Trans. of "Die Kinetik der Invertinwirkung." In: *Biochemische Zeitschrift* 49 (1913), pp. 339–369.

Monard, C., F. Martin-Laurent, O. Lima, M. Devers-Lamrani, and F. Binet (2013). "Estimating the biodegradation of pesticide in soils by monitoring pesticide-degrading gene expression." In: *Biodegradation* 24.2, pp. 203–213. DOI: 10.1007/s10532-012-9574-5.

Monnahan, C. C., J. T. Thorson, and T. A. Branch (2017). "Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo." In: *Methods in Ecology and Evolution* 8.3, pp. 339–348. DOI: 10.1111/2041-210X.12681.

Monod, J. (1949). "The growth of bacterial cultures." In: *Annual Review of Microbiology* 3.1, pp. 371–394. DOI: 10.1146/annurev.mi.03.100149.002103.

Moran, M. A., B. Satinsky, S. M. Gifford, H. Luo, A. Rivers, L.-K. Chan, J. Meng, B. P. Durham, C. Shen, V. A. Varaljay, C. B. Smith, P. L. Yager, and B. M. Hopkinson (2013). "Sizing up metatranscriptomics." In: *The ISME journal* 7.2, pp. 237–243. DOI: 10.1038/ismej.2012.94. PMID: 22931831.

Murray, A. M., J. Maillard, B. Jin, M. M. Broholm, C. Holliger, and M. Rolle (2019). "A modeling approach integrating microbial activity, mass transfer, and geochemical processes to interpret biological assays: an example for PCE degradation in a multi-phase batch setup." In: *Water Research* 160, pp. 484–496. DOI: 10.1016/j.watres.2019.05.087.

Nicolaisen, M. H., J. Bælum, C. S. Jacobsen, and J. Sørensen (2008). "Transcription dynamics of the functional *tfdA* gene during MCPA herbicide degradation by *Cupriavidus necator* AEO106 (pRO101) in agricultural soil." In: *Environmental Microbiology* 10.3, pp. 571–579. DOI: 10.1111/j.1462-2920.2007.01476.x.

Nielsen, K. M., P. J. Johnsen, D. Bensasson, and D. Daffonchio (2007). "Release and persistence of extracellular DNA in the environment." In: *Environmental Biosafety Research* 6.1-2, pp. 37–53. DOI: 10.1051/ebr:2007031.

Nokhal, T.-H. and H. G. Schlegel (1983). "Taxonomic study of *Paracoccus denitrificans*." In: *International Journal of Systematic Bacteriology* 33.1, pp. 26–37. DOI: 10.1099/00207713-33-1-26.

O'Hagan, A. (2013). "Bayesian inference with misspecified models: Inference about what?" In: *Journal of Statistical Planning and Inference* 143.10, pp. 1643–1648. DOI: 10.1016/j.jspi.2013.05.016.

Pagel, H., C. Poll, J. Ingwersen, E. Kandeler, and T. Streck (2016). "Modeling coupled pesticide degradation and organic matter turnover: from gene abundance to process rates." In: *Soil Biology and Biochemistry* 103, pp. 349–364. DOI: 10.1016/j.soilbio.2016.09.014.

Pan, Y., X. Zeng, H. Xu, Y. Sun, D. Wang, and J. Wu (2020). "Assessing human health risk of groundwater DNAPL contamination by quantifying the model structure uncertainty." In: *Journal of Hydrology* 584, p. 124690. DOI: 10.1016/j.jhydrol.2020.124690.

Papamakarios, G., E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan (2021). "Normalizing flows for probabilistic modeling and inference." In: *Journal of Machine Learning Research* 22.57, pp. 1–64. arXiv: 1912.02762.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay (2011). "Scikit-learn: machine learning in Python." In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Philippot, L. (2006). "Use of functional genes to quantify denitrifiers in the environment." In: *Biochemical Society Transactions* 34.1, pp. 101–103. DOI: 10.1042/BST0340101. PMID: 16417493.

Picioreanu, C., M. C. M. van Loosdrecht, and J. J. Heijnen (1997). "Modelling the effect of oxygen concentration on nitrite accumulation in a biofilm airlift suspension reactor." In: *Water Science and Technology* 36.1, pp. 147–156. DOI: 10.2166/wst.1997.0034.

Pronk, G. J., A. Mellage, T. Milojevic, C. M. Smeaton, K. Engel, J. D. Neufeld, F. Rezanezhad, and P. V. Cappellen (2020). "Carbon turnover and microbial activity in an artificial soil under imposed cyclic drainage and imbibition." In: *Vadose Zone Journal* 19.1, e20021. DOI: 10.1002/vzj2.20021.

Qu, Z., L. R. Bakken, L. Molstad, Å. Frostegård, and L. L. Bergaust (2015). "Transcriptional and metabolic regulation of denitrification in *Paracoccus denitrificans* allows low but significant activity of nitrous oxide reductase under oxic conditions." In: *Environmental Microbiology* 18.9, pp. 2951–2963. DOI: 10.1111/1462-2920.13128.

Rahm, B. G. and R. E. Richardson (2008). "*Dehalococcoides*' gene transcripts as quantitative bioindicators of tetrachloroethene, trichloroethene, and *cis*-1,2-dichloroethene dehalorespiration rates." In: *Environmental Science & Technology* 42.14, pp. 5099–5105. DOI: 10.1021/es702912t.

Ramkrishna, D. and H.-S. Song (2019). *Cybernetic Modeling for Bioreaction Engineering*. Cambridge Series in Chemical Engineering. Cambridge; New York, NY: Cambridge University Press. DOI: 10.1017/9780511731969.

Reed, D. C., C. K. Algar, J. A. Huber, and G. J. Dick (2014). "Gene-centric approach to integrating environmental genomics and biogeochemical models." In: *Proceedings of the National Academy of Sciences* 111.5, pp. 1879–1884. DOI: 10.1073/pnas.1313713111. PMID: 24449851.

Reed, D. C., J. A. Breier, H. Jiang, K. Anantharaman, C. A. Klausmeier, B. M. Toner, C. Hancock, K. Speer, A. M. Thurnherr, and G. J. Dick (2015). "Predicting the response of the deep-ocean microbiome to geochemical perturbations by hydrothermal vents." In: *The ISME Journal* 9.8, pp. 1857–1869. DOI: 10.1038/ismej.2015.4.

Reichert, P. and J. Mieleitner (2009). "Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters: analyzing input and structural uncertainty." In: *Water Resources Research* 45.10. DOI: 10.1029/2009WR007814.

Reid, T., S. R. Chaganti, I. G. Droppo, and C. G. Weisener (2018). "Novel insights into freshwater hydrocarbon-rich sediments using metatranscriptomics: opening the black box." In: *Water Research* 136, pp. 1–11. DOI: 10.1016/j.watres.2018.02.039.

Roberts, E., A. Magis, J. O. Ortiz, W. Baumeister, and Z. Luthey-Schulten (2011). "Noise contributions in an inducible genetic switch: a whole-cell simulation study." In: *PLoS Computational Biology* 7.3. DOI: 10.1371/journal.pcbi.1002010. PMID: 21423716.

Robinson, J. A. and J. M. Tiedje (1983). "Nonlinear estimation of Monod growth kinetic parameters from a single substrate depletion curve." In: *Applied and Environmental Microbiology* 45.5, pp. 1453–1458. DOI: 10.1128/aem.45.5.1453-1458.1983.

Rocca, J. D., E. K. Hall, J. T. Lennon, S. E. Evans, M. P. Waldrop, J. B. Cotner, D. R. Nemergut, E. B. Graham, and M. D. Wallenstein (2015). "Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed." In: *The ISME Journal* 9.8, pp. 1693–1699. DOI: 10.1038/ismej.2014.252.

Rode, M., G. Arhonditsis, D. Balin, T. Kebede, V. Krysanova, A. van Griensven, and S. E. A. T. M. van der Zee (2010). "New challenges in integrated water quality modelling." In: *Hydrological Processes* 24.24, pp. 3447–3461. DOI: 10.1002/hyp.7766.

Roden, E. E. and Q. Jin (2011). "Thermodynamics of microbial growth coupled to metabolism of glucose, ethanol, short-chain organic acids, and hydrogen." In: *Applied and Environmental Microbiology* 77.5, pp. 1907–1909. DOI: 10.1128/AEM.02425-10. PMID: 21216913.

Roels, J. A. (1983). *Energetics and Kinetics in Biotechnology*. Amsterdam; New York: Elsevier Biomedical Press. 330 pp.

Rohe, L., T. Oppermann, R. Well, and M. A. Horn (2020). "Nitrite induced transcription of *p450nor* during denitrification by *Fusarium oxysporum* correlates with the production of N_2O with a high ^{15}N site preference." In: *Soil Biology and Biochemistry* 151, p. 108043. DOI: 10.1016/j.soilbio.2020.108043.

Rubinstein, R. L., M. A. Borton, H. Zhou, M. Shaffer, D. W. Hoyt, J. Stegen, C. S. Henry, K. C. Wrighton, and R. Versteeg (2022). "ORT: a workflow linking genome-scale metabolic models with reactive transport codes." In: *Bioinformatics* 38.3, pp. 778–784. DOI: 10.1093/bioinformatics/btab753. bioRxiv: 2021.03.02.433463.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2007). *Global Sensitivity Analysis. The Primer*. Chichester, UK: John Wiley & Sons, Ltd. DOI: 10.1002/9780470725184.

Salvatier, J., T. V. Wiecki, and C. Fonnesbeck (2016). "Probabilistic programming in Python using PyMC3." In: *PeerJ Computer Science* 2, e55. DOI: 10.7717/peerj-cs.55.

Sander, R. (2015). "Compilation of Henry's law constants (version 4.0) for water as solvent." In: *Atmospheric Chemistry and Physics* 15.8, pp. 4399–4981. DOI: 10.5194/acp-15-4399-2015.

Sanford, R. A., D. D. Wagner, Q. Wu, J. C. Chee-Sanford, S. H. Thomas, C. Cruz-García, G. Rodríguez, A. Massol-Deyá, K. K. Krishnani, K. M. Ritalahti, S. Nissen, K. T. Konstantinidis, and F. E. Löffler (2012). "Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils." In: *Proceedings of the National Academy of Sciences* 109.48, pp. 19709–19714. DOI: 10.1073/pnas.1211238109. PMID: 23150571.

Sanz-Prat, A., C. Lu, R. T. Amos, M. Finkel, D. W. Blowes, and O. A. Cirpka (2016). "Exposure-time based modeling of nonlinear reactive transport in porous media subject to physical and geochemical heterogeneity." In: *Journal of Contaminant Hydrology* 192, pp. 35–49. DOI: 10.1016/j.jconhyd.2016.06.002.

Sanz-Prat, A., C. Lu, M. Finkel, and O. A. Cirpka (2015). "On the validity of travel-time based nonlinear bioreactive transport models in steady-state flow." In: *Journal of Contaminant Hydrology* 175–176, pp. 26–43. DOI: 10.1016/j.jconhyd.2015.02.003.

Sargsyan, K., X. Huan, and H. N. Najm (2019). "Embedded model error representation for Bayesian model calibration." In: *International Journal for Uncertainty Quantification* 9.4. DOI: 10.1615/Int.J.UncertaintyQuantification.2019027384. arXiv: 1801.06768.

Sargsyan, K., H. N. Najm, and R. Ghanem (2015). "On the statistical calibration of physical models." In: *International Journal of Chemical Kinetics* 47.4, pp. 246–276. DOI: 10.1002/kin.20906.

Sawyer, A. H. (2015). "Enhanced removal of groundwater-borne nitrate in heterogeneous aquatic sediments." In: *Geophysical Research Letters* 42.2, pp. 403–410. DOI: 10.1002/2014GL062234.

Sawyer, A. H., M. B. Cardenas, A. Bomar, and M. Mackey (2009). "Impact of dam operations on hyporheic exchange in the riparian zone of a regulated river." In: *Hydrological Processes* 23.15, pp. 2129–2137. DOI: 10.1002/hyp.7324.

Scheibe, T. D., R. Mahadevan, Y. Fang, S. Garg, P. E. Long, and D. R. Lovley (2009). "Coupling a genome-scale metabolic model with a reactive transport model to describe in situ uranium bioremediation." In: *Microbial Biotechnology* 2.2, pp. 274–286. DOI: 10.1111/j.1751-7915.2009.00087.x.

Scheidegger, A. E. (1974). *The Physics of Flow through Porous Media*. 3rd ed. Toronto: University of Toronto Press. 372 pp. DOI: 10.3138/9781487583750.

Scheurer, S., A. Schäfer Rodrigues Silva, F. Mohammadi, J. Hommel, S. Oladyshkin, B. Flemisch, and W. Nowak (2021). "Surrogate-based Bayesian comparison of computationally expensive models: application to microbially induced calcite precipitation." In: *Computational Geosciences* 25.6, pp. 1899–1917. DOI: 10.1007/s10596-021-10076-9.

Schweitzer, H. D., H. J. Smith, E. P. Barnhart, L. J. McKay, R. Gerlach, A. B. Cunningham, R. R. Malmstrom, D. Goudeau, and M. W. Fields (2022). "Subsurface hydrocarbon degradation strategies in low- and high-sulfate coal seam communities identified with activity-based metagenomics." In: *npj Biofilms and Microbiomes* 8.1 (1), pp. 1–10. DOI: 10.1038/s41522-022-00267-2.

Seyboldt, A. (2020). *Sunode*. In collab. with A. Störiko, M. Osthege, and L. Widmer. Version 0.1.2. Zenodo. DOI: 10.5281/zenodo.4058330.

Seyboldt, A. (2021). *Sunode*. In collab. with A. Störiko, M. Osthege, and L. Widmer. Version 0.2.2. Zenodo. DOI: 10.5281/zenodo.5213947.

Siade, A. J., B. C. Bostick, O. A. Cirpka, and H. Prommer (2021). "Unraveling biogeochemical complexity through better integration of experiments and modeling." In: *Environmental Science: Processes & Impacts* 23.12, pp. 1825–1833. DOI: 10.1039/D1EM00303H.

Sierra, C. A., S. Malghani, and M. Müller (2015). "Model structure and parameter identification of soil organic matter models." In: *Soil Biology and Biochemistry* 90, pp. 197–203. DOI: 10.1016/j.soilbio.2015.08.012.

Simon, J. and M. G. Klotz (2013). "Diversity and evolution of bioenergetic systems involved in microbial nitrogen compound transformations." In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. The Evolutionary Aspects of Bioenergetic Systems 1827.2, pp. 114–135. DOI: 10.1016/j.bbabio.2012.07.005.

Smeaton, C. M. and P. Van Cappellen (2018). "Gibbs energy dynamic yield method (GEDYM): predicting microbial growth yields under energy-limiting conditions." In: *Geochimica et Cosmochimica Acta* 241, pp. 1–16. DOI: 10.1016/j.gca.2018.08.023.

Smith, H. J., A. J. Zelaya, K. B. De León, R. Chakraborty, D. A. Elias, T. C. Hazen, A. P. Arkin, A. B. Cunningham, and M. W. Fields (2018). "Impact of hydrologic boundaries on microbial planktonic and biofilm communities in shallow terrestrial subsurface environments." In: *FEMS Microbiology Ecology* 94.12, fiy191. DOI: 10.1093/femsec/fiy191. PMID: 30265315.

Smith, T., L. Marshall, and A. Sharma (2015). "Modeling residual hydrologic errors with Bayesian inference." In: *Journal of Hydrology* 528, pp. 29–37. DOI: 10.1016/j.jhydrol.2015.05.051.

Sokol, N. W. et al. (2022). "Life and death in the soil microbiome: how ecological processes influence biogeochemistry." In: *Nature Reviews Microbiology*. DOI: 10.1038/s41579-022-00695-z.

Song, H.-S. and C. Liu (2015). "Dynamic metabolic modeling of denitrifying bacterial growth: the cybernetic approach." In: *Industrial & Engineering Chemistry Research* 54.42, pp. 10221–10227. DOI: 10.1021/acs.iecr.5b01615.

Song, H.-S., J. C. Stegen, E. B. Graham, J.-Y. Lee, V. A. Garayburu-Caruso, W. C. Nelson, X. Chen, J. D. Moulton, and T. D. Scheibe (2020). "Representing organic matter thermodynamics in biogeochemical reactions via substrate-explicit modeling." In: *Frontiers in Microbiology* 11. DOI: 10.3389/fmicb.2020.531756.

Song, H.-S., D. G. Thomas, J. C. Stegen, M. Li, C. Liu, X. Song, X. Chen, J. K. Fredrickson, J. M. Zachara, and T. D. Scheibe (2017). "Regulation-structured dynamic metabolic model provides a potential mechanism for delayed enzyme response in denitrification process." In: *Frontiers in Microbiology* 8. DOI: 10.3389/fmicb.2017.01866.

Spalding, R. F. and M. E. Exner (1993). "Occurrence of nitrate in groundwater—a review." In: *Journal of Environmental Quality* 22.3, pp. 392–402. DOI: 10.2134/jeq1993.00472425002200030002x.

Spiro, S. (2012). "Nitrous oxide production and consumption: regulation of gene expression by gas-sensitive transcription factors." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1593, pp. 1213–1225. DOI: 10.1098/rstb.2011.0309.

– (2017). "Regulation of denitrification." In: *Metalloenzymes in Denitrification: Applications and Environmental Impacts*. Ed. by I. Moura, J. J. G. Moura, S. R. Pauleta, and L. B. Maia.

Metallobiology 9. The Royal Society of Chemistry, pp. 312–330. DOI: 10.1039/9781782623762-00312.

Starke, R., N. Jehmlich, and F. Bastida (2019). "Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics." In: *Journal of Proteomics*. 10 Year Anniversary of Proteomics 198, pp. 50–58. DOI: 10.1016/j.jprot.2018.11.011.

Steefel, C. I., D. J. DePaolo, and P. C. Lichtner (2005). "Reactive transport modeling: An essential tool and a new research approach for the Earth sciences." In: *Earth and Planetary Science Letters* 240.3, pp. 539–558. DOI: 10.1016/j.epsl.2005.09.017.

Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, and S. Sörlin (2015). "Planetary boundaries: guiding human development on a changing planet." In: *Science* 347.6223. DOI: 10.1126/science.1259855. PMID: 25592418.

Stelzer, R. S., L. A. Bartsch, W. B. Richardson, and E. A. Strauss (2011). "The dark side of the hyporheic zone: depth profiles of nitrogen and its processing in stream sediments." In: *Freshwater Biology* 56.10, pp. 2021–2033. DOI: 10.1111/j.1365-2427.2011.02632.x.

Stewart, E. J. (2012). "Growing unculturable bacteria." In: *Journal of Bacteriology* 194.16, pp. 4151–4160. DOI: 10.1128/JB.00345-12.

Stief, P., D. Beer, and D. Neumann (2002). "Small-scale distribution of interstitial nitrite in freshwater sediment microcosms: the role of nitrate and oxygen availability, and sediment permeability." In: *Microbial Ecology* 43.3, pp. 367–377. DOI: 10.1007/s00248-002-2008-x.

Stoliker, D. L., D. A. Repert, R. L. Smith, B. Song, D. R. LeBlanc, T. D. McCobb, C. H. Conaway, S. P. Hyun, D.-C. Koh, H. S. Moon, and D. B. Kent (2016). "Hydrologic controls on nitrogen cycling processes and functional gene abundance in sediments of a groundwater flow-through lake." In: *Environmental Science & Technology* 50.7, pp. 3649–3657. DOI: 10.1021/acs.est.5b06155.

Störiko, A. (2021). *Adrpy: Advection-Dispersion-Reaction Models in Python*. Version 0.1.0. Zenodo. DOI: 10.5281/zenodo.5590973.

– (2022). *Quantifying Uncertainty of Hyporheic Nitrogen Cycling: Modeling Code and Data*. Version 0.0.1. Zenodo. DOI: 10.5281/zenodo.7460717.

Störiko, A., H. Pagel, A. Mellage, and O. A. Cirpka (2021a). "Does it pay off to explicitly link functional gene expression to denitrification rates in reaction models?" In: *Frontiers in Microbiology* 12. DOI: 10.3389/fmicb.2021.684146.

– (2021b). *Nitrogene: Modelling Code and Data of an Enzyme-Based Denitrification Model*. Version 0.0.2. Zenodo. DOI: 10.5281/zenodo.4620167.

Störiko, A., H. Pagel, A. Mellage, P. Van Cappellen, and O. A. Cirpka (2022). "Denitrification-driven transcription and enzyme production at the river-groundwater interface: insights from reactive-transport modeling." In: *Water Resources Research* 58.8, e2021WR031584. DOI: 10.1029/2021WR031584.

Stouthamer, A. H. (1980). "Bioenergetic studies on *Paracoccus denitrificans*." In: *Trends in Biochemical Sciences* 5.6, pp. 164–166. DOI: 10.1016/0968-0004(80)90015-8.

Suenaga, T., S. Riya, M. Hosomi, and A. Terada (2018). "Biokinetic characterization and activities of N_2O-reducing bacteria in response to various oxygen levels." In: *Frontiers in Microbiology* 9. DOI: 10.3389/fmicb.2018.00697.

Sun, X., X. Zeng, J. Wu, and D. Wang (2021). "A two-stage Bayesian data-driven method to improve model prediction." In: *Water Resources Research* 57.12, e2021WR030436. DOI: 10.1029/2021WR030436.

Thullner, M. and P. Regnier (2019). "Microbial controls on the biogeochemical dynamics in the subsurface." In: *Reviews in Mineralogy and Geochemistry* 85.1, pp. 265–302. DOI: 10.2138/rmg.2019.85.9.

Transtrum, M. K. and P. Qiu (2014). "Model reduction by manifold boundaries." In: *Physical Review Letters* 113.9, p. 098701. DOI: 10.1103/PhysRevLett.113.098701.

Triska, F. J., A. P. Jackman, J. H. Duff, and R. J. Avanzino (1994). "Ammonium sorption to channel and riparian sediments: a transient storage pool for dissolved inorganic nitrogen." In: *Biogeochemistry* 26.2, pp. 67–83. DOI: 10.1007/BF02182880.

Van de Schoot, R., S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau (2021). "Bayesian statistics and modelling." In: *Nature Reviews Methods Primers* 1.1, pp. 1–26. DOI: 10.1038/s43586-020-00001-2.

Van Verseveld, H. W., E. M. Meijer, and A. H. Stouthamer (1977). "Energy conservation during nitrate respiration in *Paracoccus denitrificans*." In: *Archives of Microbiology* 112.1, pp. 17–23. DOI: 10.1007/BF00446649.

Vavilin, V. A., B. Fernandez, J. Palatsi, and X. Flotats (2008). "Hydrolysis kinetics in anaerobic degradation of particulate organic material: an overview." In: *Waste Management* 28.6, pp. 939–951. DOI: 10.1016/j.wasman.2007.03.028.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner (2021). "Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC." In: *Bayesian Analysis* 16.2, pp. 667–718. DOI: 10.1214/20-BA1221.

Villaverde, A. F. (2019). "Observability and structural identifiability of nonlinear biological systems." In: *Complexity* 2019, e8497093. DOI: 10.1155/2019/8497093.

Virtanen, P. et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python." In: *Nature Methods* 17.3, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, G. and S. Chen (2012). "A review on parameterization and uncertainty in modeling greenhouse gas emissions from soil." In: *Geoderma* 170, pp. 206–216. DOI: 10.1016/j.geoderma.2011.11.009.

– (2013). "Evaluation of a soil greenhouse gas emission model based on Bayesian inference and MCMC: model uncertainty." In: *Ecological Modelling* 253, pp. 97–106. DOI: 10.1016/j.ecolmodel.2012.09.010.

Wang, S., W. Wang, S. Zhao, X. Wang, M. M. Hefting, L. Schwark, and G. Zhu (2019). "Anammox and denitrification separately dominate microbial N-loss in water saturated and unsaturated soils horizons of riparian zones." In: *Water Research* 162, pp. 139–150. DOI: 10.1016/j.watres.2019.06.052.

Wang, Y. and P. Van Cappellen (1996). "A multicomponent reactive transport model of early diagenesis: application to redox cycling in coastal marine sediments." In: *Geochimica et Cosmochimica Acta* 60.16, pp. 2993–3014. DOI: 10.1016/0016-7037(96)00140-8.

Wegner, C.-E., M. Gaspar, P. Geesink, M. Herrmann, M. Marz, and K. Küsel (2019). "Biogeochemical regimes in shallow aquifers reflect the metabolic coupling of the elements nitrogen, sulfur, and carbon." In: *Applied and Environmental Microbiology* 85.5. DOI: 10.1128/AEM.02346-18. PMID: 30578263.

Welsh, A., J. C. Chee-Sanford, L. M. Connor, F. E. Löffler, and R. A. Sanford (2014). "Refined NrfA phylogeny improves PCR-based *nrfA* gene detection." In: *Appl. Environ. Microbiol.* 80.7, pp. 2110–2119. DOI: 10.1128/AEM.03443-13. PMID: 24463965.

Westrich, J. T. and R. A. Berner (1984). "The role of sedimentary organic matter in bacterial sulfate reduction: the G model tested." In: *Limnology and Oceanography* 29.2, pp. 236–249. DOI: 10.4319/lo.1984.29.2.0236.

Wood, N. J., T. Alizadeh, S. Bennett, J. Pearce, S. J. Ferguson, D. J. Richardson, and J. W. B. Moir (2001). "Maximal expression of membrane-bound nitrate reductase in *Paracoccus* is induced by nitrate via a third FNR-like regulator named NarR." In: *Journal of Bacteriology* 183.12, pp. 3606–3613. DOI: 10.1128/JB.183.12.3606-3613.2001. PMID: 11371524.

Wu, Z., Y. Liu, Z. Liang, S. Wu, and H. Guo (2017). "Internal cycling, not external loading, decides the nutrient limitation in eutrophic lake: a dynamic model with temporal Bayesian hierarchical inference." In: *Water Research* 116, pp. 231–240. DOI: 10.1016/j.watres.2017.03.039.

Xu, T., A. J. Valocchi, M. Ye, and F. Liang (2017). "Quantifying model structural error: efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model." In: *Water Resources Research* 53.5, pp. 4084–4105. DOI: 10.1002/2016WR019831.

Yan, S., Y. Liu, C. Liu, L. Shi, J. Shang, H. Shan, J. Zachara, J. Fredrickson, D. Kennedy, C. T. Resch, C. Thompson, and S. Fansler (2016). "Nitrate bioreduction in redox-variable low permeability sediments." In: *Science of The Total Environment* 539, pp. 185–195. DOI: 10.1016/j.scitotenv.2015.08.122.

Ying, S., J. Zhang, L. Zeng, J. Shi, and L. Wu (2017). "Bayesian inference for kinetic models of biotransformation using a generalized rate equation." In: *Science of The Total Environment* 590–591, pp. 287–296. DOI: 10.1016/j.scitotenv.2017.03.003.

Zarnetske, J. P., R. Haggerty, S. M. Wondzell, and M. A. Baker (2011). "Labile dissolved organic carbon supply limits hyporheic denitrification." In: *Journal of Geophysical Research: Biogeosciences* 116.G4. DOI: 10.1029/2011JG001730.

Zarnetske, J. P., R. Haggerty, S. M. Wondzell, V. A. Bokil, and R. González-Pinzón (2012). "Coupled transport and reaction kinetics control the nitrate source-sink function of hyporheic zones." In: *Water Resources Research* 48.11. DOI: 10.1029/2012WR011894.

Zemb, O., C. S. Achard, J. Hamelin, M.-L. D. Almeida, B. Gabinaud, L. Cauquil, L. M. G. Verschuren, and J.-J. Godon (2020). "Absolute quantitation of microbes using 16S rRNA gene metabarcoding: a rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard." In: *MicrobiologyOpen* 9.3, e977. DOI: 10.1002/mbo3.977.

Zhang, B., C. R. Penton, Z. Yu, C. Xue, Q. Chen, Z. Chen, C. Yan, Q. Zhang, M. Zhao, J. F. Quensen, and J. M. Tiedje (2021). "A new primer set for Clade I *nosZ* that recovers genes from a broader range of taxa." In: *Biology and Fertility of Soils* 57.4, pp. 523–531. DOI: 10.1007/s00374-021-01544-6.

Zhang, H., H. Fu, J. Wang, L. Sun, Y. Jiang, L. Zhang, and H. Gao (2013). "Impacts of nitrate and nitrite on physiology of *Shewanella oneidensis*." In: *PLoS One* 8.4, e62629. DOI: 10.1371/journal.pone.0062629.

Zhang, Z. and A. Furman (2021). "Soil redox dynamics under dynamic hydrologic regimes - A review." In: *Science of The Total Environment* 763, p. 143026. DOI: 10.1016/j.scitotenv.2020.143026.

Zhou, J., X. Su, and G. Cui (2018). "An adaptive Kriging surrogate method for efficient joint estimation of hydraulic and biochemical parameters in reactive transport modeling." In: *Journal of Contaminant Hydrology* 216, pp. 50–57. DOI: 10.1016/j.jconhyd.2018.08.005.

Zhu, Y.-G., X.-M. Xue, A. Kappler, B. P. Rosen, and A. A. Meharg (2017). "Linking genes to microbial biogeochemical cycling: lessons from arsenic." In: *Environmental Science & Technology* 51.13, pp. 7326–7339. DOI: 10.1021/acs.est.7b00689.

Zumft, W. G. (1997). "Cell biology and molecular basis of denitrification." In: *Microbiology and Molecular Biology Reviews* 61.4, pp. 533–616. DOI: 10.1128/mmbr.61.4.533-616.1997. PMID: 9409151.