UNIVERSIDADE FEDERAL DO PARANÁ

LETÍCIA BECHER YAMASHITA

A DATA-CONSTRAINED FRAMEWORK FOR MARINE

BIOGEOCHEMISTRY MODELING WITH APPLICATIONS TO THE

PARANAGUÁ ESTUARINE COMPLEX

CURITIBA

2026

LETÍCIA BECHER YAMASHITA

A DATA-CONSTRAINED FRAMEWORK FOR MARINE BIOGEOCHEMISTRY MODELING WITH APPLICATIONS TO THE PARANAGUÁ ESTUARINE COMPLEX

Tese, apresentada como requisito parcial à obtenção do grau de Doutora em Matemática, no Curso de Pós-Graduação em Matemática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Orientador: Prof. Dr. Francisco de Melo Viríssimo (LSE, Reino Unido e INCLINE-USP, Brasil)

Coorientador 1: Prof. Dr. Ademir Alves Ribeiro (UFPR)

Coorientador 2: Prof. Dr. Roberto Ribeiro Santos Júnior (UFPR)

CURITIBA

2026

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MATEMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **LETÍCIA BECHER YAMASHITA**, intitulada: **A DATA-CONSTRAINED FRAMEWORK FOR MARINE BIOGEOCHEMISTRY MODELING WITH APPLICATIONS TO THE PARANAGUÁ ESTUARINE COMPLEX**, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Agosto de 2025.

Assinatura Eletrônica
01/10/2025 09:14:21.0
FRANCISCO DE MELO VIRISSIMO
Presidente da Banca Examinadora

Assinatura Eletrônica
01/10/2025 08:25:10.0
MAURICIO ALMEIDA NOERNBERG
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
01/10/2025 00:59:30.0
ELÍAS ALFREDO GUDIÑO ROJAS
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
03/10/2025 03:45:45.0
EVELIN HERINGER MANOEL KRULIKOVSKI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
02/10/2025 17:05:21.0
FILIPE PEREIRA DOS SANTOS
Avaliador Externo (UNIVERSITY OF CALIFORNIA SAN DIEGO)

AGRADECIMENTOS

Agradeço aos meus orientadores, aos membros da banca e ao coordenador do PPGM, professor Marcelo Muniz, pelo apoio, pela disponibilidade e pela confiança no meu potencial ao longo do doutorado.

Em especial, agradeço à Marines pela parceria fundamental na concepção da forma inicial deste projeto, e ao Francisco, pela supervisão atenta e constante durante todo o processo de desenvolvimento do trabalho.

Agradeço também ao Rick, meu marido, cuja presença, cuidado e incentivo tornaram este percurso possível.

Agradeço à doutora Sophy Oliver, à professora Iris Kriest e à doutoranda Rafaela Farias do Nascimento pelo suporte na primeira iteração deste projeto.

Agradeço ao Carlos Carvalho e à equipe do Laboratório Central de Processamento de Alto Desempenho (LCPAD) da UFPR pelo suporte com o uso do cluster.

RESUMO

Nas últimas duas décadas, modelos matemáticos e computacionais tornaram-se uma das principais ferramentas usadas no estudo da dinâmica de sistemas marinhos, permitindo que pesquisadores testem hipóteses e investiguem situações de interesse no presente, passado e futuro - antes impossíveis sem campanhas ocenográficas. Estas incluem o estudo de mudanças climáticas, previsão da qualidade da água, fluxos de nutrientes, risco de proliferação de algas nocivas e gestão de portos e estuários, dentre outros. Em geral, o desenvolvimento de modelos que representem de forma confiável o ambiente de interesse é uma tarefa complicada e de alto custo computacional, dificultando a aplicação destes modelos em ambientes específicos, como plumas e complexos estuarinos - tais como o de Paranaguá, na costa paranaense. Além disso, modelos biogeoquímicos possuem vários parâmetros a serem calibrados para região de interesse, em geral ajustados empiricamente, com base em heurísticas e inferências à partir de dados disponíveis, devido ao alto custo computacional envolvido nas simulações. Neste trabalho, propomos avançar o estado da arte do desenvolvimento e calibração de modelos biogeoquímicos, dentro do contexto brasileiro, com três contribuições. Primeiro, desenvolvemos um modelo conceitual para a dinâmica nutriente-fitoplâncton do Complexo Estuarino do Paranaguá (CEP) - um dos primeiros modelos desenvolvidos no Brasil para a região. O modelo é simples e de baixo custo computacional, acessível a qualquer laptop doméstico. Em seguida, o trabalho propõe uma abordagem de calibração sistemática para modelos biogeoquímicos marinhos utilizando conjuntos de dados traçadores das regiões de estudo e otimização. Finalmente, apresentamos uma aplicação prática desta abordagem ao modelo do CEP, onde a calibração de modelo é realizada utilizando dados in-situ de traçadores. O resultado mostra que o modelo, apesar de simples, é capaz de reproduzir as observações desde que propriamente calibrado, demonstrando o poder do uso de técnicas de otimização - mesmo em modelos concentuais. Como possibilidades futuras, observamos que tanto o modelo quanto a abordagem são generalizáveis, permitindo calibração multiparâmetro, variação sazonal de parâmetros e modelos bioquímicos acoplados a modelos hidrodinâmicos de maior fidelidade, com potencial de automação via aprendizado de máquina.

Palavras-chave: modelagem matemática e computacional; modelagem biogeoquímica marinha; dinâmica estuarina; otimização aplicada; otimização sem derivadas para mínimos quadrados; calibração de parâmetros; Complexo Estuarino de Paranaguá.

ABSTRACT

Over the last two decades, mathematical and computational models have become one of the primary tools used in the study of marine biogeochemical systems, enabling researchers to test hypotheses and to investigate scenarios of interest in the present, past, and future – previously impossible without dedicated oceanographic expeditions. These include the study of climate change, water quality prediction, nutrient flows, harmful algal bloom risk, and port and estuarine management, among others. In general, developing models that reliably represent the environment of interest is a complex and computationally expensive task, which hinders the application of these models in specific environments, such as plumes and estuarine complexes – such as the Paranaguá Estuarine Complex (PEC) on the Paraná State coast, in Southern Brazil. Furthermore, biogeochemical models have several parameters that need to be calibrated for the region of interest, which are generally tuned empirically, based on heuristics and inferences from available data, due to the high computational cost involved in the simulations. In this work, we propose advancing the state of the art in the development and calibration of marine biogeochemical models, within the Brazilian context, with three main contributions. First, we developed a conceptual model for the nutrient-phytoplankton dynamics of the PEC – one of the first models developed in Brazil for the region. The model is simple and computationally inexpensive, accessible to any home laptop. Next, the work proposes a systematic calibration approach for marine biogeochemical models using tracer datasets from the study regions and optimization. Finally, we present a practical application of this approach to the PEC model, where model calibration is performed using in-situ tracer data. The results show that the model, despite its simplicity, is capable of reproducing the observations provided it is properly calibrated, demonstrating the power of using optimization techniques – even in a conceptual model setting. As future opportunities, we note that both the model and the approach are generalizable, enabling multi-parameter calibration, seasonal variation of parameters, and biochemical models coupled with higher-fidelity hydrodynamic models, with further automation potential via machine learning.

Keywords: mathematical and computational modeling; marine biogeochemical modeling; estuarine dynamics; applied optimization; derivative-free optimization for least squares; parameter calibration in models; Paranaguá Estuarine Complex.

LIST OF FIGURES

LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

C_{N} Mean nitrate concentration in a given water volume (mmol m^{-3}).

C_{PHY} Mean phytoplankton concentration in a given water volume (mmol m^{-3}).

V_{max} Maximum specific growth rate of phytoplankton (d^{-1}).

λ Linear mortality rate of phytoplankton (d^{-1}).

K Monod half-saturation constant for nitrogen (mmol m^{-3}).

r Nitrogen remineralization rate (d^{-1}).

I Solar irradiance (kWh m^{-3} d^{-1}).

T Mean water temperature (^{o}C).

S Mean water salinity (dimensionless).

$\eta(C_{\mathrm{N}})$ Limiting factor on phytoplankton growth due to nitrate availability, dimensionless, $0 \le \eta(C_{\mathrm{N}}) \le 1$.

$\gamma(I)$ Limiting factor on phytoplankton growth due to solar irradiance, dimensionless, $0 \le \gamma(I) \le 1$.

$\alpha(T)$ Limiting factor on phytoplankton growth due to water temperature, dimensionless, $0 \le \alpha(T) \le 1$.

$\beta(S)$ Limiting factor on phytoplankton growth due to water salinity, dimensionless, $0 \le \beta(S) \le 1$.

$g(C_{\mathrm{N}}, T, S, I)$ Total limiting factor on phytoplankton growth, dimensionless, $0 \le g(C_{\mathrm{N}}, T, S, I) \le 1$.

L_x Horizontal extent of the estuarine box in the x-direction (m).

L_y Horizontal extent of the estuarine box in the y-direction (m).

H Mean total water depth in the model domain (m).

H_{up} Mean depth of the upper (surface) layer in the two-box representation (m).

H_{low} Mean depth of the lower (bottom) layer in the two-box representation (m).

$S_{\mathrm{up}}(t)$ Salinity in the upper box as a function of time t (dimensionless).

$S_{\mathrm{low}}(t)$ Salinity in the lower box as a function of time t (dimensionless).

$Q_{\mathrm{ocean}}(t)$ Water volume flux exchanged with the adjacent ocean as a function of time t (m^3 s^{-1}).

$Q_{\text{river}}(t)$	Freshwater inflow from rivers as a function of time t (m³ s⁻¹).
$Q_{\text{ebm}}(t)$	Net volume flux associated with the estuarine box model (EBM) forcing as a function of time t (m³ s⁻¹).
Vol_{box}	Water volume of a model box (m³).
$T(t)$	Upper-box temperature, in ⁰C, as a function of time t.
$C_{\text{N}}^{\text{river}}(t)$	Average nitrogen concentration in the river inflows to the PEC, as a function of time t (mmol m⁻³).
ω	A point in the parameter space $\mathbb{R}^n$.
n	Dimension of the parameter vector ω.
$p+1$	Number of candidate parameters in the candidate pool.
ω_j^k	Candidate parameter at iteration k of DFO-LS, $j = 0, 1, \ldots, p$.
$\boldsymbol{\Omega}^k$	Candidate pool in the parameter space $\mathbb{R}^n$ at iteration k, $\boldsymbol{\Omega}^k = \{\omega_0^k, \omega_1^k, \ldots, \omega_p^k\}$.
$f_{\text{Misfit}}(\omega)$	Misfit function minimized by DFO-LS, $f_{\text{Misfit}} : \mathbb{R}^n \to \mathbb{R}_+$.
$r_{\text{Misfit}}(\omega)$	Misfit residual vector for $f_{\text{Misfit}}(\omega)$, $r_{\text{Misfit}} : \mathbb{R}^n \to \mathbb{R}^m$.
m	Dimension of the misfit residual vector $r_{\text{Misfit}}(\omega)$.
PEC	Paranaguá Estuarine Complex.
SMS	Source-minus-sink term describing net biogeochemical sources and sinks.
NP	Nutrient-phytoplankton interactions.
TMM	Transport Matrix Method.
KEBM	Knudsen Estuarine Box Model.
ZMT	Zone of Maximum Turbidity.
ODE	Ordinary Differential Equation.
PDE	Partial Differential Equation.
DFO	Derivative-Free Optimization.
DFO-LS	Derivative-Free Optimization for Least Squares.
CMA-ES	Covariance Matrix Adaptation Evolution Strategy.
MOPS	Model of Oceanic Pelagic Stoichiometry.

CONTENTS

1 Introduction

This thesis concerns the development of conceptual mathematical models in marine biogeochemistry, and the practical use and implementation of state-of-the-art derivative-free optimization techniques to calibrate the parameters in the model. Specifically, the goal of the thesis is to develop a bespoke, data-constrained conceptual model for the nutrient-phytoplankton dynamics of the Paranaguá Estuarine Complex (PEC), which involves deriving governing equations for the biogeochemistry, and to use in-situ observations from phytoplankton and nitrate (as the nutrient) to calibrate the model using a systematic optimization framework.

Marine biogeochemistry is an interdisciplinary branch of science that investigates how biological, chemical, and geophysical processes regulate the transformation and transport of biological and chemical components in the ocean, such as planktonic organisms and carbon. Marine biogeochemical models, in turn, translate these interactions into mathematical equations, which can then be solved (usually with the help of a computer), enabling the simulation of scenarios, hypothesis testing, and informed decision-making [17, 49, 57, 67]. These models range from simple "box" models, such as the ones presented in this thesis, to highly complex four-dimensional models representing in detail the marine biogeochemical cycles at global scales [14, 20, 25, 26, 28, 30, 31, 32, 33, 34, 35, 45, 47, 70, 77].

The applications of marine biogeochemical models are also diverse and include estimating water quality, carbon and nutrient cycles, risk of harmful algal blooms, and port management [5, 13, 14, 17, 49, 57, 73]. In particular, these models are also widely used to assess the potential effects of anthropogenic climate change and pollution on the ocean ecosystem, informing public policy and management [28, 30, 32, 33, 45, 47, 77]. The latter is particularly relevant to PEC, where the anthropogenic discharge of nutrients into PEC from the surrounding urban and port environments leads to the bloom of toxic algae, which impacts the local ecosystem and compromises the quality of water [4, 40, 43, 44, 50, 61, 76].

Indeed, the Paranaguá Estuarine Complex has long been recognized as a system highly sensitive to nutrient enrichment, mainly because rivers, urban effluents, port activities, and hydrodynamic modifications intensify the input and retention of nitrogen- and phosphorus-rich waters [40, 41, 43, 44, 50]. Several studies in the region have documented elevated nutrient concentrations, alterations in phytoplankton structure, and recurrent episodes of harmful algal blooms [4, 61, 62]. However, most of this knowledge is based on in situ measurements collected at specific times and locations. While these observations are essential, they often provide a fragmented view of a dynamic, highly variable system strongly modulated by tides, circulation, and

estuarine mixing. In this context, marine biogeochemical models allow researchers to test hypotheses and to explore the consequences of different management interventions before they occur in the real system [17, 49, 57, 73]. This thesis is motivated precisely by the need to fill the gap between observational studies and predictive understanding through the development of a simple conceptual model, designed to represent the local biogeochemical dynamics of the PEC in a realistic, interpretable, and computationally efficient manner [27, 28, 31]. Such a framework provides a basis for future scenario analyses and can ultimately support evidence-based environmental policies and estuarine management.

Through this approach, this thesis emphasizes data-fitting and optimized modeling, seeking reproducible procedures for calibrating parameters and quantifying uncertainties under settings relevant to PEC. Before proceeding, we will briefly expand on the two main elements of this thesis: biogeochemistry and modeling.

Biogeochemistry for mathematicians

Before exploring the concept of a biogeochemical cycle, we define tracers and compartments, using examples that are likely familiar to most readers. Some of the most well-known chemical element cycles are the water cycle, the phosphate cycle, and the carbon cycle. In any biogeochemical cycle, we refer to the elements being tracked, in their various forms and locations, as *tracers*. Thus, in the examples mentioned, the tracers could be identified as H_2O, PO_4, and CO_2, respectively. A biogeochemical cycle typically considers a set of tracers, some of which have biological origins. For example, phytoplankton populations, which are algae and therefore realize photosynthesis, can be tracked by the amount of Chlorophyll-A present on the surface of a marine region. In this case, we could define Chlorophyll-A as the tracer, although it would be equivalent to considering the phytoplankton itself as a tracer.

On the other hand, we will now discuss *compartments*. Returning to the example of the water cycle, as it passes through different states of matter, we can affirm that water also travels through different compartments, which is the definition of *cycling*. In its liquid state, it is present in bodies of water, such as rivers and the ocean. In its gaseous state, it is present in the form of clouds in the Earth's atmosphere. We could then consider the ocean as one large compartment, and the atmosphere as another large compartment. Thus, a possible model for the global water cycle would involve the transport of this tracer between these two compartments.

The complexity of a biogeochemical cycle increases even further when multiple tracers and compartments are taken into account. Not only are tracers transported between compartments (geophysical part), but they also interact with one another (biochemical part). Due to their complex nature, it is common for mathematical models representing marine biogeochemical cycles to be obtained by coupling a biochemical conceptual model, described in diagrammatic form, with a physical transport model that represents advection and diffusion processes, described as a Dirichlet boundary problem. The resulting mathematical model is a set of ODEs (or PDEs, depending on the approach), which rely on a set of variables and parameters. More

details on the process of mathematical modeling of a marine-biogeochemical cycle are presented in *Chapter 2*. Additionally, the values to be set in each category are not static and can vary depending on the purpose for which the model is being used. This will be explored in the optimization framework steps of validation and calibration, in *Chapter 5*.

Mathematical modeling for biogeochemists

In our context, modeling refers to the mathematical representation of the interactions between elements of a biogeochemical cycle. It's virtually impossible to represent everything that occurs in an environment, just as it's unrealistic to expect results without any precision errors. Thus, creating a mathematical model involves identifying the most relevant elements of a biogeochemical cycle and quantifying the primary interactions between them.

The biochemical part of quantifying interactions between elements is generally performed based on empirical estimates obtained through laboratory experiments, while the geophysical part usually derives from a computational mathematical model that uses input data, such as ocean current flows in a few locations, to produce accurate estimates about the current flow at each point in an entire region. Even so, it is also common to use theoretical hypotheses in modeling as a way of obtaining more information from less data.

The strategy of assuming hypotheses is particularly useful when data is scarce. In some cases, data scarcity occurs due to difficult access to sampling regions. However, it is also common for the motivation behind sampling data to be using it as input in a mathematical model. In this case, the development of a mathematical model can be carried out prior to data sampling, without this factor being an impediment. In this context, in *Chapter 3*, we present an application on the biogeochemical mathematical modeling of a marine region.

Once a biogeochemical cycle has been translated into a mathematical model, this model is rewritten in the form of a computer program and processed computationally. The way in which model and data are processed by the computer is not unique and can indeed be refined. *Chapter 6* presents an application on how to apply the framework presented in *Chapter 5* for refining a marine biogeochemical model using optimization and the data available.

In the following, we provide a concise overview of the thesis's context. Some technical terms may be unfamiliar to the reader, depending on their background in research. Nevertheless, more details about each of the presented concepts will be explored in the subsequent chapters.

Motivation, Objectives, and Contributions

Biogeochemical modeling spans from conceptual biochemical models, aimed at addressing mechanistic questions and hypothesis testing, to high-fidelity three-dimensional configurations in which conceptual biochemistry is coupled to complex hydrodynamics. Typical marine biogeochemical models include dozens of parameters. Additionally, the high dimensionality of the parameter space induces correlations and non-identifiability, as different parameter combinations can fit observations equally well (equifinality), as documented across simple and 3-D calibrated models [28, 31, 32, 33]. Many parameters are not directly observable at the rates and scales required for operational applications. Time series may be short, sparse, and heterogeneous; laboratory estimates do not always translate to natural settings.

Particularly, in coastal and estuarine environments, strong spatial–temporal variability and anthropogenic forcings necessitate careful scoping and process selection to ensure the model remains parsimonious and computationally tractable, producing reliable and reproducible estimates [41, 44].

Due to the level of complexity in determining parameter values, it has historically been common for parameters to be inferred theoretically, based on information obtained from previous studies. However, recently an approach has been presented that estimates a set of parameters indirectly through data fitting [31].

While manual trial-and-error tuning remains common, it is hard to reproduce, is dependent on the modeler's experience, and tends to explore only a small fraction of the parameter space - the latter being a crucial issue in the context of climate modeling [46]. On the other hand, systematic calibration specifies an objective function (model–data misfit), defines search rules (global and local), and documents choices (weights, normalizations, constraints), enabling comparability across scenarios and methodological review [7, 31, 55].

When dealing with complex models, a computational approach for calibrating parameters can be costly and may exhibit numerical noise, as well as unreliable or unavailable gradients. Beyond computational expense, multiple local minima and process–resolution interactions affect fitting. These realities motivate cost-reduction strategies (for example, transport-matrix methods and accelerated spin-up) and derivative-free optimization tailored to least-squares structure [7, 25, 26, 34, 55]. In this context, studying parameter calibration in a simple model provides a controlled laboratory environment for addressing targeted questions, generating mechanistic insights, and facilitating rapid experimentation at a low cost. They enable broad parameter-space exploration, structural testing, and intuition-building that can inform more complex set-ups [36, 73]. In estuaries and ports – where interest often centers on specific processes and seasonal/anthropogenic responses – such models are particularly useful [41, 44, 79].

At lower computational cost, we combine heuristic global search (to locate promising initial points) with local derivative-free refinement, such as DFO-LS, a Gauss–Newton-type method for nonlinear least squares minimization [7, 8, 55]. This arrangement facilitates studies of identifiability, noise robustness, and

objective-function design, and supports post-fit sensitivity analysis [31]. In this sense, the guiding questions are: (i) Which combination of objective function, search strategy, and constraints yields reproducible and interpretable fits? (ii) Which parameters (or parameter groups) are identifiable with available data, and under what conditions? (iii) How do observational uncertainties and normalization/weighting choices affect fits and process inference? This focus aligns with prior PEC studies on circulation and biogeochemistry [41, 44, 50, 68].

The central purpose of this thesis is to connect enthusiasts from different research areas, such as mathematics and oceanography, by providing a step-by-step guide on how to develop a systematic, data-informed parameter-calibration methodology and apply it to build a conceptual, data-constrained model of marine biogeochemical dynamics.

The scope centers on (a) conceptual models applied to PEC; (b) multi-parameter calibration using tracer data; (c) derivative-free strategies suitable for potentially noisy objectives; and (d) cross-validation and sensitivity protocols.

The outcome is an interdisciplinary and practice-oriented study that bridges different scientific languages, making the content approachable to readers from multiple fields and pointing to several possible extensions of this work.

Structure of this Thesis

This thesis is organized to move from modeling basics to optimization methods and, finally, to the application of these principles to develop data-constrained model for the PEC.

Chapter 2 reviews marine biogeochemical modeling at a conceptual level with illustrative experiments.

Chapter 3 develops a conceptual nutrient-phytoplankton model tailored to the PEC. It defines the domain, derives biogeochemical equations and simple circulation components, couples them, and implements the model numerically; results and limitations are discussed.

Chapter 4 introduces the optimization background used throughout the thesis, formulates parameter calibration as a (weighted) least-squares problem from tracer observations, and explores the theory and computational usage of a derivative-free strategy – DFO-LS – motivating its use when gradients are unreliable or simulations are noisy/expensive.

Chapter 5 presents a systematic calibration workflow for conceptual models. It details the construction and minimization of a quadratic regression model within a trust-region framework, and demonstrates the pipeline (problem set-up, heuristic search, and local refinement) through calibration tests and case studies, with results and diagnostics.

Chapter 6 applies the calibration framework to PEC in a data-constrained setting, including a twin benchmark, a fit to one-year nitrate and phytoplankton observations, and a scenario simulation with increased riverine nitrate load.

Lastly, we summarize the main contributions of this thesis and outline potential directions for future research.

2 Modeling marine biogeochemical cycles

Biogeochemical mathematical models represent the processes that govern the cycling of biological and chemical elements in an environment. They are widely used to study environmental issues such as climate change and pollution [49, 57, 67], serving as a useful tool for gaining greater insights into the human impact on nature and, thus, facilitating the conscious planning of these interactions. Since the 1960s, the advent of computers has enabled the development of the first computational biogeochemical models. Given the limited knowledge and computing power of the time, these were relatively simple conceptual models that simulated only a few processes [57]. Over time, both scientific understanding and computational capacity increased, enabling the development of more complex biogeochemical models. In this chapter, we will review the foundations of biogeochemical modeling, exemplified by simple conceptual models. This basic understanding is fundamental and generalizable for the study and development of more complex models, which may be of interest to the reader in the future [28, 31].

In the following section, we will introduce some key elements mentioned in the introduction that are essential for creating a conceptual biogeochemical model.

2.1   Tracers and parameters

For context, let us imagine a situation where a swimming pool presents a leak. A technician is called, and he pours an amount of colored dye into the pool. Tracking the path taken by the dye allows the technician to estimate the location of the leak. Now, imagine someone collects a bucket of water from this pool after the technician has added the dye. Analyzing the color of the collected water allows us to infer the average concentration of the dye present in the water at the time of collection, at least in the part of the pool where the water was collected.

In the context of marine biogeochemical modeling, a tracer is a chemical or biological component that exhibits behavior similar to that of a dye added to a swimming pool. The tracer dynamics is modeled using a geophysical model, which typically considers advection and diffusion, in addition to the environmental geography. Furthermore, the concentration of a tracer in a given region can be quantified through various sampling or identification methods (see Box 2.1). In addition to the behavior described, tracers also interact with each other. Mathematically, a computational marine-biogeochemical model predicts or estimates tracer

concentrations in given marine regions over time.

Box 2.1: Gathering tracer data:

How are phytoplankton concentrations estimated in marine regions?

Chlorophyll-A is the most abundant photosynthetic pigment in phytoplankton and is useful as an indirect indicator of these microorganisms' biomass [58]. Different sampling and analysis techniques are applied to estimate its concentration in ocean waters:

- In-situ collections: A CTD (Conductivity, Temperature, and Depth) system is an oceanographic instrument used to measure the electrical conductivity, temperature, and depth of seawater throughout the water column. A Niskin bottle is a PVC or acrylic cylinder with end caps that remain open during descent and are closed remotely (by a mechanical messenger or electrical system) at the desired depth, ensuring that other water layers do not contaminate a water sample. Traditionally, researchers use oceanographic cylinders (such as Niskin bottles) attached to CTD systems to collect water samples at different depths. In the laboratory, Chlorophyll-A is extracted with solvents (usually acetone or methanol) and quantified by fluorometry or spectrophotometry (which are precise methods, but require time and careful handling [72]).

- Fluorometers and optical sensors: Submersible equipment allows real-time measurements of fluorescence associated with Chlorophyll-A. These sensors are widely used in vertical profiles and in continuous monitoring systems on board ships, buoys, or gliders, providing high-resolution spatial and temporal data [72].

- Remote sensing: Satellites such as MODIS and Sentinel-3 estimate surface Chlorophyll-A concentrations based on light reflectance in the ocean, allowing for mapping of large ocean areas. Although less accurate locally, this is a crucial tool for collecting data in areas of difficult access [22].

Each of these methods has its own advantages and limitations: while laboratory analyses are more accurate, sensors and satellites expand the range of observations. Together, they determine the baseline data on phytoplankton concentrations in different marine regions.

On the other hand, parameters are model features that usually cannot be measured directly. They are inferred values based on models and theoretical estimates that may be constant, depend on the tracer's features, or depend on interactions between tracers. For example, the phytoplankton growth rate is a parameter that varies among species and environmental conditions; however, in most models, it is represented by an average laboratory-derived estimate rather than a directly measurable field value.

A conceptual model is a diagram that represents the main components of the biogeochemical system and how they interact with each other [14, 17, 57, 67]. Mathematically, it can be translated into a set of differential equations representing changes in tracers' concentrations depending on time and interactions between them [17, 57, 67].

The complexity of a conceptual model implies a necessary level of complexity in the mathematical model obtained from it. In this sense, conceptual models that consider a simplified domain, with few processes and parameters, often can be translated to simple mathematical models, with low computational costs [27, 28, 30, 32, 33]. To illustrate a simple model, we present the following example.

Example 2.1 (Conceptual model for nitrate-phytoplankton dynamics)**.** *Consider a well-mixed, homogeneous oceanic region represented by a single box for the entire water column. The model tracks two tracers over time: nitrate and phytoplankton. Phytoplankton grow by assimilating nitrate during the photosynthesis process. When phytoplankton die, a fraction of their organic nitrogen is remineralized to dissolved inorganic nitrate within the box. In this example, we assume that* 70% *of dead phytoplankton is remineralized locally (parameter* $r = 0.7$*), while the remaining* 30% *is lost from the modeled region (for example, export/sinking). A conceptual model for this situation is presented in Figure 2.1.*

Figure 2.1 – Diagram representing a conceptual biochemical NP-model within an aquatic environment.
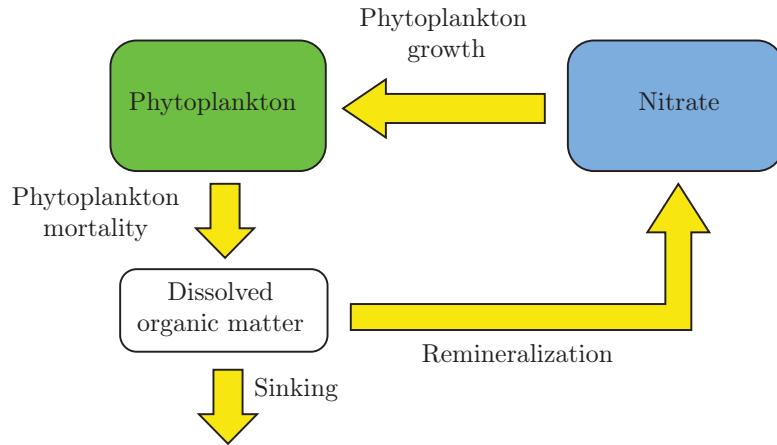
The mean concentrations of the tracers in the environment are represented by colored boxes, namely nitrate and phytoplankton. The arrows represent interactions between model components that alter the concentration of each tracer. Arrows pointing toward a box indicate processes that increase the concentration of that component, while arrows leaving a box indicate processes that decrease it. Dissolved organic matter is not represented as a tracer but as a conceptual component that explains a process modeled implicitly.
Source: the author (2026).

In Example 2.1, the conceptual model presented is treated as a closed aquarium containing a set of interacting tracers, so the physical processes are dealt implicitly: all elements within the modeled domain are assumed to interact instantaneously. Due to this simplification, the geophysical framework can be omitted from the model representation, which can then be summarized as a diagram of biochemical interactions between tracers. The system represented in Example 2.1 corresponds to a simple NP (Nutrient-Phytoplankton) model (see Box 2.2 for a broader discussion of NP and NPZ models).

Box 2.2: NP and NPZ models:
Simplified Representations of Marine Ecosystem Dynamics

The NP (Nutrient-Phytoplankton) and NPZ (Nutrient-Phytoplankton-Zooplankton) models are among the most widely used simplified representations of marine biogeochemical interactions. Nutrients are chemical substances that living organisms need to grow and survive. In the ocean, nutrients feed phytoplankton - microscopic algae that float near the surface and use sunlight to produce energy through photosynthesis. Although tiny, phytoplankton are the base of the marine food web and play a crucial role in producing oxygen and absorbing CO_2 from the atmosphere.

An NP model is a system comprising two tracers: a main nutrient (N), which, for example, can be set as nitrate or phosphorus, depending on data availability, and phytoplankton (P), a simplified representation of an entire population of algae containing different species. Phytoplankton consume the nutrients in a process known as primary production, which depends on light, temperature, and the availability of nutrients. The loss of phytoplankton - through respiration, mortality, or aggregation - returns nutrients to the inorganic pool. This minimal structure captures the basic feedback between nutrient limitation and phytoplankton growth.

On the other hand, an NPZ model additionally considers zooplankton (Z) as a tracer, explicitly representing the grazing and predation of phytoplankton. Phytoplankton serve as food for zooplankton, whose metabolic losses (excretion, mortality, and sloppy feeding) regenerate nutrients and particulate organic matter. By considering zooplankton as a tracer, nonlinear trophic interactions, such as predator-prey cycles, are introduced.

These models strike a balance between simplicity and ecological realism. While an NP model helps explore nutrient limitation and productivity under equilibrium conditions, an NPZ model provides a more dynamic and detailed view of ecosystem regulation and energy transfer within the planktonic food web [67].

In the following section, we explore details about the mathematical representation obtained from the diagrammatic form of a conceptual model.

2.2 Source Minus Sink equations

In biogeochemical modeling, a *source* refers to any process that increases the concentration of a tracer, whereas a *sink* refers to any process that decreases it. For example, in the diagram depicted in Figure 2.1, each arrow entering the box representing nitrate tracer concentrations represents a source of this tracer, while each arrow exiting this box represents a sink. The source-minus-sink (SMS) equation of a tracer quantifies the aftermath concentrations of the tracer as the source and sink processes occur. Below, we discuss how to obtain SMS equations for the conceptual model presented in diagrammatic form on Example 2.1.

Example 2.2. *The conceptual model represented in Figure 2.1 can be mathematically translated by the following ODE equations:*

$$\frac{dC_{\mathrm{N}}}{dt} = -V_{max} \cdot \frac{C_{\mathrm{N}}}{K + C_{\mathrm{N}}} \, C_{\mathrm{PHY}} + r \cdot \lambda \cdot C_{\mathrm{PHY}} \tag{2.1}$$

$$\frac{dC_{\mathrm{PHY}}}{dt} = V_{max} \cdot \frac{C_{\mathrm{N}}}{K + C_{\mathrm{N}}} \, C_{\mathrm{PHY}} - \lambda \cdot C_{\mathrm{PHY}} \tag{2.2}$$

In the system of equations (2.1)-(2.2), the nitrate concentration decreases due to phytoplankton uptake, $V_{\max} \frac{C_{\mathrm{N}}}{K+C_{\mathrm{N}}} \, C_{\mathrm{PHY}}$, and increases through remineralization of a fraction r of phytoplankton mortality, that is, $r \, \lambda \, C_{\mathrm{PHY}}$. Phytoplankton increases via Monod-limited growth, $V_{\max} \dfrac{C_{\mathrm{N}}}{K + C_{\mathrm{N}}} \, C_{\mathrm{PHY}}$, and decreases due to mortality, $\lambda \, C_{\mathrm{PHY}}$. In these equations, we considered the following notations:

Tracers:

- C_{N}*: nitrate concentration (mmol m^{-3});*

- C_{PHY}*: phytoplankton nitrogen concentration (mmol m^{-3});*

Parameters:

- V_{max}*: maximum specific uptake/growth rate (day^{-1});*

- *K: half-saturation constant for nitrate uptake (mmol m^{-3});*

- *λ: phytoplankton mortality rate (day^{-1});*

- *r: remineralization fraction returning mortality to nitrate (dimensionless, $0 \leq r \leq 1$).*

The right-hand sides of Eqs. (2.1)-(2.2) are written in the *source-minus-sink* (SMS) form: each tracer's net tendency equals the sum of processes that increase it (sources) minus the sum of processes that decrease it (sinks).

For the nitrate concentration, C_{N}, we have

$$\frac{dC_{\mathrm{N}}}{dt} = \underbrace{r \, \lambda \, C_{\mathrm{PHY}}}_{\text{source: remineralization}} - \underbrace{V_{\max} \frac{C_{\mathrm{N}}}{K + C_{\mathrm{N}}} \, C_{\mathrm{PHY}}}_{\text{sink: uptake by phytoplankton}} \;\; .$$

Thus, nitrate increases when remineralization ($r \, \lambda \, C_{\mathrm{PHY}}$) exceeds biological uptake, and decreases otherwise.

For the phytoplankton concentration, C_{PHY}, we have

$$\frac{dC_{\mathrm{PHY}}}{dt} = \underbrace{V_{\max} \frac{C_{\mathrm{N}}}{K + C_{\mathrm{N}}} \, C_{\mathrm{PHY}}}_{\text{source: Monod-limited growth}} - \underbrace{\lambda \, C_{\mathrm{PHY}}}_{\text{sink: mortality}} \;\; .$$

Hence, phytoplankton biomass increases when the specific growth rate $V_{\max} \frac{C_{\mathrm{N}}}{K+C_{\mathrm{N}}}$ exceeds the mortality rate λ; if mortality dominates, biomass declines. In both equations, each term has units of mmol m^{-3} day^{-1}, so positive SMS implies tracer accumulation and negative SMS implies tracer loss. We can obtain the prediction for the concentrations of the tracers from the model by solving a Cauchy problem, as illustrated in the following example.

Example 2.3. *Recalling the mathematical model presented in Example 2.2, we now attribute initial values to the tracer concentrations and treat it as a Cauchy problem: $C_{\mathrm{N}}(0) = 80$ mmol m^{-3}, $C_{\mathrm{PHY}}(0) = 50$ mmol m^{-3}. If we additionally set the parameter values mentioned above, for example: $V_{max} = 1.4$ day^{-1}, $K = 0.1$ mmol m^{-3}; $r = 0.7, \lambda = 0.05$ day^{-1}, the ODE system (2.1)-(2.2) can then be solved numerically to simulate the evolution of the concentrations of nitrate and phytoplankton inside the box over time (Figure 2.2). For more details on the computational experiments, see the Appendix.*

2.3 Increasing model complexity: how to couple a biochemical conceptual model to a geophysical box model

The level of complexity required for a biogeochemical model depends on the research question and the available data. While simple models are useful for exploring basic concepts and relationships, making predictions for relatively short time periods, and as a teaching tool, more complex models are necessary for simulating complex biogeochemical cycles, making long-term predictions, and studying the impacts of environmental change on ecosystem processes. A basic increase in the complexity of the model presented in Example 2.1 involves adding physical processes while considering the domain divided into two boxes, rather than a single homogeneous box. The idea behind doing so is to achieve greater accuracy in describing the distinct features of different regions and how these features interact with one another.

A box model of a marine region is a representation obtained by either directly discretizing or conceptually depicting the region as a finite set of three-dimensional boxes, where each box is modeled as being uniformly

Figure 2.2 – Evolution of nitrate (upper plot) and phytoplankton (lower plot) concentrations over time for Example 2.3.
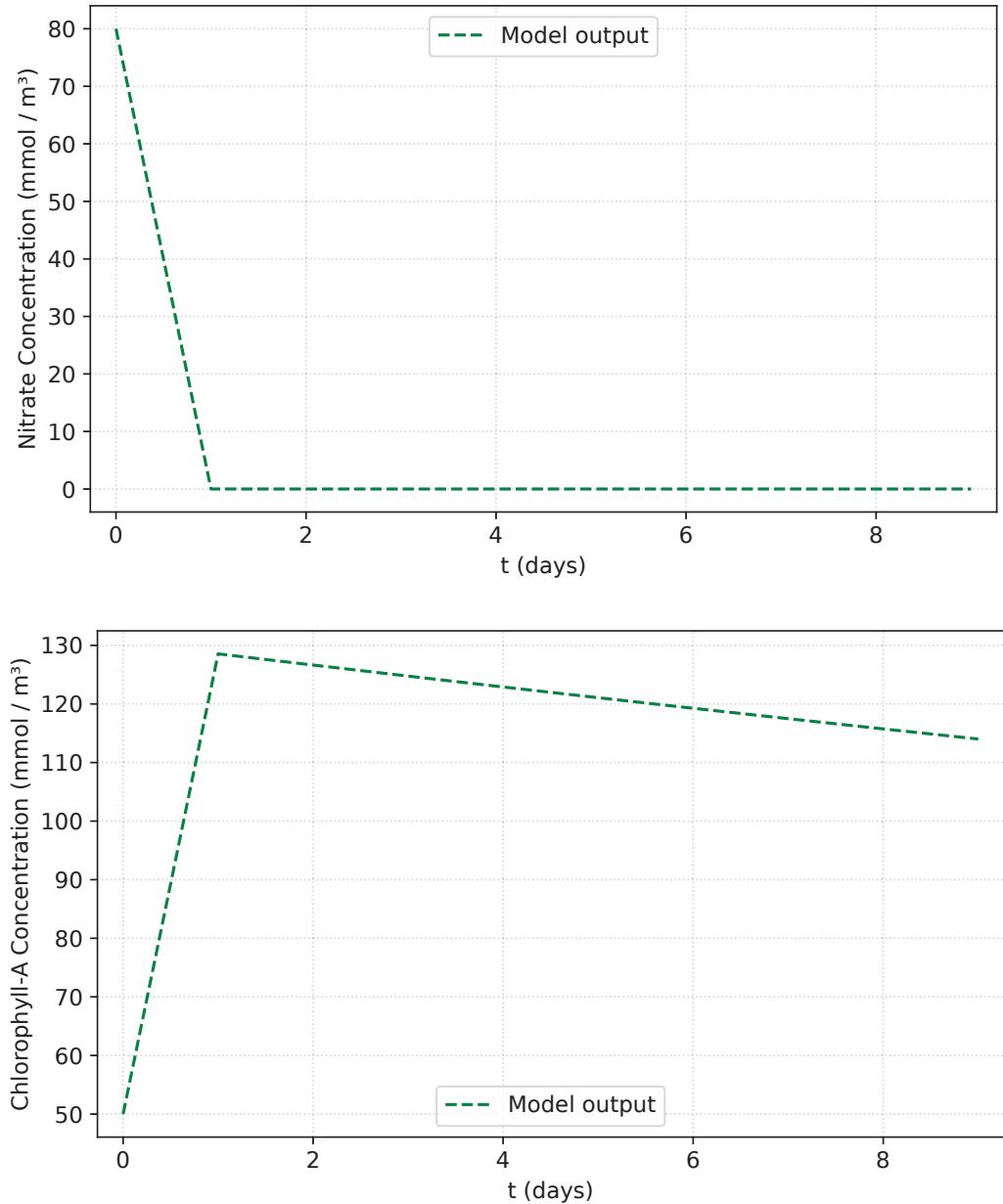
Source: the author (2026).

mixed, meaning that properties such as temperature, salinity, and tracer concentration are assumed to be uniform throughout the box volume. These simplified assumptions are considered for the practical implementation of the model, where the exchange between boxes represents the circulation and mixing processes that occur in the real environment. In a biogeochemical box model, the tracers within each box have their own sources and sinks, which can arise from both biochemical interactions and geophysical circulation.

Following the classical derivation of the tracer conservation equation for a fixed control volume [67, s. 19–23], we represent transport as the sum of advection and diffusion and group all in-box biogeochemical and external processes into a source–minus–sink (SMS) biochemical term. In a box-discretized aquatic model, advection corresponds to fluxes across box faces driven by the resolved velocity field, whereas diffusion represents down-gradient fluxes parameterized by an effective diffusivity. The SMS term accounts for local production and loss (for example, biological uptake, remineralization, external inputs), ensuring mass conservation at the box scale. For the discrete/matrix treatment of tracer transport in such models, see also [25].

In general, once we have an aquatic model where the space is discretized into boxes, advection and diffusion are the two fundamental processes that describe how substances move and spread in the fluid contained in each box. Advection is the process by which a substance is transported by the flow of a fluid, as occurs when the current of a river carries materials along its course (Figure 2.3). Advection is influenced by the speed and direction of fluid flow. Diffusion is the process by which a substance spreads from a region of high concentration to a region of low concentration, due to the random movement of particles (Figure 2.3). Even in the absence of macroscopic fluid flow, the individual particles of the substance are always in thermal motion, resulting in their gradual dispersion. Diffusion is influenced by the concentration of the substance, the diffusion coefficient (which determines how quickly dispersion occurs), and the concentration gradient. In summary, advection refers to the transport of the substance by fluid flow, while diffusion refers to the dispersion of the substance due to the random motion of particles. The tracer conservation equation for the volume of fluid contained in a fixed box of the discretized domain has the form:

$$\frac{\partial C}{\partial t} = \left.\frac{\partial C}{\partial t}\right|_{advection} + \left.\frac{\partial C}{\partial t}\right|_{diffusion} + SMS(C) \ , \tag{2.3}$$

where C is the concentration of such tracer, the first two terms of the sum are the variation of C due to advection and diffusion, and the last term is a source-minus-sink function of C within the box.

Example 2.4. *Consider the advection scheme represented in Figure 2.3. When the boxes' dimensions are large enough, it is common to ignore the diffusion contributions. In this illustration, the region modeled, that is, the middle box, will have both a source and a sink contribution due to advective flux.*

In a box model, while physical processes are incorporated within the sources and sinks of each tracer, the geographical features of the environment are captured through the boxes' configuration. For example, a complex coastal region can be represented by multiple interconnected boxes, each corresponding to different

Figure 2.3 – Illustrative representation of the advection and diffusion processes.

Advection
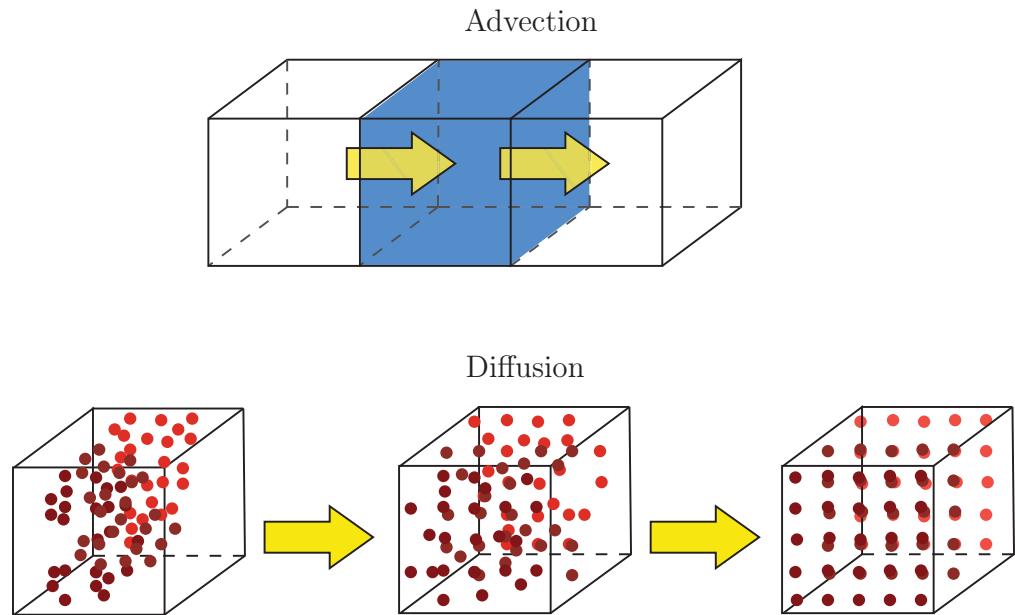
Diffusion

In the upper scheme, advective flux occurs in the positive x-direction; the blue region represents the modeled domain, while the arrows point in the flux direction. In the lower scheme, diffusion occurs over time; dots represent the molecules of a chemical tracer, while the arrows indicate their spatial arrangement when time passes.

Source: the author (2026).

zones such as the open ocean, continental shelf, and estuarine areas. Or, when more computational processing is available, even small regions can be directly discretized into boxes.

A concrete example of a global biogeochemical box model that combines source-minus-sink formulations with physical transport is given by the Model of Oceanic Pelagic Stoichiometry (MOPS), briefly described in Box 2.3.

Box 2.3: MOPS: a global laboratory for marine biogeochemistry

The Model of Oceanic Pelagic Stoichiometry (MOPS) [30, 55] is a marine biogeochemical model designed to represent the interactions between specific tracers in the global ocean. It represents the coupled carbon, nitrogen, phosphorus, and oxygen cycles within the ocean interior, combining physical transport with local biogeochemical transformations.

The dynamics of each tracer are governed by a system of ODEs representing the SMS equations of tracers, derived from both biochemical and physical processes. Each of these equations is written in

---

the form:

$$\frac{\partial C_i}{\partial t} = \underbrace{- \nabla \cdot (uC_i) + \nabla \cdot (\varepsilon_i \nabla C_i)}_{\text{Transport via TMM}} + \underbrace{\text{SMS}_i(C_1, \ldots, C_\ell, x, y, z, t)}_{\text{Local biogeochemical processes}},$$

where C_i is the concentration of tracer i (for example, nitrate, phosphate, oxygen, organic matter), u denotes the velocity field, ε_i the effective diffusivity, and SMS_i the nonlinear source-minus-sink function representing local biological and chemical interactions.

In its reference configuration (MOPS-1.0) [30], the model includes six adjustable parameters, along with several fixed ones, including maximum growth rates, half-saturation constants, and stoichiometric ratios. Once coupled to a general circulation model for the ocean dynamics, the nonlinearity and high dimensionality of the resulting discretized system make MOPS an ideal benchmark for automated calibration and optimization techniques, such as DFO-LS or CMA-ES, particularly under noisy or computationally expensive conditions.

MOPS is implemented using the use of the Transport Matrix Method (TMM) [25], which precomputes the global ocean's advection-diffusion operator into a sparse matrix. This approach decouples the biogeochemical model from the full dynamical ocean circulation model, reducing computational cost by orders of magnitude and enabling large-scale sensitivity and parameter studies.

Overall, MOPS represents a computational laboratory for global marine biogeochemistry, combining physical realism, stoichiometric parsimony, and numerical efficiency, which is a cornerstone framework for exploring parameter calibration and large-scale carbon-nutrient dynamics. As such, it has been used in numerous applications and studies [23, 29, 45, 47, 53].

2.4 Developing a model in practice

A realistic model can be simple or complex, depending on the goals of its representation. For example, although the values assumed for the initial concentrations and constants at Example 2.1 were arbitrary, for this model to represent a real environment, that is, be a realistic model, it would need to be initialized with measured or estimated initial values of C_N and C_{PHY}, and the model parameters $V\text{max}$, K, R, and λ would need to be estimated based on data or literature values. This task is known as the **calibration** or **tuning** of the parameters of the model, and will be explained in detail in Chapters 3 and 4.

Developing a biogeochemical model tailored for real-world applications is generally a difficult and long process, which is complicated by the fact that the inclusion of processes in such models is not an objective process [42]. In practice, such development involves several steps:

- Define the system: The first step is to define the system being modeled. This includes specifying the spatial and temporal scales, the components and processes being modeled, and the specific goals of the model.

- Gather data: The next step is to gather data on the system being modeled. This data can include observations of environmental variables, measurements of chemical concentrations, and data on the behavior of individual processes.

- Develop conceptual model: Based on the data and knowledge of the system, a conceptual model is developed that describes the main processes and interactions between the components of the system.

- Formulate mathematical model: The conceptual model is then translated into a mathematical representation. This involves developing equations that describe the behavior of each process and the interactions between the components. The equations may be based on empirical relationships or more complex, physically based models.

- Validate the model: The model is then validated by comparing its outputs with observed data. This step is crucial for ensuring that the model accurately represents the system being modeled.

- Calibrate the model: After the model has been validated, it may be necessary to calibrate it to fine-tune its parameters. This involves adjusting the model's parameters to ensure that its predictions align as closely as possible with the observed data. As final step, some models might require a sensitivity analysis to ensure the model is well-behaved in the spatiotemporal domain and parameter region of interest.

These steps are not always followed in a linear fashion and may be repeated several times as the model is refined and improved. The creation of a biogeochemical model is a complex and iterative process that requires a deep understanding of the system being modeled, as well as a strong background in mathematics and computer programming. Once the model has been validated and calibrated, it can be used to make predictions or to explore the system's behavior under various conditions. The model can also be used to evaluate the impacts of different management strategies or to gain insights into the underlying processes.

Next, we present the development of a conceptual model for the Paranaguá Estuary, based on the criteria outlined in this chapter.

3 A conceptual model for the Paranaguá Estuarine Complex

In this chapter, our primary goal is to formulate a conceptual model for the dynamics between nutrients and phytoplankton in the Paranaguá Estuarine Complex (PEC). By doing so, we also provide a detailed illustration of the mathematical modeling applied to marine biogeochemistry. The initial version of the model presented in this chapter was developed in collaboration with Dr. Marines M. Wilhelm during the postgraduate course "Tópicos de Matemática II: Aspectos teóricos e matemáticos do ciclo do carbono no oceano" taught by Prof. Dr. Francisco de Melo Viríssimo, in 2021. That early formulation served as the foundation for the expanded and fully revised model presented here.

3.1 The PEC environment and data availability

We begin this section by defining the model domain. The PEC is located on the central-north coast of Paraná (25°00'S - 25°35'S; 48°15'W - 48°15'W), in southern Brazil, and its ecological and economic relevance has been reinforced by recent geochemical assessments of sedimentary organic matter [76]. Having an area of approximately 612km^2, the PEC is subdivided into two axes, the main axis being called 'East-West', 56 km long, where the largest bulk port in Latin America is located, the Port of Paranaguá, and a 'North-South' axis, about 30 km long, where there is interference from artisanal fishing and agriculture (Figure 3.1) [37, 41, 43, 44].

The development of a conceptual NP model for PEC was motivated by the availability of in situ and literature datasets for the model tracer concentrations and related parameters. This data availability makes the NP formulation both practical and representative of the dominant nutrient-phytoplankton dynamics observed in the system.

The adoption of a two-box circulation model for the PEC is motivated by its classification as a partially mixed estuary, where both riverine inflow and tidal exchange contribute significantly to water renewal and vertical salinity gradients [41, 73]. In addition, salinity data are available [40], supporting the parametrization and validation of a simple stratified-exchange framework. In our conceptual model, the two-box representation

Figure 3.1 – Model domain: Paranaguá Estuarine Complex.
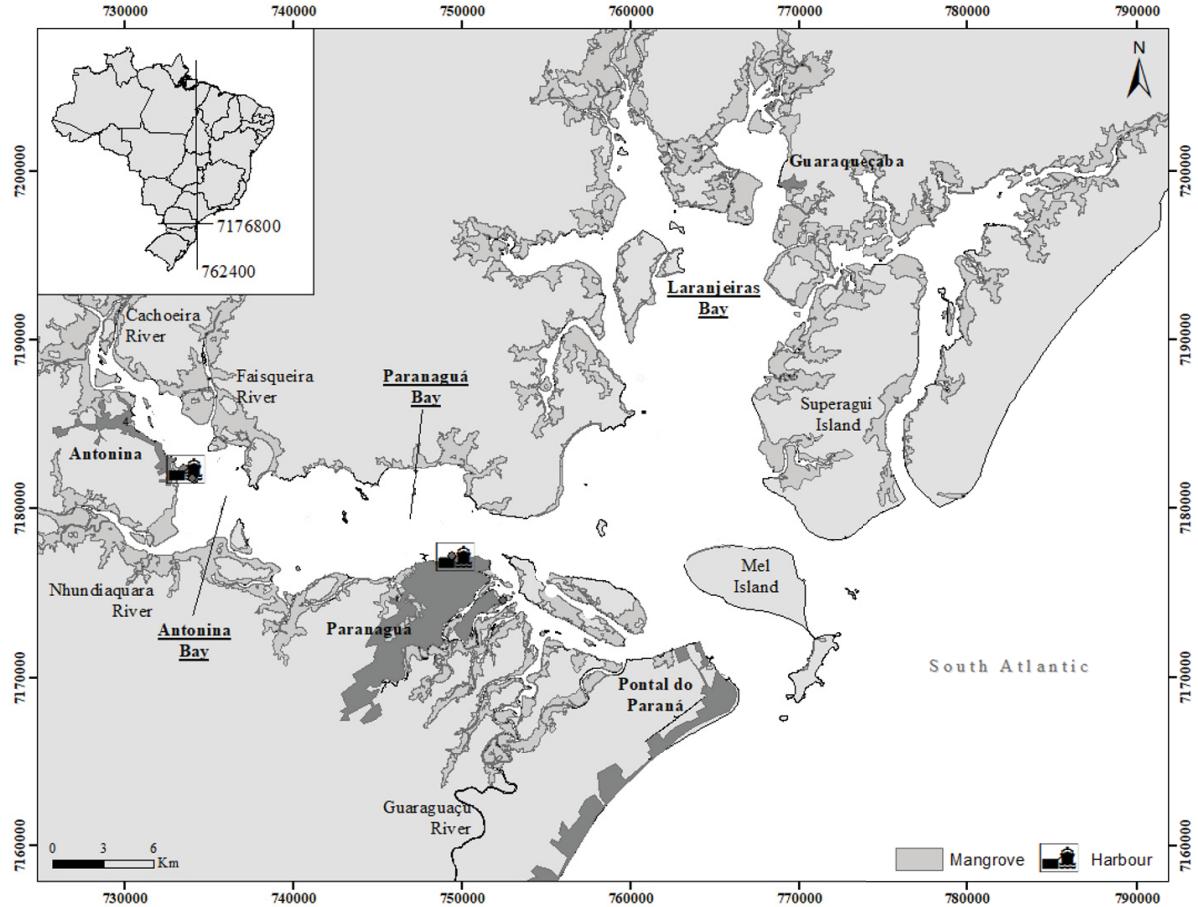
constitutes a deliberate simplification of PEC geophysical features, capturing the essential vertical structure of the estuary (surface and bottom layers) while remaining computationally inexpensive, providing a low-cost structure for testing the optimization framework presented in the following chapters by running experiments on a personal computer.

3.2   Model development

As in the previous chapter, we start by defining the biogeochemical SMS equations for this model. As part of that, we calibrate the model "manually" by finding parameter values based on the literature and available data. We then present a conceptual circulation box model representing the geophysical aspects of PEC. Finally, we couple both models to obtain the final set of biogeochemical SMS equations.

The data used for the development of the PEC biogeochemical model are:

1. Biochemical parameters, K, $V_{\max}$ and λ, here set as constants, although the last two are based on the PEC phytoplankton population data found in the literature [40, 62].

2. Physical forcing data, as salinity, S, and water fluxes, Q_{river}, Q_{ocean} and Q_{ebm}. This data was inferred from values found in the literature [40, 71], based on theoretical circulation equations [73].

3. Initial values for the tracers' concentrations [40], C_{N} and C_{PHY}, once the mathematical model is already stated as an ODE system.

Further details on the model and the data considered are explored in the following.

3.2.1 A conceptual biochemical model for the NP dynamics

Phytoplankton populations, in general, are extremely sensitive to physical and chemical changes in the aquatic environment, resulting in rapid variations in their rates of reproduction and mortality. In PEC waters, it is known that phytoplankton is dominated by diatoms, especially during the driest season, which occurs in winter. This prevalence may be related to the high competitiveness of diatoms in nutrient-rich environments, such as estuaries. Although estuarine conditions vary significantly in terms of salinity and turbidity [61], diatoms thrive and often dominate these ecosystems due to their adaptability and efficient nutrient processing. In addition to diatoms, other phytoplankton groups can also be found in the PEC [62], as we can see in Figure 3.2), but due to their occurrence in smaller proportions (Figure 3.3), they will not be considered in this model. The reproduction of phytoplankton in the PEC and its estimated lifespan depend on several factors, including the incidence of sunlight, salinity, water temperature, and the residence time of water in the PEC (see Box 3.1).

Box 3.1: Residence time of a marine environment

The residence time of a marine environment, such as the Paranaguá Estuarine Complex, is defined as the average time interval required for the entire volume of water contained in the estuary, as well as the materials dissolved in it, to be completely renewed through exchanges with the ocean. It is estimated that the residence time of the PEC is on the order of 3 to 3.5 days (Lana et al., 2001), a value considered short and indicative of a highly dynamic, efficiently renewed environment.

Figure 3.3 shows estimates on the diatom contributions for the chlorophyll-a mass in PEC. Let C be the cellular carbon on phytoplankton cells and V the volume of a phytoplankton cell. For converting the total observed cells into carbon mass, we considered the equations:

1. For diatoms:

$$C = 0.288/V^{0.811}$$

Figure 3.2 – Boxplot of the main phytoplankton groups observed in the PEC waters.
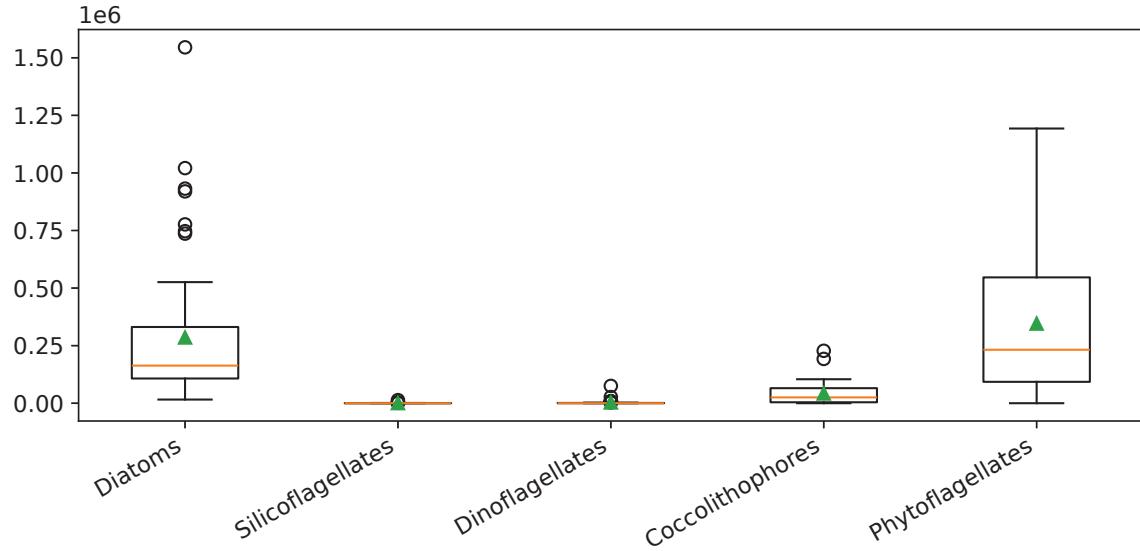
The figure shows the distribution of cell abundance (cells m^{-3}) across the approximately one-year period time series for each phytoplankton group (Diatoms, Silicoflagellates, Dinoflagellates, Coccolithophores, and Phytoflagellates). The box central line denotes the median, while the lower and upper edges correspond to the first and third quartiles. The whiskers represent values within $1.5 \times$ IQR (interquartile range), and points beyond these limits indicate outliers.

Source: the author (2026).

Data: Rezende; Brandini (2022).

2. For the other phytoplankton groups:

$$C = 0.216/V^{0.939}$$

For each group, we calculate V as the volume of a sphere with the cell's diameter. Last, we use the chlorophyll-a: Carbon ratio, $Chl : C = 0.02$, to calculate the chlorophyll-a mass per phytoplankton cell.

Let us recall the NP-model diagram, presented in Figure 2.1. With the intention of carrying out numerical tests, we will consider nitrate as the nutrient tracer not only due to data availability, but also because it is the most limiting nutrient in marine environments and a strong indicator of eutrophication in coastal areas. Let C_N be the concentration of N and C_{PHY} the concentration of phytoplankton measured in mmol m^{-3}, I the light intensity due to solar incidence measured in lm, T the temperature in ${}^oC$, S is the salinity (dimensionless). For describing the growth and mortality of phytoplankton depending on the nutrient availability, we follow [67] and define the SMS equation for phytoplankton as:

$$\frac{dC_{PHY}}{dt} = [V_{max} \cdot g(C_N, T, S, I) - \lambda] \cdot C_{PHY} \;, \tag{3.1}$$

The maximum specific growth rate of phytoplankton V_{max} is defined as the maximum growth of the phy-

Figure 3.3 – Estimated chlorophyll-a proportion by phytoplankton group (size-based) on PEC.
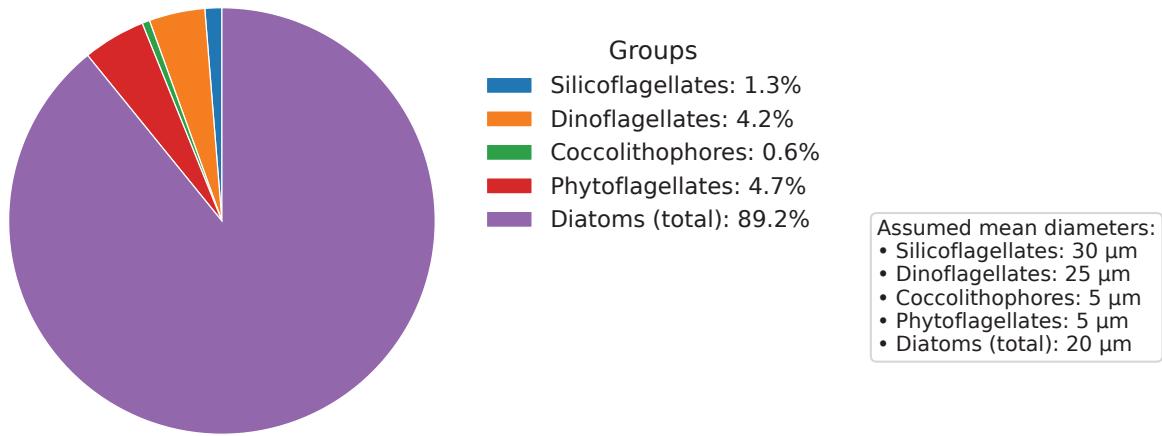
Data derived from the mean cell abundance and the volume-to-carbon conversion based on the assumed mean cell diameter for each group, assuming a constant chlorophyll-to-carbon ratio among groups. Percentages in the legend indicate each group's relative contribution.

Source: the author (2026).

Data: Machado *et al.* (1997); Menden-Deuer; Lessard (2000).

toplankton population depending solely on the existing population, under ideal conditions. The function $g(C_{\mathrm{N}}, T, S, I)$ represents the limitation of P growth due to the availability of nutrients and sunlight, variations in temperature, and salinity. We consider the mortality rate of phytoplankton as a linear function with constant rate λ. For defining the limiting function g, we consider a set of four limiting functions corresponding to each of the limiting elements. We have:

$$g(C_{\mathrm{N}}, T, S, I) = \eta(C_{\mathrm{N}}) \cdot \alpha(T) \cdot \beta(S) \cdot \gamma(I) \ , \tag{3.2}$$

with

$$\eta(C_{\mathrm{N}}) = \frac{C_{\mathrm{N}}}{C_{\mathrm{N}} + K} \ .$$

The Monod constant K is the concentration of nitrate such that the growth of phytoplankton is equivalent to half of the maximum possible growth, V_{max}. For nitrate, we have K between 0.1 and 0.3 mmol m^{-3}. In this model we will also consider $V_{\mathrm{max}} = 1.4$ day^{-1}, $K = 0.1$ mmol m^{-3} , $\lambda = 0.05$ day^{-1} [67].

The average water temperature varies between 23°–30° in summer and 18°–25° in winter [40], so the water temperature in winter is not a limiting factor on the phytoplankton growth [24]. However, in summer, phytoplankton production can be limited to up to 60% of its maximum if the temperature reaches 30°. Since T is the average temperature in ^{o}C of the water in the upper tank during the period that the model will

predict, we then define

$$\alpha(T) = \min\left\{1, \ 0.6 + 0.08 \cdot (30 - T)\right\} . \tag{3.3}$$

Equation (3.3) is obtained by assuming that temperatures in the range 18ºC to 25ºC doesn't present a limiting impact on the PEC phytoplankton growth, thus $\alpha(T) = 1$ for these temperatures, and by linear interpolation between temperatures 25ºC to 30ºC, considering $\alpha(25) = 1$ and $\alpha(30) = 0.6$ (Figure 3.4).

Figure 3.4 – Temperature limiting impact on PEC phytoplankton growth.
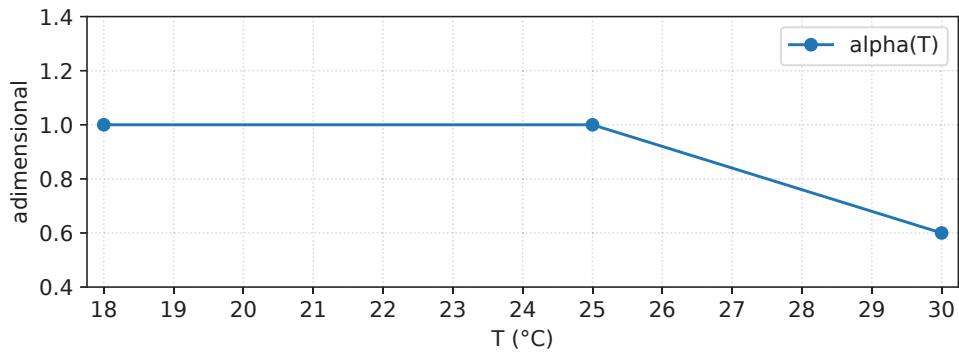
Salinity can also be a limiting factor for phytoplankton growth. Khan et al. [24] state that the ideal salinity varies from 20 to 35. As in the PEC the salinity does not exceed 33, we define

$$\beta(S) = \max\left\{0, \min\left\{1, \ 1 - 0.07 \cdot (20 - S)\right\}\right\} . \tag{3.4}$$

Equation (3.4) is obtained by assuming that salinities in the range 20 to 35 doesn't present a limiting impact on the PEC phytoplankton growth, thus $\beta(S) = 1$ for these salinities, and by linear interpolation between salinities 10 to 20, considering $\beta(10) = 0$ and $\beta(20) = 0$. We also ensured $\beta(S) \geq 0$, which could result in errors in extreme scenarios, as $S \leq 5$ (Figure 3.5).

The sunlight intensity reaching the PEC surface varies from the minimum value 2.9 $kWh \ m^{-2} \ d^{-1}$ in winter to the maximum value 5.6 $kWh \ m^{-2}d^{-1}$ in summer [21]. Therefore, the incidence of sunlight is not a limiting factor for the growth of phytoplankton in the estuary. We consider

$$\gamma(I) = 1 .$$

Figure 3.5 – Salinity limiting impact on PEC phytoplankton growth.
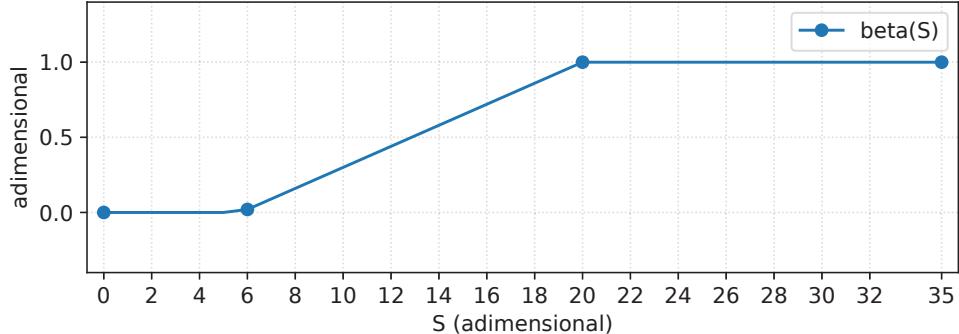
In relation to phytoplankton, the SMS nitrate equation is written as:

$$\frac{dC_{\mathrm{N}}}{dt} = [-V_{\max} \cdot g(C_{\mathrm{N}}, T, S, I) + r \cdot \lambda] \cdot C_{\mathrm{PHY}} \ , \tag{3.5}$$

The percentage remineralization of dissolved organic matter into nitrate is represented by the parameter r and set 0.7 [67]. All the parameter values used in the modeling of the PEC are presented in Table 3.1.

Table 3.1 – Parameter values set for the PEC conceptual model developed in this chapter.

Parameter	Value	Unit	Description
$V_{\max}$	1.4	day^{-1}	Maximum specific growth rate of phytoplankton
λ	0.05	day^{-1}	Linear mortality rate of phytoplankton
K	0.1	mmol m^{-3}	Monod constant for nitrogen
r	0.7	day^{-1}	Nitrogen remineralization rate

Sediment core studies indicate that the PEC functions as an efficient sediment trap, where the estuarine turbidity maximum promotes the retention of fine, organic-rich particles [76]. These conditions support the assumptions in our model regarding light limitation, stratification, and the high remineralization rates applied to phytoplankton biomass.

The set of ODE equations $\left\{\dfrac{dC_{\mathrm{N}}}{dt}, \dfrac{dC_{\mathrm{PHY}}}{dt}\right\}$, described on equations (3.5) and (3.1), represents the biochemical SMS equations for our model. Since we want these equations to depend only on the time t, the function $g(C_{\mathrm{N}}, T, S, I)$ needs to be described in terms of t only. As $\gamma(I)$ is taken as constant and C_{N} is com-

puted when integrating the ODE system, it lasts to create functions $T(t)$ and $S(t)$ based on the interpolation of sampled data found in the literature or theoretical assumptions. Both of these choices are possible. Here we consider the first of these approaches, which is better described in the following subsection, once it is intrinsically related to the physical model portraying the PEC environment.

3.2.2   A conceptual box-model representing geophysical interactions

We will now present a simple circulation model for the PEC waters. Considering that the PEC is a partially mixed estuary, we will use a three-dimensional two-box model for water circulation based on Verri et al. [73], represented in Figure 3.6.

Figure 3.6 – Circulation model for the PEC.
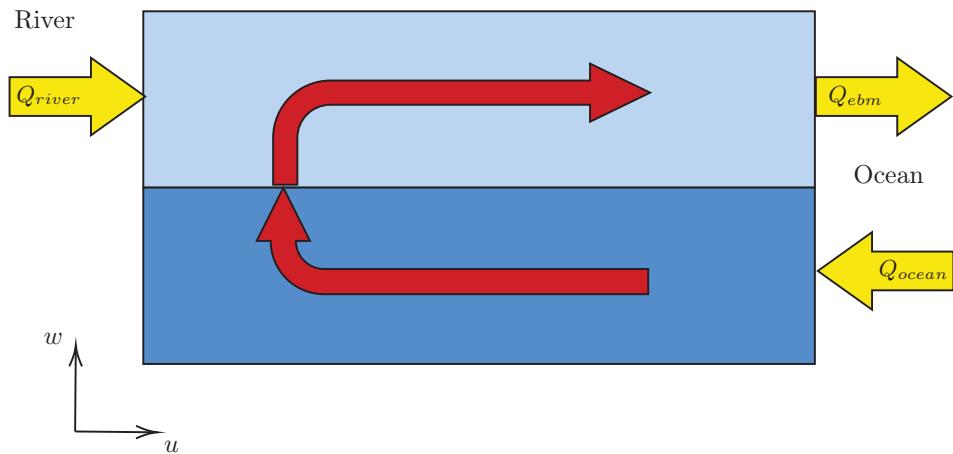
The blue rectangles represent a side view of the PEC aquatic environment, with the estuarine head on the left and the mouth on the right, defining the model domain. The arrows point in the direction of the advective flux. Yellow arrows indicate interactions with rivers and the ocean, whereas red arrows represent water circulation within the estuary complex.
Source: the author (2026).
Data: Verri *et al.* (2020).

Boxes dimensions   The surface area of the PEC is estimated as $601 \times 10^6 \text{m}^2$, its volume is estimated as $2 \times 10^9 \text{m}^3$, and the length of its three tidal inlets combined is estimated as $5.7 \times 10^3 \text{m}$ [9]. In our model, we represent the PEC environment with two stacked tridimensional boxes whose dimensions are represented in Figure 3.7. The motivations for the choice of the boxes' dimensions as well as the dimensions themselves, are described in the following:

L_y: Approximate sum of the lengths of the tidal inlets of the PEC, that is, $L_y = 5.7 \times 10^3$ m.

L_x: Approximate PEC surface area divided by L_y, that is, $L_x = 105.4 \times 10^3$ m. This choice comes from the fact that the depth itself does not directly interfere with the amount of nutrients in the water since the maximum depth reached in the estuary is less than 35m, but it directly interferes with the incidence of sunlight, which affects the growth of phytoplankton. Due to the features of the internal circulation of the PEC, phytoplankton will be modeled only in the upper box, which is another motivation to consider that the surface of the entire estuary is contained in the upper box.

H: The depth of the PEC waters is not homogeneous, with an average between 3 m and 5.4 m, reaching more than 15 m in some regions. For simplicity, we consider that the two boxes have the same height, that is, $H_{up} = H_{low} = H/2$, where H_{up} and H_{low} denote the heights of the upper and lower boxes, respectively. We define H as the approximate volume of water in the divided by its surface area, which is $H = 3.4$ m.

Figure 3.7 – Discretization of the PEC domain into two boxes.
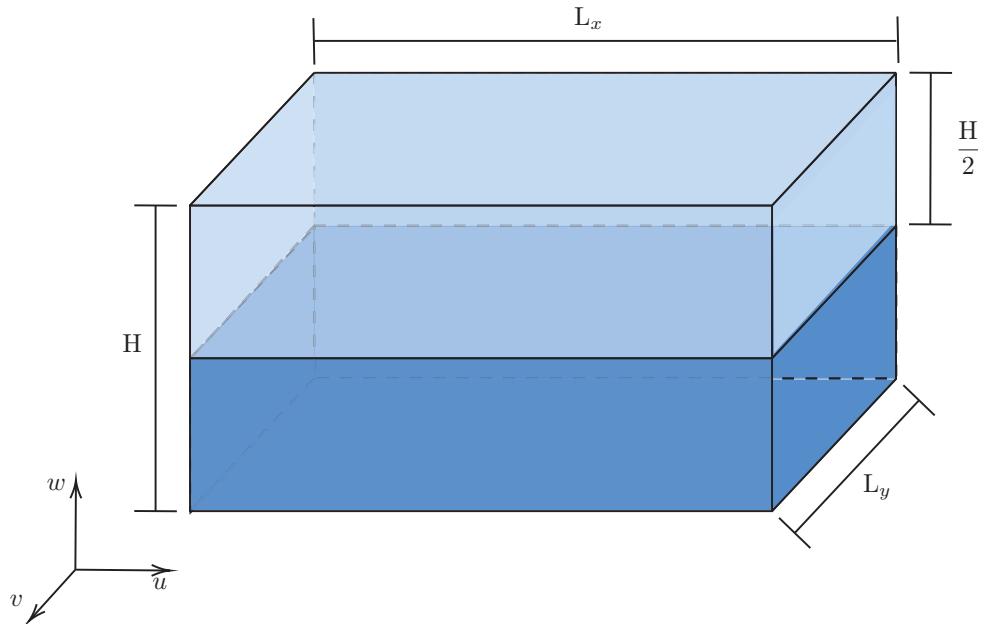
L_x is the horizontal dimension representing the path from the PEC head to its mouth, L_y is the horizontal dimension representing the width of PEC margins and H is the vertical dimension representing PEC depth. The domain is subdivided into two regions with equal volume, each with dimensions L_x, L_y and $\dfrac{H}{2}$.

Source: the author (2026).

Data: Cattani (2009).

Volumetric flow and salinity   For the approximate calculation of water volume flow and salinity, we will consider the Knudsen model (KEBM), reviewed by Verri et al. [73]. We will calculate an approximation

of horizontal averages for flow and salinity, where we disregard heat fluxes, precipitation, wind and water temperature. We will also consider the average over the tidal cycle, approximating the lunar day by the solar day with the average tide during it, which is consistent with partially mixed estuaries like the PEC. KEBM considers the following water flows entering the estuary: Q_{river} represents the flow coming from rivers entering the estuary at the head of the estuary; in the upper box, Q_{ocean} represents the flow coming from the ocean entering the estuary through the lower box and Q_{ebm} represents the water flow leaving the estuary through the estuary tidal inlets, in the upper box (Figure 3.6).

All these flows are measured in $\text{m}^3 \, \text{day}^{-1}$ and, as by hypothesis, the volume of water in the estuary is constant, it is worth the relation (3.6). Furthermore, we assume that the salinity S_{low} of the lower box is equal to that of the ocean salinity, and the salinity in the upper box, denoted S_{up}, is calculated as a solution to the equation (3.7). We consider as input data for this model:

1. $S_{\text{low}}(t)$ and $S_{\text{up}}(t)$ the mean salinity in the lower and upper boxes of the model [40] (Figure 3.8), respectively, at time instant t;

2. The flow Q_{river} is derived from the PEC residence time, which is approximately 3.49 days [38];

3. Estimates for the flows Q_{ocean} and Q_{ebm} obtained from the equation (3.6) and available data on current velocities [9] (Figure 3.9).

$$Q_{\text{ebm}}(t) = Q_{\text{river}}(t) + Q_{\text{ocean}}(t) \, , \tag{3.6}$$

$$S_{\text{up}}(t) \, Q_{\text{ebm}}(t) = S_{\text{low}}(t) \, Q_{\text{ocean}}(t) \, . \tag{3.7}$$

Estimates for nitrate concentration in rivers based on weighted average data from [50] (Figure 3.11) and estimates for water flows are shown in Figure 3.9, with the rainy season corresponding to the summer and the dry season corresponding to the winter. Apart from the nitrate input from ocean water, we consider the hypothesis that rivers are not responsible for the totality of the remaining inorganic nitrate input in the PEC. This way, we estimate the concentration of nitrate coming from the atmosphere, human activity, and other sources, including rivers, as proportional to the mass of nitrate coming from the rivers, based on calibration results, presented in *Chapter 6*. Recent geochemical analyses of sedimentary organic matter in the PEC [76] show that the system is strongly dominated by terrigenous inputs, with marine-derived material becoming more relevant only near the estuarine mouth and in the northern sector of the complex. This pattern is consistent with the structure of the nutrient sources considered in our model, where riverine nitrogen supply plays a central role relative to oceanic inputs.

As an aside, Box 3.2 briefly describes an alternative approach for estimating salinity when no observational data are available.

3.2.3   A conceptual biogeochemical model for PEC

We consider some additional assumptions to create a model that couples the SMS equations and the circulation model. We consider the tracers' supply from the lower box to the upper box to occur exclusively through advection proportional to the transport of salinity[1]. We also consider a remineralization rate of phytoplankton that dies equal to 70%, and the other possible factor for reducing its concentration in the upper box is its export to the ocean, since in our model, there is no advective flow that transports it to the lower box. We will also make use of the following notations:

- Vol_{box} : The volume of the upper box (which, in this case, is equal to the volume of the lower box), calculated as $\text{Vol}_{\text{box}} = L_x \cdot L_y \cdot H/2 = 10^9 \text{ m}^3$.

- T : The upper box temperature (Figure 3.10), in ^{o}C.

- $C_{\text{N}}^{\text{river}}$: Average nitrate concentration in the Nhundiaquara river, used for estimate the riverine and pluvial water drained to the PEC (Figure 3.11), in mmol m^{-3}.

- $C_{\text{N}}^{\text{low}}(t)$: The lower box average nitrate concentration (Figure 3.12), in mmol m^{-3}.

In Figure 3.12, note that the behavior of the blue plot (Upper box) is what we want to reproduce from our model for the nitrate tracer concentration throughout the year, while the other data (Lower box) is considered as data input for the model.

We also considered the entry of phytoplankton into the upper tank via advective flow from the lower box. This factor is based on the available data (Figure 3.13) and may be a cause for the seasonal variation in

---

[1]We do this in order to reduce the number of differential equations in the model, thus keeping it computationally inexpensive to be explored on a personal computer.

the composition of the phytoplankton population in the PEC. In Figure 3.13, also note that the behavior of the blue plot (Upper box) is what we want to reproduce from our model for the nitrate tracer concentration throughout the year, while the other data (Lower box) is considered as data input for the model.

Figure 3.8 – PEC daily data estimates on salinity over a one-year period.
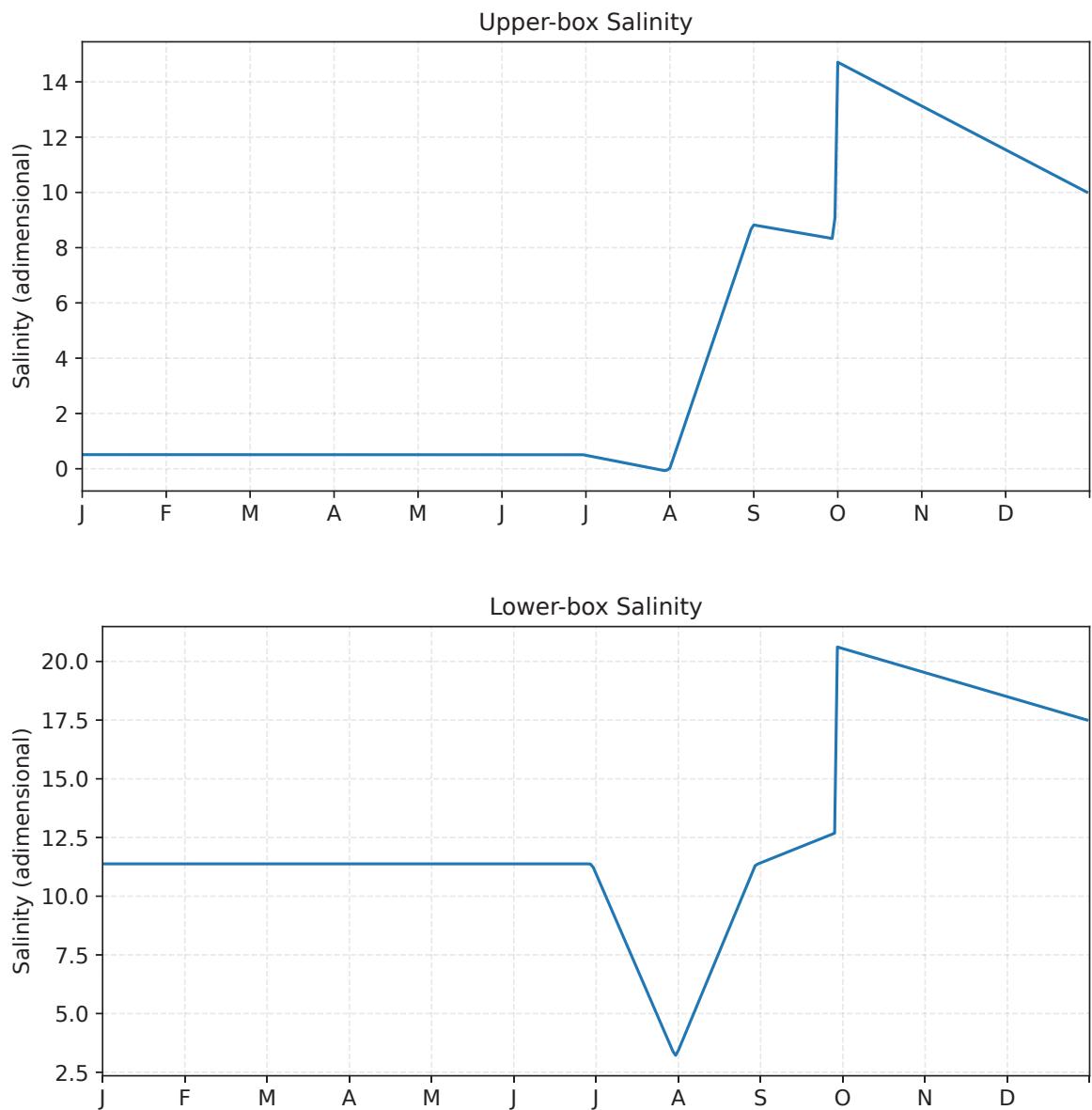
Figure 3.9 – **PEC daily data estimates on circulation fluxes over a one-year period.**

Flows over the Year
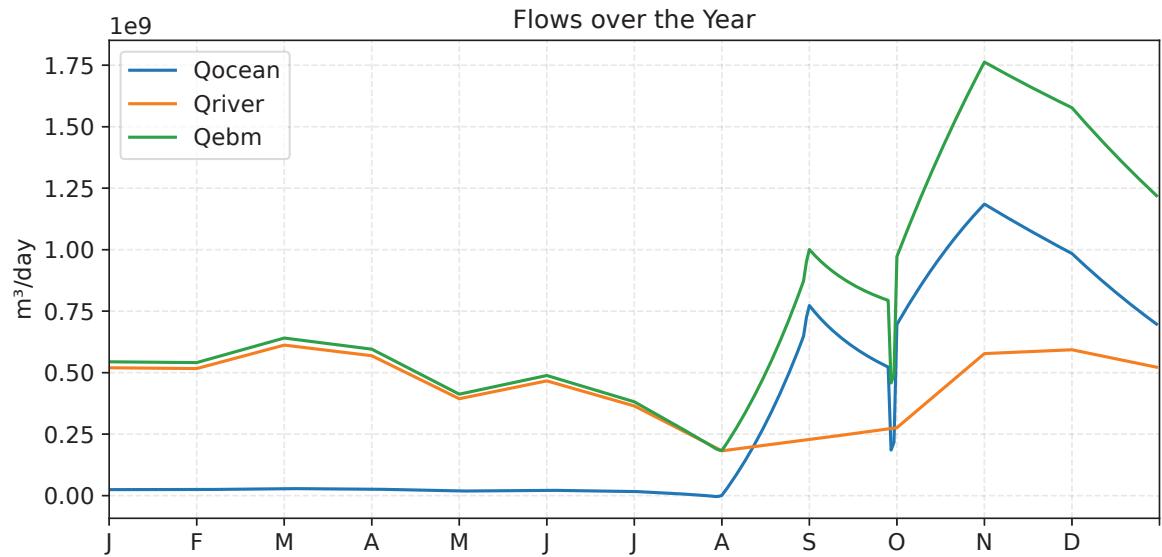
Figure 3.10 – **PEC daily data estimates on the water temperature of the upper region of PEC over a one-year period.**

Upper-box Temperature
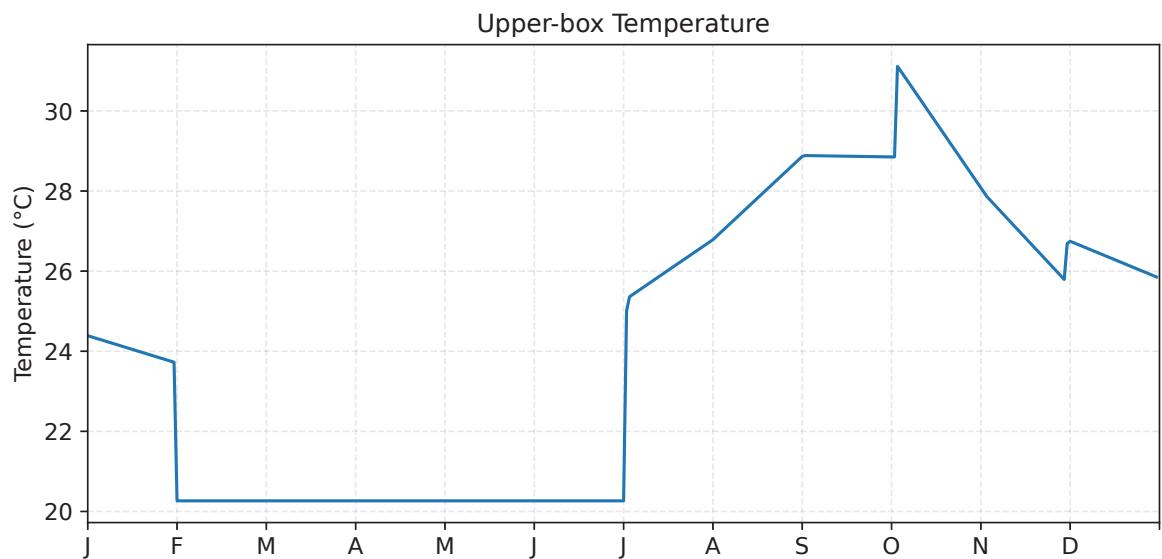
Figure 3.11 – PEC daily data estimates on the Nhundiaquara river nitrate concentration over a one-year period.
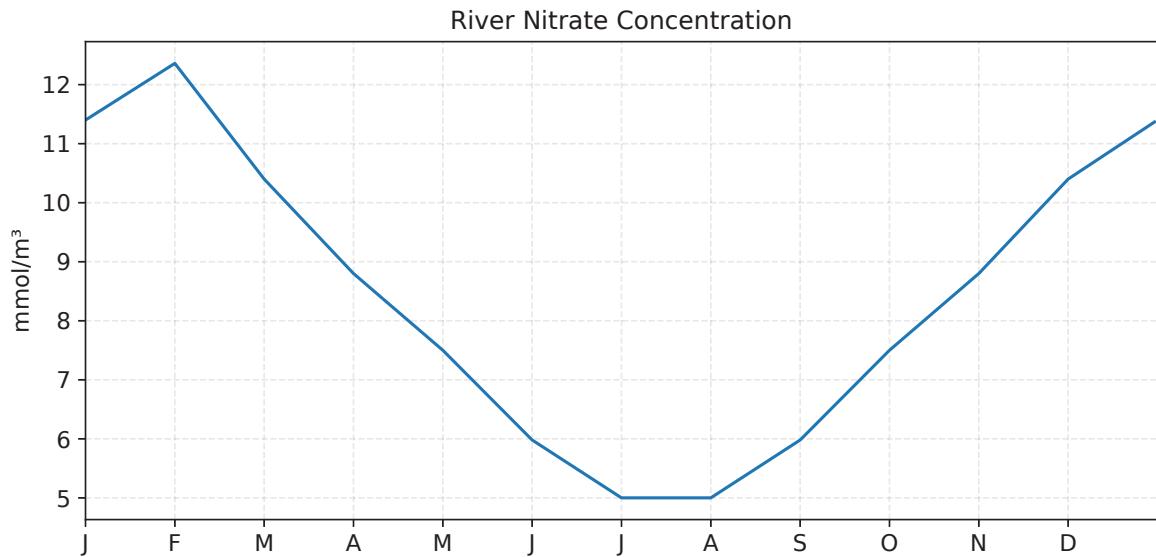
Figure 3.12 – PEC daily data estimates on nitrate concentrations over a one-year period.
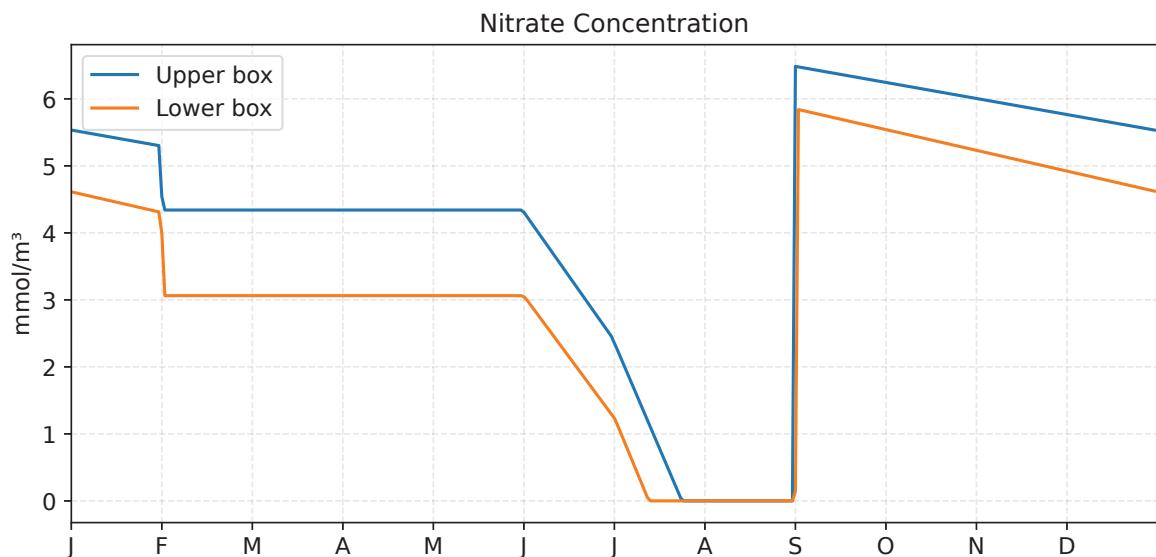
Recalling that our initial objective was to model the NP dynamics in the PEC and considering that in

Figure 3.13 – PEC daily data estimates on chlorophyll-a concentrations over a one-year period.

Phytoplankton Concentration
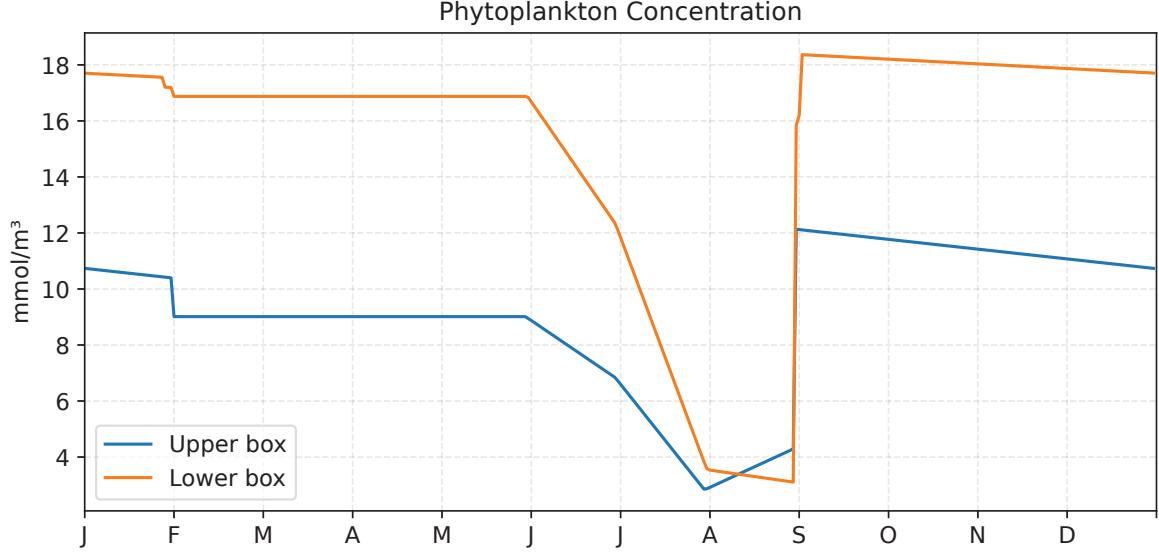
our circulation model these processes only occur in the upper box, the coupling of the circulation model to the conceptual model we developed before is represented in Figure 3.14.

Considering the model schematically represented in Figure 3.14, we obtain the equations for the concentration variations of nutrient (3.8) and phytoplankton (3.9) in the upper box and consider the stationary model (3.10).

$$
\frac{dC_{\mathrm{N}}(t)}{dt} = \frac{0.588\ C_{\mathrm{N}}^{\mathrm{river}}(t) \cdot Q_{\mathrm{river}}(t)}{\mathrm{Vol}_{\mathrm{box}}} + \frac{C_{\mathrm{N}}^{\mathrm{low}}(t) \cdot Q_{\mathrm{ocean}}(t)}{\mathrm{Vol}_{\mathrm{box}}} + r \cdot \lambda \cdot C_{\mathrm{PHY}}(t)
$$

$$
- \alpha(T(t))\ \beta(S_{\mathrm{up}}(t)) \frac{V_{\max} \cdot C_{\mathrm{PHY}}(t)\ C_{\mathrm{N}}(t)}{C_{\mathrm{N}}(t) + K} - \frac{C_{\mathrm{N}}(t) \cdot Q_{\mathrm{ebm}}(t)}{\mathrm{Vol}_{\mathrm{box}}}\ . \tag{3.8}
$$

$$
\frac{dC_{\mathrm{PHY}}(t)}{dt} = C_{\mathrm{PHY}}(t) \cdot \left(\alpha(T(t))\ \beta(S_{\mathrm{up}}(t)) \frac{V_{\max} \cdot C_{\mathrm{N}}(t)}{C_{\mathrm{N}}(t) + K} - \lambda - \frac{Q_{\mathrm{ebm}}(t)}{\mathrm{Vol}_{\mathrm{box}}} \right) +
$$

$$
+ \frac{C_{\mathrm{PHY}}^{\mathrm{low}}(t) \cdot Q_{\mathrm{ocean}}(t)}{\mathrm{Vol}_{\mathrm{box}}}\ . \tag{3.9}
$$

45

$$\frac{dC_{\mathrm{N}}(t)}{dt} = 0 \ , \ \frac{dC_{\mathrm{PHY}}(t)}{dt} = 0 \ . \tag{3.10}$$

Before presenting the results, we briefly introduce the concept of spin-up time (Box 3.3), which is used to interpret model equilibrium.

> **Box 3.3: Spin-up time in marine biogeochemical models**
>
> The spin-up time of a marine biogeochemical model is the period required for the model to reach a dynamic and thermodynamic state close to equilibrium under a fixed set of observational forcings [67]. When modeling an estuarine environment, the spin-up time tends to be shorter, as these environments are dynamic and variations occur on short time scales.
>
> Mathematically, the spin-up time can be viewed as an integration interval long enough for the differential equations defining the model to reach equilibrium.
>
> Conceptually, we can say that during the spin-up time, the model "forgets" its initial conditions and begins to respond only to observations and to itself.

Figure 3.14 – A conceptual biogeochemical NP-model for PEC.
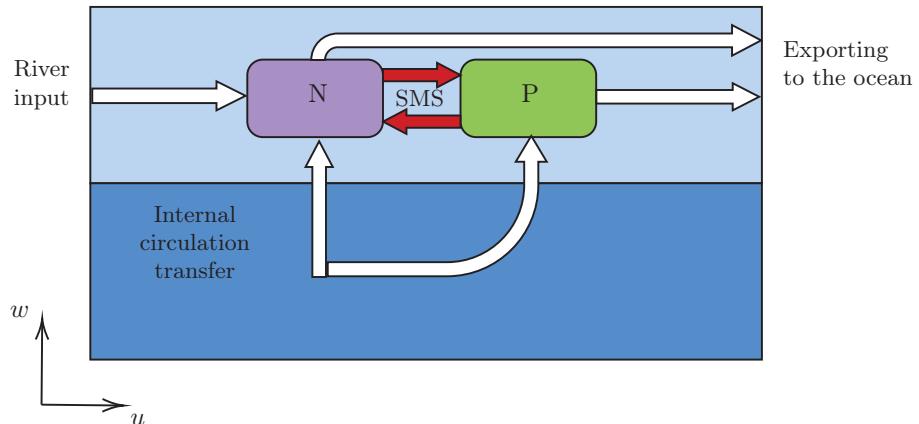
Red arrows represent biochemical interactions, while white arrows represent geophysical circulation transfers. N and P denote the mean concentration of the tracers nitrate and phytoplankton in the upper box, respectively. Arrows pointing towards a tracer box represent a source of the tracer, whereas arrows pointing outward from a tracer box represent a sink of the tracer.
Source: the author (2026).

3.3 Results and discussion

This section summarizes the results obtained with the conceptual NP model developed for PEC. The simulations aim to assess the model's internal consistency and its ability to reproduce the expected seasonal behavior of nutrients and phytoplankton in the PEC environment. Rather than providing quantitative forecasts, the results serve to evaluate how well the simplified structure captures the dominant processes and interactions described in previous field and modeling studies.

This model was implemented and run in Python with some adjustments (see the *Appendix*).

Although the model presented employs several simplifications and rough estimates, the results obtained are consistent with the available literature data. In Figure 3.15, we present a forecast of the concentrations, in mmol m^{-3}, of nitrate and phytoplankton in the upper box of the estuary throughout one year. As expected, phytoplankton concentration is generally higher during the dry season than during the rainy season. Estimates for nitrate concentration were also close to the observational data.

According to Mizerkowski et al. [50], there is no consistent relation between the volume of water discharged by rivers in the PEC and the amount of nutrients. Therefore, nutrient concentrations in the rivers that drain the PEC exhibit significant variability, necessitating weekly or even daily monitoring for improved model accuracy. This would also make it possible to identify the input of nutrients transported from the coast to the estuary by rainwater and possibly find a relationship between rainwater discharge and nutrient input in the PEC. The salinity of the estuary can vary due to precipitation, evaporation, and discharge of water and salts through rivers. Due to this change, salinity can damage cell membranes and change the density of seawater and, therefore, the buoyancy of certain organisms due to the water balance of cells [15], so the data availability for the average salinity in PEC waters is essential for the purposes of limiting phytoplankton growth.

Low nutrient concentration is the main limiting factor for phytoplankton growth in our model, as any excess will be converted into biomass, while nitrate concentration values below 0.00037 mmol m^{-3} inhibit the growth of phytoplankton [67]. The the primary reason for the time-step in our model being counted in days, rather than fractions of a day, is that representing variations in phytoplankton growth and mortality at different daily times would necessitate a significantly greater degree of complexity. The main sources of nitrate for the estuary are the rivers that drain their respective watersheds, the ocean itself through advection and turbulent mixing – which is not explored in our model, due to complexity increasing – and the recycling of nearby photoautotrophic organisms [49, 67]. Furthermore, anthropogenic input is a crucial factor to consider when modeling estuarine environments (see Box 3.4).

Figure 3.15 – NP-model: Predicted behavior for the nitrate and phytoplankton concentrations along one year on PEC.
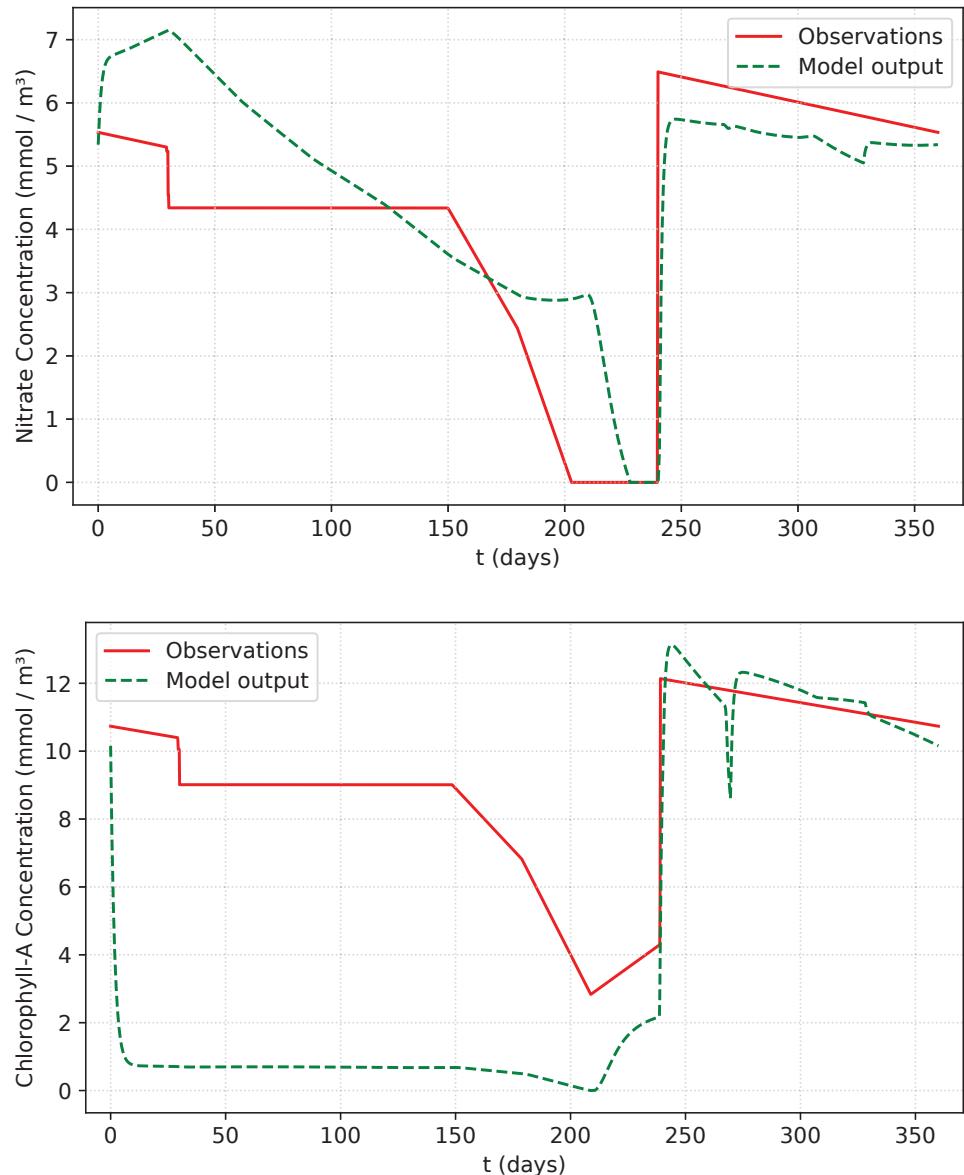
Considering the model defined by equations (3.8) - (3.10), the settings of Table 3.1, and the data presented in figures 3.8 - 3.13.

Source: the author (2026).

Box 3.4: Human impacts on estuarine environments

Estuaries are bodies of water characterized by the interconnection between marine and river environments. They play a fundamental role in human development, as, in addition to providing natural protection for ports and industrial facilities by facilitating access to food and energy resources, estuaries also offer a range of ecological benefits, including nurseries for various marine species, water filtration, and flood protection [4, 38, 61]. In this context, mathematical modeling is a powerful tool for studying the impacts of human occupation on estuaries and developing effective measures for their preservation [17, 57, 73]. The intensification of human occupation in areas close to estuaries generates a series of environmental impacts, putting the natural balance of these ecosystems at risk. Pollution caused by the discharge of domestic and industrial effluents without adequate treatment, in addition to the flow of pesticides and other contaminants, is called *anthropogenic input* and contributes to the degradation of water quality in estuaries, which can lead to eutrophication, death of marine organisms, and harmful proliferation of algae [43, 44, 50, 60, 78]. This disturbance of the natural balance of estuaries can have serious consequences for the environment and for the communities that depend on these ecosystems for their sustenance, as it can lead to a decrease in fishing productivity and the loss of habitat for threatened species [5, 12].

Regarding the nitrogen input into PEC coming from human activity, we can outline two main types: Diffuse sources come from surface runoff from agricultural and urban areas, from the fertilizer industry and docks in ports (Figure 3.16), while point sources (in concentrated form) of domestic or industrial origin come from sewage that flows into water bodies without prior treatment [50, 65].

Recent historical reconstructions of the PEC reveal strong anthropogenic influences, where activities such as agriculture, urban tourism, river diversion, and maintenance have reshaped the local hydrodynamics and modified its nutrient deposition patterns [76], suggesting the need to consider diffuse and point sources of anthropogenic nitrate supply when developing an NP model for the PEC.

The euphotic zone in the coastal zone may be restricted to just a few centimeters of the water column, as is the case in estuaries, where the waters are turbid by the presence of sediments, particulate matter, and humic acids [16]. This fact validates the design of our model, which considers the growth of phytoplankton only in the upper box. In addition to light being a limiting factor in this case, the scarcity of nutrients also plays a prominent role, regulating the development of the phytoplankton community and, consequently, the other organisms that make up the food web in that habitat [6]. The zone of maximum turbidity (ZMT) present in partially mixed estuaries, such as the PEC, functions as a sediment trap, as it maintains the fine particles that make up the sediment for a long time, generally composed of material particulates, nutrients, organic matter, and pollutants, as they constitute the part that adheres or is adsorbed to these substances. Furthermore, the change in the ZMT can erode the bottom when the tide is falling, and at high tide, it can deposit this material [79]. This fact supports the choice of the parameter for phytoplankton remineralization

rate as 70%, even though the model does not explicitly consider flows indicating this rate.

The phytoplankton concentration is estimated through the analysis of the chlorophyll-a concentration in the water, and we use here the seasonal values presented by [4] for comparison with the results of our model. The results obtained by simulating these scenarios with our model confirmed these assumptions. Furthermore, studies indicate that a longer residence time, salinity, and anthropogenic input directly influence the increase in phytoplankton concentration, which is an indication of environmental imbalance, while frequent rains decrease salinity, nutrient concentration, and residence time of water in an estuary, which can cause a drastic reduction in the total phytoplankton concentration [3]. Studies indicate that a longer residence time, salinity, and anthropogenic input directly influence the increase in the phytoplankton concentration, which is an indication of environmental imbalance, while frequent rains decrease the salinity, nutrient concentration, and residence time of water in an estuary, causing a drastic reduction in the local phytoplankton concentration [3].

Mathematical modeling is a helpful tool for studying the dynamics of estuaries. However, in order to make an effective contribution, it is necessary to know the processes involved to facilitate the creation and maintenance of the model, for example, by understanding how effectively nitrate limits phytoplankton productivity, because this conclusion is based on a few studies that evaluate nitrate limitation indirectly [78].

3.4   Model limitations and development opportunities

In this modeling study, the flows were determined based on assumptions, rather than direct measurements. As our goal is to exemplify the development and application of a simple biogeochemical model when considering assumptions for data extrapolation, even with some data available on the river fluxes, there are still unpredictable factors, such as the input of rainwater and non-mapped water sources. In this context, the flow Q_{ebm} was estimated based on the average residence time of the estuary, while the flow Q_{river} was calculated proportionally to the river discharge with the greatest contribution to the system, including the estimated water input from rainfall. The flows Q_{ebm}, Q_{river}, and Q_{ocean} were obtained from equations that relate flow rate and salinity, also using observed salinity data in the estuary. In the context of modeling, incorporating data from direct measurements of flows in the estuary could increase the model's representativeness in relation to the circulation dynamics of the PEC.

A promising possibility would be to replace the simplified two-box model with a more detailed circulation model of the PEC. This more complex model could include a more refined spatial discretization and, ideally, the use of time series of observational data of the flows, which involves increased computational cost.

Another parameter estimated in this study was the riverine input of nitrate into the PEC waters. Although precisely quantifying this input is challenging, the estuary model can be adapted to generate daily estimates of nitrate input, considering the sum of both fluvial and anthropogenic contributions. The use of a more sophisticated circulation model could improve the accuracy of these estimates.

Figure 3.16 – Diffuse anthropogenic supply simulation on PEC.
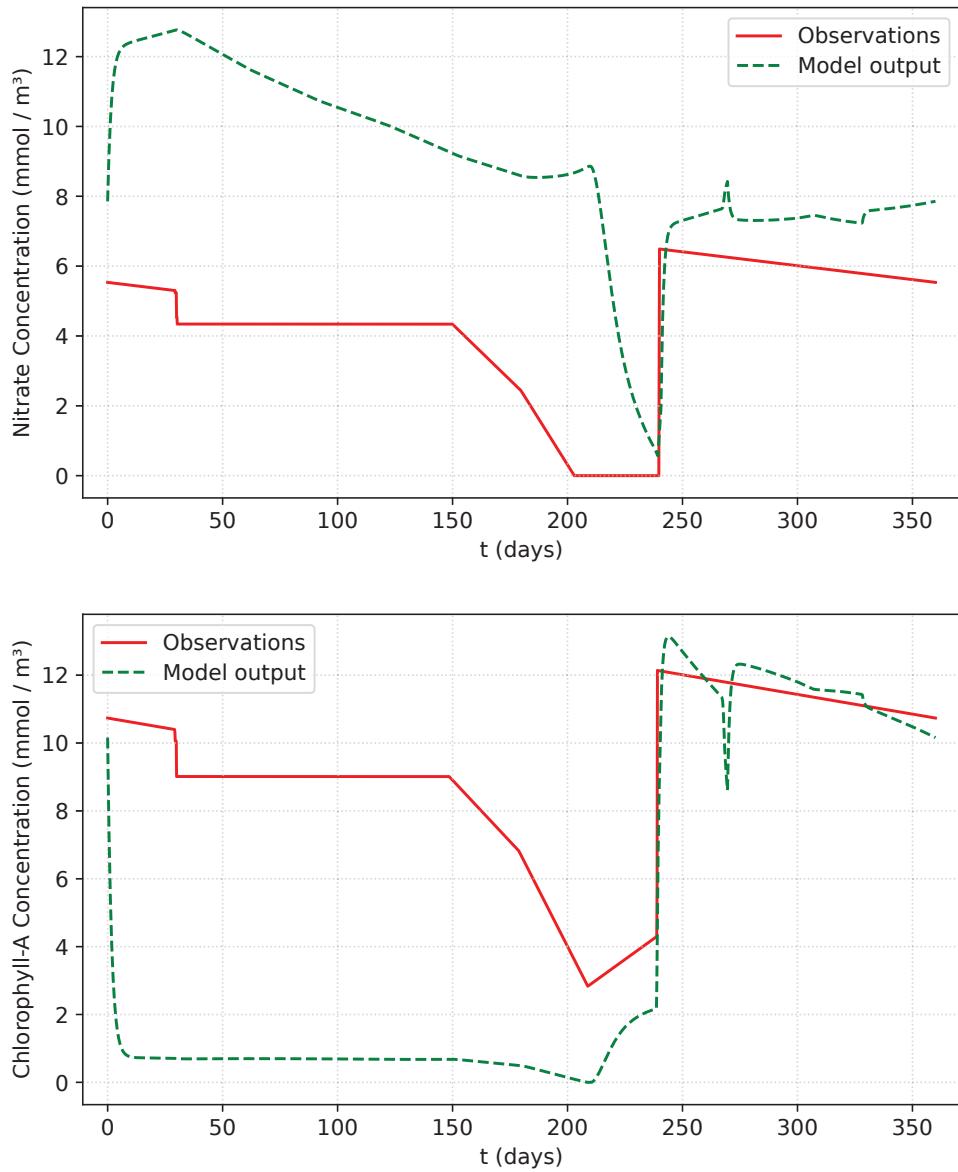
The diffuse supply of nutrients directly influences a greater concentration of phytoplankton. To illustrate this situation, we increased the concentration data input from the rivers in 10 mmol m^{-3}, while keeping all the other model settings and equations as in Figure 3.15. The plot shows the model's forecast during the year.

Source: the author (2026).

The availability of up-to-date and sufficient data for model execution, in order to reduce the need for data extrapolation, would be ideal for better output performance. The modeling of SMS functions is also an issue that can be revisited in the future.

Finally, we note that the data used for validating this model come from different sources and refer to different years. This impairs the representativeness of the model because, ideally, all data would be related to the same days of the same year. Even so, when simulating different theoretical scenarios, the model showed the expected behavior consistent with what is described in the literature. During dry periods, there was an increase in phytoplankton, which agreed with the expected estimate. Recent sedimentary reconstructions [76] suggest that long-term changes in the PEC are closely linked to land-use evolution, supporting the notion that simplified conceptual models can still capture the dominant local processes shaping nutrient and phytoplankton dynamics, while showing that the PEC has been well-monitored from a geochemical perspective, which is an interesting aspect to be explored further. Thus, rather than aiming for predictive accuracy, this model is intended as a computational testbed for the optimization and calibration procedures presented in *Chapter 6*. Although more complex biogeochemical models could represent the system in greater detail, they also require higher computational resources and remain constrained by limited observational coverage.

4 Model calibration and derivative-free optimization

In this chapter, we present the optimization strategies used to calibrate the model parameters, which will be applied in the following chapter. Parameter calibration is treated as an adjustment process: we seek parameter values that make the model outputs match the observed data as closely as possible. These methods are particularly suitable for problems where the objective function is expensive to evaluate, noisy, or non-smooth, which exploration is a natural progression of the work presented here. We also discuss how to define a misfit function that properly measures the difference between model results and observations, considering normalization, regional weighting, and data uncertainty.

First, we review the basic concepts of optimization, including local and global minima, constrained and unconstrained problems, and how multi-objective problems can be reformulated as least-squares problems. Next, we describe the construction of the misfit function and the techniques used to minimize it. At last, we review the Derivative-Free Optimization for Least Squares (DFO-LS) algorithm [32, 55]. Practical aspects are also discussed, setting the stage for the applications presented in the next chapter.

4.1 Basic concepts of optimization

Optimization is a branch of Mathematics that deals with problems of minimization or maximization of an objective under possible constraints. Such a problem must be described within a general formulation as

$$
\begin{aligned}
&\text{minimize} \ \ f(\omega) \\
&\text{subject to:} \ \ \omega \in U \subset \mathbb{R}^n
\end{aligned} \tag{4.1}
$$

where $U = \{\omega \in \mathbb{R}^n \ : \ g_i(\omega) \leq 0 \ , \ h_j(\omega) = 0 \ ; \ i = 1, \ldots, k; j = 1, \ldots, \ell\}$ is the feasible set, $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^n \to \mathbb{R}^\ell$ are the functions that define the constraints.

There are two standard solutions to (4.1): local and global minimizers. We can define them as follows.

Definition 4.1 (Local and global minimizers). *Considering the minimization problem (4.1), we say that $\bar{\omega}_L$ is a local minimizer or a local solution for the problem if, and only if*

$$f(\bar{\omega}_L) \leq f(\omega), \ \forall \omega \in B(\bar{\omega}_L, \epsilon) \cap U,$$

for some $\epsilon > 0$. Also, we say that $\bar{\omega}_G$ is a global minimizer or a global solution for the problem (4.1) if, and only if

$$f(\bar{\omega}_G) \leq f(\omega), \ \forall \omega \in U.$$

For practical reasons, it is unusual to search for a global minimizer for an optimization problem. In practice, finding global minimizers is rarely guaranteed; thus, a common strategy is to explore multiple local minimizers and select the best among them.

It is also possible to use previous knowledge to estimate which optimization algorithms are faster in calculating a solution for the optimization problem. This would be done by analyzing the features of the functions f, g, and h (for example, smoothness, noise, convexity) and finding an algorithm known by producing good results related to such features, but also by empirical testing on representative problems.

Another situation arises when we are dealing with *multi-objective optimization*. In this case, we would want to find a local optimizer for the problem (4.1), with the addition that $f : \mathbb{R}^n \to \mathbb{R}^m$, for some $m \in \mathbb{N}$. Several optimization techniques are directly applicable to such problems, although not every optimization method is specifically designed to handle multi-objective optimization. Yet, sometimes, it is possible to convert a multi-objective optimization problem into a *single-objective* optimization problem, as is the case of a least squares problem. This reformulation will be explored in the following sections.

4.2 Calibrating parameters from observations

In many applications, the forward model to be calibrated is, in practice, treated a *black-box*: the optimization algorithm sees only inputs (parameters) and outputs (diagnostics), but does not have access to the analytical form of the objective function. In ocean biogeochemistry, for example, the inputs are parameter values, and the outputs are predicted tracer fields. Calibration then seeks parameters whose model outputs best match an observation data set.

The known strategies for automatic parameter calibration [32, 55] involve optimizing an objective function known as the *misfit function*. This function quantifies the difference between model outputs and observation data and can be seen as a measure of calibration quality. We outline a generic construction and weighting scheme suitable for global applications (and used in complex settings such as [55]).

Let ω be a vector whose entries are the parameters we want to calibrate; $\mathcal{O}_{q,i,j}$ the observed concentration

of the tracer A_q on a subregion $R_{i,j}$, with volume $V_{i,j}$, inside a larger region R_j, with volume V_j, at a fixed moment in time; $m_{q,i,j}(\omega)$ the model prediction for the concentration of the tracer A_q on $R_{i,j}$, at the moment corresponding to the observation $\mathcal{O}_{q,i,j}$; where $\omega \in \mathbb{R}^{\ell_{\text{parameters}}}$, $q = 1, \ldots, \ell_{\text{tracers}}$, $i = 1, \ldots, \ell_{\text{j}}$, $j = 1, \ldots, \ell_{\text{regions}}$. We want to simultaneously minimize a set of functions $f_{q,i,j} : \mathbb{R}^{\ell_{\text{parameters}}} \to \mathbb{R}_+$, defined as:

$$f_{q,i,j}(\omega) = |m_{q,i,j}(\omega) - \mathcal{O}_{q,i,j}| , \tag{4.2}$$

for $\omega \in \mathbb{R}^{\ell_{\text{parameters}}}$, $q = 1, \ldots, \ell_{\text{tracers}}$, $i = 1, \ldots, \ell_j$, $j = 1, \ldots, \ell_{\text{regions}}$. Taking into account the fact that the parameters must lie on some defined intervals to make sense, there is also a set of constraints that must be satisfied, which we will not define yet, but call it $U \subset \mathbb{R}^{\ell_{\text{parameters}}}$. We could handle this task by using a strategy for simultaneously solving the constrained minimization problems:

$$\begin{aligned} \text{minimize} \quad & f_{q,i,j}(\omega) \\ \text{subject to:} \quad & \omega \in U , \end{aligned} \tag{4.3}$$

for $\omega \in \mathbb{R}^{\ell_{\text{parameters}}}$, $q = 1, \ldots, \ell_{\text{tracers}}$, $i = 1, \ldots, \ell_{\text{j}}$, $j = 1, \ldots, \ell_{\text{regions}}$.

One can notice that the problem (4.3) may be converted into a *least-squares problem*. This strategy reduces the problem's dimensionality, transforming it from a multi-objective optimization problem into a single-objective one. Considering that we can accommodate the problem constraints within the strategies to solve the optimization problem, we will now describe how to rewrite the objective function of the problem (4.3) as one corresponding to a least-squares constrained problem:

$$\begin{aligned} \text{minimize } f(\omega) = \sum_{q=1}^{\ell_{\text{tracers}}} r_q(\omega)^2 \\ \text{subject to: } \omega \in U \subset \mathbb{R}^{\ell_{\text{parameters}}} \end{aligned} \tag{4.4}$$

where r_q is the residual function corresponding to the tracer A_q, which will be better described in the following. We now discuss details on the residual formulation presented in [33]. First, consider a tracer A_q, a region R_j with its corresponding subregions $R_{i,j}$, and a set of corresponding observations. A first attempt to write a residual function $r_{q,j}(\omega)$ could be:

$$r_{q,j}(\omega) = \sqrt{\sum_{i=1}^{\ell_j} f_{q,i,j}(\omega)^2} \tag{4.5}$$

Mathematically, this is a valid attempt. But, numerically, it may lead to *ill-conditioning* due to several factors. For instance, in the context of marine biogeochemistry models, the observations for some tracers can

be typically very small in comparison to other tracers, so the minimization of $f(\omega)$ may end up not taking into account these observations. Moreover, the impact of the observations over the misfit function can be disproportional to the volume of the region sampled. To prevent both of these possible outcomes, we can consider the attribution of weights, leading to the following formulation for the residual function:

$$\tilde{r}_{q,j}(\omega) = \sqrt{\frac{V_j}{\sum_{j=1}^{\ell_{\text{regions}}} V_j}} \frac{\sqrt{\sum_{i=1}^{\ell_j} f_{q,i,j}(\omega)^2 \frac{V_{i,j}}{V_j}}}{\sum_{i=1}^{\ell_j} \mathcal{O}_{q,i,j} \frac{V_{i,j}}{V_j}} \tag{4.6}$$

where it is easy to notice that the weightings satisfy $\sum_{j=1}^{\ell_{\text{regions}}} \frac{V_j}{\sum_{j=1}^{\ell_{\text{regions}}} V_j} = 1$ and $\sum_{i=1}^{\ell_j} \frac{V_{i,j}}{V_j} = 1$.

Thus, we may define the residual function corresponding to a tracer A_q as:

$$r_q(\omega) = \sqrt{\sum_{j=1}^{\ell_{\text{regions}}} \tilde{r}_{q,j}(\omega)^2} \ . \tag{4.7}$$

During the calibration and validation steps of the biogeochemical model, the sparsity of data and noise pose significant obstacles to obtaining precise and reliable tracer observations. While the first problem can be mitigated by filtering on the optimized parameters, the latter can be handled by modifying the function $f_{q,i,j}$ to account for some noise. If we consider the modification of $f_{q,i,j}$ of the function $f_{q,i,j}^{\text{Noise}} : \mathbb{R}^{\ell_{\text{parameters}}} \to \mathbb{R}_+$:

$$f_{q,i,j}^{\text{Noise}}(\omega) = |m_{q,i,j}(\omega) - (\mathcal{O}_{q,i,j} + \varepsilon_{q,i,j})| \ , \tag{4.8}$$

we can now solve a minimization problem that accounts for uncertainties over the observations by solving the problem:

$$\text{minimize} \ \ f^{\text{Noise}}(\omega) = \sum_{q=1}^{\ell_{\text{tracers}}} r_q^{\text{Noise}}(\omega)^2$$

$$\text{subject to:} \ \ \omega \in U \ , \tag{4.9}$$

where the residual function corresponding to a tracer A_q is defined as:

$$r_q^{\text{Noise}}(\omega) = \sqrt{\sum_{j=1}^{\ell_{\text{regions}}} \tilde{r}_{q,j}(\omega)^2} \ , \tag{4.10}$$

with

$$\tilde{r}_{q,j}^{\text{Noise}}(\omega) = \sqrt{\frac{V_j}{\sum_{j=1}^{\ell_{\text{regions}}} V_j}} \frac{\sqrt{\sum_{i=1}^{\ell_j} f_{q,i,j}^{\text{Noise}}(\omega)^2 \frac{V_{i,j}}{V_j}}}{\sum_{i=1}^{\ell_j} \mathcal{O}_{q,i,j} \frac{V_{i,j}}{V_j}} \ . \tag{4.11}$$

In applied studies, specific index sets and data choices instantiate f^{Noise}; diverse DFO algorithms [8, 39, 63] have been used to minimize (4.9). Due to their advantages, these methodologies have already been applied to optimize models in more complex settings, such as in [31], where a sensitivity analysis reduced parameter dimension, and in [32, 55], where the optimization strategies Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Derivative Free Optimization for Least Squares (DFO-LS) delivered competitive fits under limited computational budgets.

4.3 Performance measure for an optimization algorithm

The performance of an optimization algorithm in model calibration depends mainly on two aspects: its computational cost and the quality of the fit it produces. At the same time, the accuracy with which the calibrated parameters reproduce the observed data must be evaluated using problem-dependent metrics, such as the least squares error or observational parameter recovery, as discussed in the following chapter. A reliable algorithm must therefore balance these two dimensions: obtaining a sufficiently accurate fit while keeping the computational costs compatible with the application's constraints. In well-defined calibration problems, this translates into convergence to a consistent set of parameters that performs well across all chosen metrics.[33].

When comparing the performance of two algorithms, several common situations must be considered. Some of these situations are directly related to the modeling and implementation of the fitting strategy, such as the model's or optimizer's ability to handle noise in the observational data and their sensitivity to parameter variations.

Another common problem is the scarcity of observational data, especially in the context of marine biogeochemistry. In the context of global ocean modeling, collecting data in remote environments, such as the deep ocean and oceanic regions near the poles, poses a significant challenge. Even when data collection is possible, observations (such as tracer concentrations) can vary significantly in time and space, making it difficult to obtain a complete dataset. Still, even in ideal situations in the aforementioned contexts, data may not be available because they are not yet being monitored. These limitations should also be taken into account when evaluating the performance of the algorithms.

A standard comparison protocol involves running both algorithms under identical conditions and recording the real-time iteration count and error obtained. These methodologies have already been applied to systematically compare the CMA-ES and DFO-LS algorithms in the context of global ocean biogeochemical modeling.[55].

Evolutionary computation algorithms are optimization methods inspired by the process of natural selection. They work by iteratively generating, evaluating, and improving a population of candidate solutions to a problem, mimicking biological evolution through mechanisms such as mutation, recombination, and selection. In this context, contrasting with the DFO-LS algorithm, which is based on a deterministic trust-region approach, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [18, 19] is a stochastic algorithm that relies on systematic sampling and the gradual improvement of the best-performing candidates for the parameters being optimized. Figure 4.1 shows a schematic representation of the CMA-ES algorithm iteration near convergence.

Figure 4.1 – Iterative scheme of the CMA-ES algorithm.
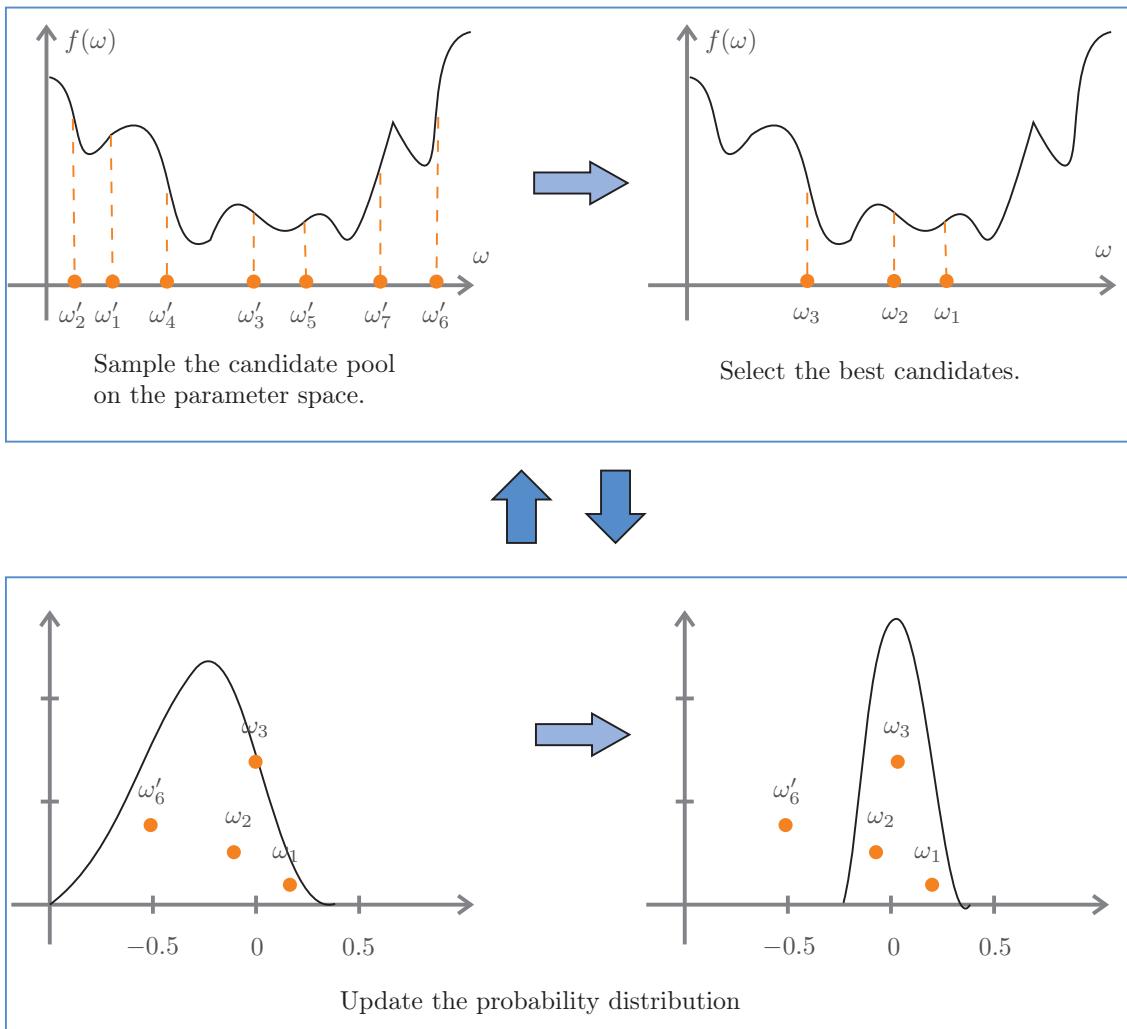
4.4   DFO-LS: Derivative Free Optimization for Least-Squares

An optimization strategy recently explored for parameter calibration is DFO-LS, which is a derivative-free trust-region method explicitly tailored to nonlinear least-squares objectives. It builds local surrogate models of the residual vector from function evaluations only and chooses steps by minimizing a quadratic approximation within a trust region, which can markedly reduce the number of expensive model evaluations when the objective is a sum of squared residuals [7, 55]. Thanks to its residual-aware design, DFO-LS is well suited to noisy, nonconvex black-box calibrations with simple constraints.

In the following, we outline the core elements of DFO-LS, as interpolation-set management, linear residual modeling, quadratic objective modeling, trust-region subproblems, acceptance/updates, and restarts. We also note that methods of this class have been applied in complex biogeochemical settings [55], illustrating their utility beyond conceptual problems.

DFO-LS is a trust-region least-squares minimization method. Compared with evolutionary strategies (see Box 4.1), DFO-LS exploits the least-squares structure explicitly, often reducing the number of expensive model evaluations on noisy, nonconvex black-box problems [55]. Owing to these advantages (no gradients required, robustness to noise, and explicit exploitation of the residual structure), DFO-LS has been used to calibrate increasingly complex models. Applications include global ocean biogeochemistry (for example, MOPS) [55], climate and atmospheric chemistry models [51], among many more [1, 75].

Consider a continuous function $r(\omega)$, where

$$r : \mathbb{R}^n \to \mathbb{R}^m \tag{4.12}$$
$$\omega \mapsto (r_1, ..., r_m)$$

and the constraints $g(\omega) \leq 0$, $h(\omega) = 0$, where

$$g : \mathbb{R}^n \to \mathbb{R}^{m'} \tag{4.13}$$
$$x \mapsto (g_1(\omega), ..., g_{m'}(\omega))$$
$$h : \mathbb{R}^n \to \mathbb{R}^{m''} \tag{4.14}$$
$$x \mapsto (h_1(\omega), ..., h_{m''}(\omega))$$

The DFO-LS method is an iterative algorithm that searches for a local solution for a problem equivalent to the least-squares one for $r(\omega)$ under the constraints $g(\omega) \leq 0$, $h(\omega) = 0$. That is, considering the parameter space

$$U = \{\omega \in \mathbb{R}^n : g(\omega) \leq 0, h(\omega) = 0\} \ , \tag{4.15}$$

DFO-LS iteratively search for a local minimizer for the problem:

$$\min_{\omega \in U} f(\omega) = \sum_{i=1}^{m} [r_i(\omega)]^2. \qquad (4.16)$$

In our setting, the objective f is precisely the least-squares misfit, and feasibility is addressed within the trust-region framework.

DFO-LS is a development of the earlier Derivative-Free Gauss-Newton method (DFO-GN) [7], with enhancements in geometry management and restarts. The DFO-LS outline can be seen as:

1. Choose an initial point $\omega_0 \in \mathbb{R}^n$.

2. Choose n points around ω_0.

3. Iteratively:

    1. Build a quadratic regression model for $f(\omega)$ around ω_k, based on the n+1 points already selected.

    2. Exclude one old point and add another using a trust-region strategy.

    3. Choose the best point in the pool to be ω_{k+1}.

These steps are illustrated in Figure 4.2. Details of each component of the DFO-LS algorithm are presented next.

DFO-LS must be given a starting location within the parameter space to use for initialization. The strategy chosen to increase confidence that DFO-LS had found a good minimizer was to start from multiple initial points. DFO-LS can also be restarted to improve global exploration: *hard restarts* reinitialize the entire interpolation set at new locations; *soft restarts* shift part of the set toward geometry-improving points [55].

4.4.1 Building the quadratic regression model

From a regression standpoint, DFO-LS builds its surrogate by posing a *local least-squares fit* of the residual vector. Given an iteration k, we sample a set of points $Y_k = \{y_1, \ldots, y_p\}$ such that $y_t = \omega_k + s_t$ and $\|s_t\|$ is small, for $t = 1, \ldots, p$, that is, we sample a set of points in a close neighborhood of ω_k. Based on these evaluations, we build the equivalent of a linear regression model for the misfit function gradient (which is not defined, since it is a black-box function) in a neighborhood of ω_k. This approximation is then used to construct the equivalent of a Hessian for the misfit function at ω_k. Based on both of these approximations, DFO-LS builds the equivalent of a second-order Taylor approximation for the misfit function in a neighborhood of ω_k. Minimizing this local quadratic surrogate model for the misfit function defines the trust-region subproblem. This regression view clarifies both efficiency (one model serves all residuals) and robustness to noise (least-squares averaging), while keeping constraint handling natural via the trust-region geometry [7, 55].

Figure 4.2 – Iterative scheme of the DFO-LS algorithm.

Make the mini-linear approximations.

Calculate quadratic approximation and find a minimizer within the trust region.

Eliminate the worst old point, add the new point.
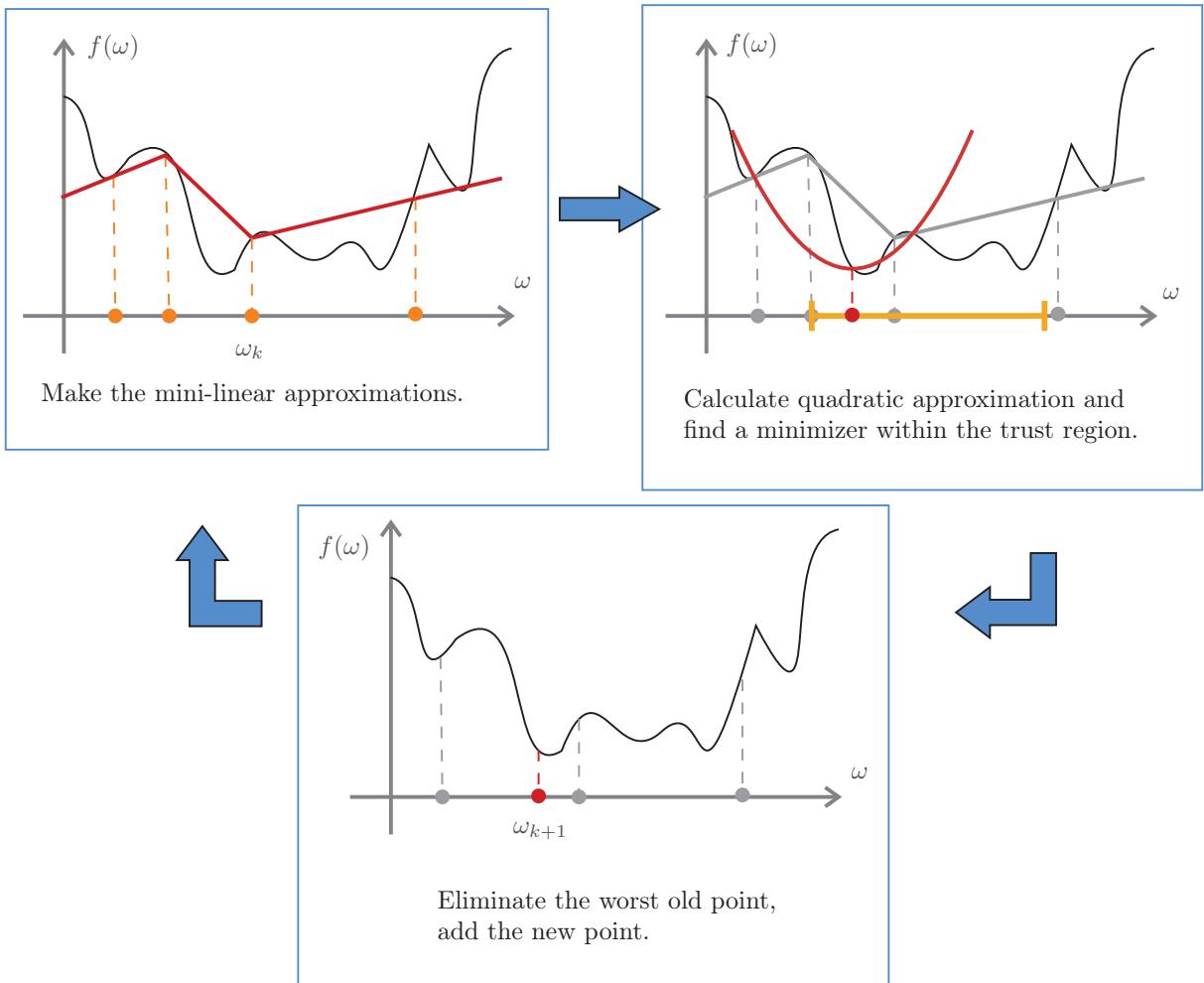
After evaluating the objective function at a set of points, the DFO-LS algorithm iteratively performs the sequence of steps represented by the upper-right panel, upper-left panel, and lower panel, respectively. In all panels, dots represent candidates ω, the background black curve represents the plot of the black-box misfit function $f(\omega)$. In the upper-left panel, the dots represent each element of the candidate pool for the iteration k, where ω_k is the best fit at this iteration, while the red solid line segments represent the information used for building a quadratic local model for $f(\omega)$. In the upper-right panel, the red curve represents the quadratic local model for $f(\omega)$ in a neighborhood of ω_k, while the interval marked by the yellow horizontal line represents the search bounds for its minimizer. The highlighted red point in the lower panel represents the candidate considered the best fit at iteration $k+1$, while a removed gray point corresponds to the least fit at this iteration.

Source: the author (2026).

Consider the functions $r(\omega)$, $f(\omega)$, and the constraints set S as described in (4.12)-(4.16). The DFO-LS method chooses a trust region method to search for a good local solution[1] to the problem (4.16), which is equivalent to finding a solution to:

$$\min_{\omega \in U} \; \frac{1}{2} \, \|r(\omega)\|^2 \; . \tag{4.17}$$

We can notice that when $U = \mathbb{R}^n$, the problem (4.17) is the least squares minimization one, which is an unconstrained minimization problem. Despite this, we will now see how DFO-LS seeks to solve a constrained problem.

Consider given p points[2] in the sample space S, that is, $\{y_1, \ldots, y_p\} \in U$, where $p \in \mathbb{N}$, $p \geq n$, and a point $\omega \in U$ (for simplicity of notation, in this section we are going to write just ω instead of ω^k). First, we define, for convenience, $y_0 := \omega$. We now have a set of $p + 1$ points, $\{y_0, y_1, \ldots, y_p\} \in U$. So we are going to build a linear local model $m_\ell(\cdot)$ for the function $r(\cdot)$ around ω, that is:

$$r(\omega + s) \approx m_\ell(s) := \bar{F} + \bar{J}s \; , \tag{4.18}$$

where $\bar{F} \in \mathbb{R}^m$ and $\bar{J} \in \mathbb{R}^{m \times n}$. The entries of $\bar{F}$ and $\bar{J}$ are found by solving the following regression problem, which is equivalent to an (unconstrained) least squares problem:

$$\min_{\bar{r}, \bar{J}} \sum_{t=0}^{p} \|m_\ell(y_t - \omega) - r(y_t)\|^2 \; . \tag{4.19}$$

Consider the notation:

- $\bar{r}_i$ being the i-th entry of the vector $\bar{F}$;

- $\bar{J}(i, j)$ to the entry at the i-th row and j-th column of the matrix $\bar{J}$;

- $\bar{J}(i, :)$ to the i-th row of the matrix $\bar{J}$;

- $\bar{J}(:, j)$ to the j-th column of the matrix $\bar{J}$.

Proposition 4.2. *Problem (4.19) is equivalent to the linear system of equations:*

$$W \begin{bmatrix} \bar{r}_i \\ \bar{J}(i, :)^T \end{bmatrix} = \begin{bmatrix} r_i(y_0) \\ \vdots \\ r_i(y_p) \end{bmatrix} \; , \tag{4.20}$$

---

[1] A better characterization of a good solution is given in the next section.

[2] Here p is a positive integer, whose value can be set accordingly to different purposes. For example, when seeking exploration in the early iterations, it could be useful to choose $p < n$ for reducing the cost of the algorithm, while for exploitation it could be best to choose $p \geq n$. DFO-LS initialize with $p = n + 1$ as default.

where the matrix $W \in \mathbb{R}^{(p+1) \times (n+1)}$ is defined as

$$W := \begin{bmatrix} 1 & (y_0 - \omega)^T \\ \vdots & \vdots \\ 1 & (y_p - \omega)^T \end{bmatrix} . \tag{4.21}$$

Proof of Proposition 4.2:

Note that, for each $t = 0, 1, \ldots, p$, we have:

$$m_\ell(y_t - \omega) - \mathrm{r}(y_t) = \bar{F} + \bar{J}\,(y_t - \omega) - \mathrm{r}(y_t) = \begin{bmatrix} \bar{F} & \bar{J} \end{bmatrix} \begin{bmatrix} 1 \\ (y_t - \omega) \end{bmatrix} - \mathrm{r}(y_t) . \tag{4.22}$$

From (4.22), we have that the i-th entry of the vector $E_t := (m_\ell(y_t - \omega) - \mathrm{r}(y_t)) \in \mathbb{R}^m$, $i = 1, 2, \ldots, m$, is given by:

$$E_t(i) = \begin{bmatrix} \bar{r}_i & \bar{J}(i, :) \end{bmatrix} \begin{bmatrix} 1 \\ (y_t - \omega) \end{bmatrix} - r_i(y_t) . \tag{4.23}$$

So, by (4.23) we have that:

$$\|m_\ell(y_t - \omega) - \mathrm{r}(y_t)\|^2 = \sum_{i=1}^{m} E_t(i)^2 .$$

And, therefore,

$$\begin{aligned} \sum_{t=0}^{p} \|m_\ell(y_t - \omega) - \mathrm{r}(y_t)\|^2 &= \sum_{t=0}^{p} \sum_{i=1}^{m} E_t(i)^2 \\ &= \sum_{i=1}^{m} \sum_{t=0}^{p} E_t(i)^2 \\ &= \sum_{i=1}^{m} \left\| \begin{bmatrix} 1 & (y_0 - \omega)^T \\ \vdots & \vdots \\ 1 & (y_p - \omega)^T \end{bmatrix} \begin{bmatrix} \bar{F}(i) \\ \bar{J}(i, :)^T \end{bmatrix} - \begin{bmatrix} r_i(y_0) \\ \vdots \\ r_i(y_p) \end{bmatrix} \right\|^2 . \end{aligned}$$

Then, solving the problem (4.19) by least squares method is equivalent to solve m independent problems in the form:

$$\left\| \begin{bmatrix} 1 & (y_0 - \omega)^T \\ \vdots & \vdots \\ 1 & (y_p - \omega)^T \end{bmatrix} \begin{bmatrix} \bar{r}_i \\ \bar{J}(i, :)^T \end{bmatrix} - \begin{bmatrix} r_i(y_0) \\ \vdots \\ r_i(y_p) \end{bmatrix} \right\|^2 ,$$

$i = 1, \ldots, m$ by least squares method. To achieve that, we have to find solutions to the overdetermined linear

63

systems:

$$
W \begin{bmatrix} \bar{r}_i \\ \bar{J}(i,:)^T \end{bmatrix} := \begin{bmatrix} 1 & (y_0 - \omega)^T \\ \vdots & \vdots \\ 1 & (y_p - \omega)^T \end{bmatrix} \begin{bmatrix} \bar{r}_i \\ \bar{J}(i,:)^T \end{bmatrix} = \begin{bmatrix} r_i(y_0) \\ \vdots \\ r_i(y_p) \end{bmatrix}.
$$

$\square$

Proposition 4.3. *The set $\{y_1 - \omega, \cdots, y_p - \omega\}$ spans $\mathbb{R}^n$ iff the matrix W defined in (4.21) has full column rank.*

Proof of Proposition 4.3:

Let the matrices $W' \in \mathbb{R}^{p \times n}$, $0_n \in \mathbb{R}^{1 \times n}$ and $1_p \in \mathbb{R}^{p \times 1}$ be defined as:

$$
W' := \begin{bmatrix} (y_1 - \omega)^T \\ \vdots \\ (y_p - \omega)^T \end{bmatrix}, \quad 0_n := \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}, \quad 1_p := \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.
$$

Then we can view the W as a block matrix:

$$
W = \begin{bmatrix} 1 & 0_n \\ 1_p & W' \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1_p & W'(:,1) & \cdots & W'(:,n) \end{bmatrix}.
$$

($\Rightarrow$) Assuming that $\mathrm{span}\,(\{y_1 - \omega, \cdots, y_p - \omega\}) = \mathbb{R}^n$, we have that $\mathrm{rank}\,(W') = \mathrm{rank}\,(W'^T) = n$. Since $W' \in \mathbb{R}^{p \times n}$, we have that W' has full column rank and so its columns are linearly independent. That implies that the set of vectors:

$$
W_{col} := \left\{ \begin{bmatrix} 1 \\ 1_p \end{bmatrix}, \begin{bmatrix} 0 \\ W'(:,j) \end{bmatrix} : j = 1, \ldots, n \right\}
$$

is linearly independent. That is, the columns of W are linearly independent and so W has full column rank.

($\Leftarrow$) On the other hand, assuming that W has full column rank we have that W_{col} is a linearly independent set. In particular, that implies that the columns of W' are linearly independent, which is equivalent to $rank(\{y_1 - \omega, \cdots, y_p - \omega\}) = n$. Therefore $\mathrm{span}\,(\{y_1 - \omega, \cdots, y_p - \omega\}) = \mathbb{R}^n$.

$\square$

As a consequence of Proposition 4.3, we have the following two corollaries.

Corolary 4.4. *The linear system (4.20) is overdetermined iff the set $\{y_1 - \omega, \cdots, y_p - \omega\}$ spans $\mathbb{R}^n$.*

Corolary 4.5. *When $p < n$, the linear system (4.20) has at least one exact solution.*

So, in the case where $p \geq n$ it is possible that there are no exact solutions to the system (4.20), which justifies using the least-squares method to solve it. On the other hand, when we take $p < n$, there will always be some exact solution to this system, which can reduce the computation costs in the early iterations while providing sufficient exploration of the solution space.

Given these results, the first question is how exactly DFO-LS builds the mini local linear approximations represented by $m_\ell(s)$. The answer, already given, is:

- When $p < n$, we can make use of a direct solver to the linear system (4.20);

- and, when $p \geq n$, we use the least-squares strategy to solve it.

In the second of these cases, we solve the new problem:

$$
W^T \, W \begin{bmatrix} \bar{r}_i \\ \bar{J}(i,:)^T \end{bmatrix} = W^T \begin{bmatrix} r_i(y_0) \\ \vdots \\ r_i(y_p) \end{bmatrix} , \tag{4.24}
$$

which always has at least one solution. Once we have found $\bar{F}$ and $\bar{J}$, we can use them in the model $m_\ell(s)$ described by (4.18), which will be our mini local linear models for $\mathrm{r}(y_t + s)$, $t = 1, \ldots, p$. Considering that this procedure is made at every iteration k of the DFO-LS algorithm, we can see ω mentioned in this section as the iterate ω_k and the model $m_\ell(s)$ as a local model around ω_k, namely $m_\ell^k(s)$. The other points and matrices used in this step are somewhat disposable.

Now that we obtained $m_\ell(s)$, it's time to create the quadratic regression model, $m_q(s) \approx \|\mathrm{r}(\omega + s)\|^2$. This model is created in the most intuitive way possible:

$$
m_q(s) := \|m_\ell(s)\|^2 .
$$

We can notice that, as in classical quasi-Newton methods, one maintains an iterate-dependent matrix $B_k \approx \nabla^2 \phi(\omega_k)$ updated by secant conditions (see Box 4.2).

Box 4.2: Hessian Approximation strategies in Optimization

Given a function $f : \mathbb{R}^n \to \mathbb{R}^n$, computing its Hessian $\nabla^2 f(x)$ directly may be expensive, unavailable, or unstable in the presence of noise. Quasi-Newton methods offer a practical alternative to computing the Hessian matrix explicitly when solving nonlinear optimization problems, as they build an approximation that is iteratively refined as the algorithm progresses.

One of the most widely used quasi-Newton schemes is the **BFGS** (Broyden-Fletcher-Goldfarb-Shanno) update, which constructs a symmetric positive-definite matrix B_k meant to approximate the true

---

Hessian. The update combines curvature information extracted from successive iterates,

$$s_k = x_{k+1} - x_k,$$

and

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k),$$

through the formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{y_k^\top s_k}.$$

This construction ensures that B_{k+1} satisfies the secant condition $B_{k+1} s_k = y_k$, while preserving positive definiteness whenever $y_k^\top s_k > 0$. As a result, BFGS often achieves superlinear convergence without requiring second-order derivatives.

Other quasi-Newton updates follow similar principles but differ in numerical properties and robustness. The **DFP** (Davidon-Fletcher-Powell) update is an earlier alternative with a dual structure to BFGS, while the **SR1** (Symmetric Rank-1) update allows curvature corrections that may violate positive definiteness but can provide useful approximations in problems with highly non-quadratic structure. Limited-memory variants such as **L-BFGS** (Limited-memory BFGS) store only a small number of past curvature pairs, making quasi-Newton methods applicable to large-scale problems.

A thorough introduction to these updates and their theoretical foundations can be found in *Nocedal & Wright, Numerical Optimization* [54].

Here, $B_k = \bar{J}_k^T \bar{J}_k$ arises *implicitly* from the regression fit of residuals and is refreshed whenever the interpolation set is updated. Thus DFO-LS behaves like a Gauss–Newton-type quasi-Newton method whose curvature matrix is built from data-driven Jacobian surrogates rather than explicit secant updates; regularization (for example, $B_k + \lambda I$) can be used if $\bar{J}_k$ is rank-deficient.

Now that we already built the quadratic local model $m_q^k(s)$ around ω_k, which for simplicity we will again be calling $m_q(s)$ and ω, respectively. At this step, we choose a new point to replace an old one in our candidate pool $\{y_0, \ldots, y_p\}$. For simplicity, we will be calling this new point $\bar{x}$. This point will be the next iterate, that is, $\omega_{k+1} := \bar{x}$. We find the point $\bar{x}$ by finding a solution $\bar{s}$ to the problem:

$$\min_{\|s\| \leq \Delta} m_q(s) \,, \tag{4.25}$$

and setting $\bar{x} := \omega + \bar{s}$, where Δ is the trust-region radius chosen for this iteration.

The DFO-LS method steps are shown in Algorithm 1, in the *Appendix*.

4.4.2 The DFO-LS algorithm explained

We now proceed to describe the DFO-LS scheme in detail. In the following, we will consider the notation $\mathbf{\Omega}^k \subset \mathbb{R}^n$ to denote the set of candidates considered in iteration k, that is, $\mathbf{\Omega}^k = \{\omega_0^k, \omega_1^k, \ldots, \omega_p^k\}$. The exception is the first set, denoted $\mathbf{\Omega}^{\mathrm{Ini}}$, which is not associated with any iteration $k = 0, 1, 2, \ldots$, but has the exactly same structure as a candidate pool containing p candidates. A short version of the DFO-LS algorithm we will approach is the following pseudocode:

1. Set the initial candidate pool $\mathbf{\Omega}^{\mathrm{Ini}} = \{\omega_0^{-1}, \ldots, \omega_p^{-1}\}$ (the last candidate does not need to be chosen yet).

   Set $k = -1$.

While the stop criteria is not met :

   Set $k = k + 1$.

2. Choose a candidate $\tilde{\omega}^k$ for replacing the last column of the previous candidate pool, ω_p^{k-1}, which will lead to a new candidate pool $\mathbf{\Omega}^k = \{\omega_0^k, \ldots, \omega_p^k\}$.

   Reorder the columns of $\mathbf{\Omega}^k$ to ensure property (5.1).

3. Create a quadratic local model for $f_{\mathrm{Misfit}}(\omega)$ around ω_0^k.

4. Perform a trust-region strategy using the quadratic model to find a local minimum $\bar{\omega}^k$ for $f_{\mathrm{Misfit}}(\omega)$ around ω_0^k.

If the resulting point is a good enough improvement :

   5. Replace ω_p^k with $\bar{\omega}^k$ in the set $\mathbf{\Omega}^k$.

      Reorder the columns of $\mathbf{\Omega}^k$ to ensure property (5.1).

In the following, we provide some details about this algorithm, which are especially useful when calibrating a biogeochemical model. The DFO-LS algorithm ensures the convergence (or else the termination or restart of the algorithm executing) of a sequence of points $\{\omega^k\}$ in the parameter space to a calibrated parameter $\bar{\omega}$ in the parameter space; such sequence is represented by $\{\omega_0^k\}$. Although for $k \geq 1$ all the points ω_0^k are generated by the algorithm, the first point ω_0^0 is required for initializing the algorithm. Here, we always choose to set ω_0^0 as the best parameter calibration we have at the moment, but any point in the parameter space is also a possible choice. All the columns of $\mathbf{\Omega}^{\mathrm{Ini}}$ in step 1, except for the first one, may be chosen randomly; this also applies to the choice of $\tilde{\omega}^k$ in step 2.

4.4.3 Creating the quadratic model

When creating a quadratic local model for the misfit function in step 3 of the pseudocode presented in the previous subsection, we follow the procedure described by *C. Cartis et al. (2019, p.32:6)* [7]. We first

create a linear local model for the residual function $r_{\text{Misfit}}(\omega)$ around ω_0^k, which can be seen as mimicking or predicting the behavior of the first order Taylor approximation for the residual function around ω_0^k in situations where its first derivatives are available. This is a regression model using data from the columns of $\mathbf{\Omega}^k$. Therefore, when defining the local linear models for the residual function, we will be considering the following notation:

s :   A step (that is, a vector) in the parameter space.

$m_\ell^k(s)$:   Linear local model for the residual function $r_{\text{Misfit}}(\omega)$ around the point ω_0^k. See (4.26).

$m_q^k(s)$:   Quadratic local model for the misfit function $f_{\text{Misfit}}(\omega)$ around the point ω_0^k. See (4.30).

For the linear local model of the residual function, we have:

$$m_\ell^k(s) \approx r_{\text{Misfit}}(\omega_0^k + s) \ , \tag{4.26}$$

for every s sufficiently small. We take the following notation:

r_k :   Approximation for $r_{\text{Misfit}}(\omega_0^k)$.

J_k :   Approximation for the Jacobian matrix of $r_{\text{Misfit}}(\omega)$ at ω_0^k.

Thus, in our linear local model, we will use a linear regression strategy for defining r_k and J_k at each iteration and set:

$$m_\ell^k(s) = r_k + J_k s \ . \tag{4.27}$$

For defining r_k and J_k, we solve the following linear regression problem about the residual of the misfit function evaluated over the candidate pool:

$$\min_{r_k, J_k} \sum_{i=0}^{\lambda} \|m_\ell^k(\omega_i^k - \omega_0^k) - r_{\text{Misfit}}(\omega_i^k)\|^2 \ . \tag{4.28}$$

Solving problem (4.28) is equivalent to finding a solution for the linear system

$$W_k \begin{bmatrix} r_{k,i} \\ J_{k,i}^T \end{bmatrix} = b_{k,i} \ , \ \forall i = 1, 2, \ldots, N \ , \tag{4.29}$$

where

$$W_k = \begin{bmatrix} 1 & (\omega_0^k - \omega_0^k) \\ 1 & (\omega_1^k - \omega_0^k) \\ \vdots & \vdots \\ 1 & (\omega_p^k - \omega_0^k) \end{bmatrix} \text{, and } b_{k,i} = \begin{bmatrix} r(\omega_0^k) \\ r(\omega_1^k) \\ \vdots \\ r(\omega_p^k) \end{bmatrix} \text{ for all } i = 1, \dots, N \text{ ,}$$

$$r_k = \begin{bmatrix} r_{k,1} \\ r_{k,2} \\ \vdots \\ r_{k,N} \end{bmatrix}_N \text{, and } J_k = \begin{bmatrix} J_{k,1} & J_{k,2} & \cdots & J_{k,N} \end{bmatrix}_{n \times N} .$$

For the quadratic local model of the misfit function, we have:

$$m_q^k(s) \approx f_{\text{Misfit}}(\omega_0^k + s) \text{ ,} \tag{4.30}$$

for every s sufficiently small. We take the following notation:

f_k :   Approximation for $f_{\text{Misfit}}(\omega_0^k)$.

g_k :   Approximation for the gradient vector of $f_{\text{Misfit}}(\omega)$ at ω_0^k.

H_k :   Approximation for the Hessian matrix of $f_{\text{Misfit}}(\omega)$ at ω_0^k.

We set our quadratic model as:

$$m_q^k(s) = f_k + g_k^T s + \frac{1}{2} s^T H_k s \text{ ,} \tag{4.31}$$

where we derive f_k, g_k and H_k from some operations with r_k and J_k obtained for the linear model $m_\ell^k(s)$. They are defined as:

$$f_k = r_k^T r_k \text{ , } g_k = 2 J_k^T r_k \text{ , and } H_k = 2 J_k^T J_k \text{ .}$$

4.4.4   Minimizing the quadratic model with a trust-region strategy

At the iteration k, we want to find a minimizer step s for the following constrained optimization problem:

$$\min m_q^k(s)$$
$$\text{Subject to: } s \in \Gamma^k \tag{4.32}$$

where the set of constraints, namely the trust region, is defined as:

$$\Gamma^k = \left[\Delta_{\ell 1}^k, \Delta_{u1}^k\right] \times \left[\Delta_{\ell 2}^k, \Delta_{u2}^k\right] \times \cdots \times \left[\Delta_{\ell n}^k, \Delta_{un}^k\right] \ , \tag{4.33}$$

for some $\Delta_{\ell i}^k, \Delta_{ui}^k, \ i = 1, \ldots, n$, satisfying the condition:

$$\Delta_{\ell i}^k < \Delta_{ui}^k, \ \forall i = 1, \ldots, n \ .$$

If we consider the notations $s = (s_1, \ldots, s_n)$, note that the constraints for problem (4.32) are equivalent to:

$$\Delta_{\ell i}^k \leq s_i < \Delta_{ui}^k, \ \forall i = 1, \ldots, n \ . \tag{4.34}$$

The set of inequalities (4.34) characterizes each interval $\left[\Delta_{\ell i}^k, \Delta_{ui}^k\right]$ as search bounds for s_i, for $i = 1, 2, \ldots, n$.

For examples on the basic usage of DFO-LS for parameter calibration on Python, see the *Appendix*. The DFO-LS algorithm searches for the vector of parameters that minimizes the misfit function, leading to what we are calling the calibration of such parameters. The output obtained by DFO-LS function consists of convergence info, including the vector of parameters found by this method and the number of misfit function evaluations. All of these will become clearer with the examples presented in the next chapter.

In this chapter, we explored how to frame calibration as an optimization task, why we use a least-squares misfit, and how derivative-free methods such as DFO-LS approaches black-box, noisy, and constrained problems. Building on that, the next chapter outlines a step-by-step approach for calibrating calibrating models and test the approach in practice with few examples.

5 A systematic approach for calibrating conceptual models

The aim of chapter is to describe in detail how the optimization tools introduced in *Chapter 4* can be used in practice for systematically calibrating models, with a focus on conceptual models based on systems of ordinary differential equations (ODEs).

In general, the success of a calibration strategy for the parameters of a mathematical model using optimization methods depends on three main ingredients:

- A sufficiently accurate mathematical model for the phenomenon of interest;

- An adequate framing of the fitting as an optimization problem;

- A proper strategy for solving the optimization problem.

To illustrate and test this approach in practice, we consider a sequence of increasingly complex test problems that resemble the structure and dimension of conceptual models such as the PEC model. We focus on the appropriate choice of residual function, initial guess, and bounds. Then, we present a calibration case study, in which we design and analyze several numerical experiments built around a small set of representative ODE models. In all the computational experiments presented in this section, we used the DFO-LS optimization Python solver [64].

Throughout this chapter, the term *parameters* denotes the unknown quantities to be calibrated, *observational parameters* denotes the reference values we aim to recover via calibration, *model output* denotes the dataset produced by the model when setting a particular choice of parameters, and *observations* as the dataset generated for reference, which is used for recovering the observational parameters.

Chapter notation and DFO-LS setting: In the following, we recall the basic DFO-LS setting in a matrix-based notation that is close to the Python implementation used later. We do not consider advanced options related to restarts or noise on the DFO-LS solver, since the basic usage and the advanced options related to bounds are sufficient for the examples presented in this chapter.

At an iteration k, DFO-LS builds and updates a *candidate pool* of parameter vectors. We represent this pool by the matrix:

$$\mathbf{\Omega}^k = \begin{bmatrix} \omega_0^k & \omega_1^k & \cdots & \omega_p^k \end{bmatrix}_{n \times (p+1)} ,$$

where each column $\omega_j^k \in \mathbb{R}^n$ is a candidate parameter vector (that is, a full set of values for the n parameters being calibrated), and $p+1$ is a fixed number of candidates considered at each iteration k.

For each candidate ω_j^k, we evaluate the misfit function:

$$f_{\text{Misfit}} : \mathbb{R}^n \to \mathbb{R}_+ ,$$

and collect the corresponding values into the vector:

$$f_{\text{Misfit}} \left(\mathbf{\Omega}^k \right) = \begin{bmatrix} f_{\text{Misfit}}(\omega_0^k) \\ f_{\text{Misfit}}(\omega_1^k) \\ \vdots \\ f_{\text{Misfit}}(\omega_p^k) \end{bmatrix}_{p+1} .$$

We order the candidates so that:

$$f_{\text{Misfit}}(\omega_j^k) \geq f_{\text{Misfit}}(\omega_{j-1}^k) \geq 0 , \ \forall j = 1, 2, \ldots, p . \tag{5.1}$$

Thus, ω_0^k is the current best candidate at iteration k.

The misfit function itself is defined as the squared norm of a residual misfit function $r_{\text{Misfit}} : \mathbb{R}^n \to \mathbb{R}^m$,

$$f_{\text{Misfit}}(\omega) = \|r_{\text{Misfit}}(\omega)\|^2 ,$$

where m is the number of scalar residual entries of the residual function (typically related to the number of data points considered during its definition).

At iteration k, the residual misfit function evaluated at each candidate is represented by the matrix:

$$r_{\text{Misfit}} \left(\mathbf{\Omega}^k \right) = \begin{bmatrix} r_{\text{Misfit}}(\omega_0^k) & r_{\text{Misfit}}(\omega_1^k) & \cdots & r_{\text{Misfit}}(\omega_p^k) \end{bmatrix}_{m \times (p+1)} ,$$

whose columns are the m-dimensional residual vectors $r_{\text{Misfit}}(\omega_j^k)$.

The DFO-LS algorithm builds, at each iteration, a local quadratic model for f_{Misfit} based on these residual evaluations and solves a trust-region subproblem to obtain the next candidate. The details of this process are discussed in *Chapter 4*; here we focus on how to construct r_{Misfit} in a way that is meaningful for our models and data. The following example illustrates one possible way of defining r_{Misfit} for a conceptual

biogeochemical model.

Example 5.1. *Consider the two-box NP model for the Paranaguá Estuarine Complex presented in Chapter 3 (Equations (3.8)-(3.10)), where the observations for the monthly mean concentration of two different tracers in the upper box, nitrate C_N and phytoplankton C_P, are available for 12 different times, $t_1, t_2, \ldots, t_{12}$, summing up to the total of $m = 24$ observations. These observations will be denoted $N_{obs}(t_i)$ and $P_{obs}(t_i)$, respectively. Let us also denote the concentrations of nitrate and phytoplankton, respectively, predicted by our model for each time t_i and depending on the choice of parameters ω, as $N_{predicted}(\omega, t_i)$ and $P_{predicted}(\omega, t_i)$. In this example, we denote the entries of the residual function as*

$$r_{Misfit}(\omega) = \big(r_{1,Misfit}(\omega), \ r_{2,Misfit}(\omega), \ \ldots, \ r_{m,Misfit(\omega)}\big) \ ,$$

and set

$$r_{i,Misfit}(\omega) \ = \begin{cases} \dfrac{N_{predicted}(\omega, t_i) - N_{obs}(t_i)}{24 \cdot N_{obs}(t_i)}, \ \forall i = 1, 2, \ldots, 12, \\[4mm] \dfrac{P_{predicted}(\omega, t_{i-12}) - P_{obs}(t_{i-12})}{24 \cdot P_{obs}(t_{i-12})}, \ \forall i = 13, 14, \ldots, 24. \end{cases}$$

Thus, the misfit function is defined as:

$$f_{Misfit}(\omega) = \sum_{i=1}^{24} \left[r_{i,Misfit}(\omega)\right]^2 = \|r_{Misfit}(\omega)\|^2 \ .$$

Example 5.1 will serve as a template for the construction of residual functions in the subsequent sections, where we will adapt it to simpler ODE models that are easier to analyze in detail.

In the following section, we perform a sequence of tests designed to clarify the role of the residual function, initial guess, and bounds in parameter calibration, starting with a very simple model. We focus on ODE models that share some key characteristics with conceptual models such as the PEC model:

- The model is defined by a system of ODEs' equilibrium, which is usually periodic.

- The parameters are calibrated based on data from observations.

- Each calibrated parameter is constrained by predefined lower and upper bounds.

In what follows we briefly recall the concepts of fixed point and limit cycle, then introduce a first test problem with a scalar ODE.

A **fixed point** (also called an *equilibrium point* or *steady state*) of a dynamical system described by a set of ordinary differential equations (ODEs)

$$\dot{x} = F(x),$$

is a point $x^* \in \mathbb{R}^n$ such that

$$F(x^*) = 0.$$

At this point, the system does not change in time: if the system starts at x^*, it remains there for all t. A stable fixed point acts as an *attractor*: trajectories starting near it move toward and remain near that equilibrium, as fixed points represent time-invariant solutions.

On the other hand, a **limit cycle** is a closed, isolated periodic orbit. Formally, it is a trajectory $\gamma(t)$ satisfying

$$\gamma(t + T) = \gamma(t),$$

for some period $T > 0$. We say that a limit cycle is *stable* when trajectories starting near the cycle converge to $\gamma(t)$ as $t \to \infty$.

The use of observations for parameter calibration is more clearly illustrated in the following example.

Example 5.2 (A constant equilibrium)**.** *Let $a, b > 0$ be constants. Consider the linear ODE*

$$\frac{dx}{dt} = a - bx \ . \tag{5.2}$$

The equilibrium solution is $\bar{x} = \dfrac{a}{b}$. Every solution of (5.2) converges to the equilibrium, meaning that, when integrating this ODE system for a sufficiently large time interval, we obtain:

$$x(t) \approx \bar{x} \ .$$

We use this simple model to illustrate the basic DFO-LS calibration scheme. We fix a value $a > 0$ and an observational equilibrium $\bar{x} > 0$, and we use DFO-LS to calibrate the parameter b. Since $\bar{x} = a/b$, we expect b to converge to $a/\bar{x}$. Equivalently, we can think of the observational parameter vector as:

$$[a_{Obs}, b_{Obs}] = [2, 10] \ ,$$

so that $\bar{x} = a_{Obs}/b_{Obs} = 0.2$, and the calibration aims to recover b_{Obs}.

Because this model has a single state variable, the residual can be defined as a direct comparison between the model equilibrium $\bar{x}$ (considered as observational data) and one sampled value from the system integration:

$$r_{Misfit}(b) = x(b) - \bar{x},$$

where $x_{(}b)$ denotes the numerical solution obtained by integrating system (5.2) and sampling the last integra-

tion point. The misfit function is then

$$f_{Misfit}(b) = \left[r_{Misfit}(b) \right]^2.$$

For generating the model output to be compared with the observational data, we performed the numerical integration with:

- t_{max} *large enough for the solution to be close to equilibrium,*

- *a fixed time step* Δt,

- *an initial condition for the integration,* $x(0) = x_0$.

When varying a, b and $\bar{x}$, it is important to verify that the choices of t_{max}, Δt and x_0 remain appropriate. We then choose an initial guess b_0 for the parameter calibration and run DFO-LS. The Python implementation and the detailed settings for this test are given in the Appendix. The DFO-LS search for b is shown in Figure 5.1.

Figure 5.1 – DFO-LS calibration of parameter b in Example 5.2.
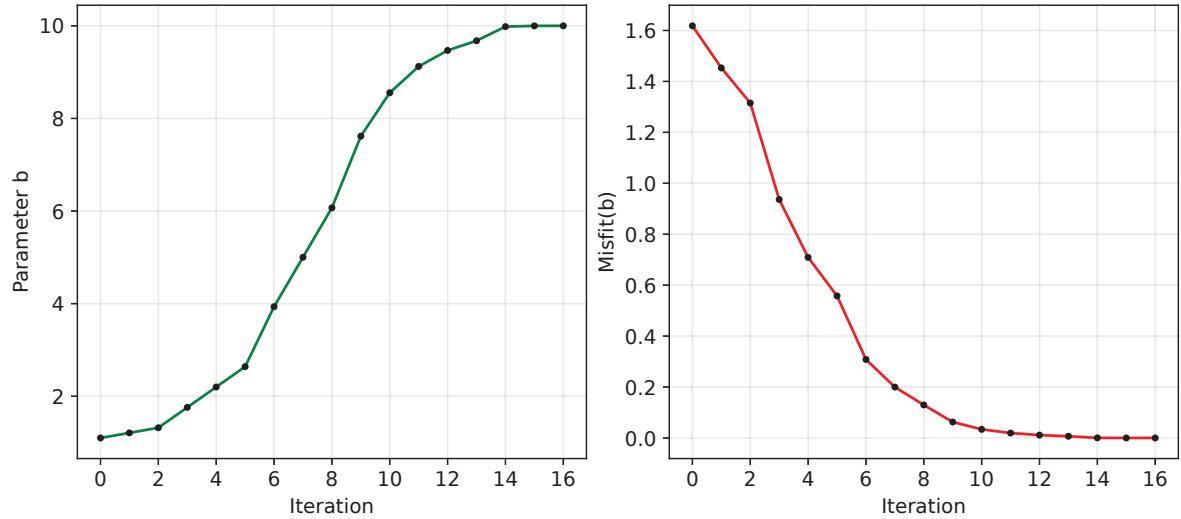
In the left-hand plot, the points represent the values of b obtained at each iteration. In the right-hand plot, the points represent the corresponding misfit values $f_{\mathrm{Misfit}}(b)$ attained at each iteration. The observational parameter $b_{\mathrm{Obs}} = 10$ is recovered in 16 iterations, with $f_{\mathrm{Misfit}}(b_{\mathrm{Obs}}) = 0$.
Source: the author (2026).

This example shows that even a minimal residual definition can be sufficient for parameter calibration when the dynamics are simple and the equilibrium is well-behaved. In the next section, we extend this

framework to nonlinear systems with limit cycles and more parameters.

5.1 Calibration case study

The previous scalar example illustrates the basic mechanics of parameter calibration with DFO-LS. We now consider a case study based on ODE systems with cyclic or nonlinear behavior, designed to emulate some key features of conceptual biogeochemical models:

- the presence of limit cycles (periodic equilibria),

- the simultaneous calibration of several parameters,

- possible ill-conditioning of the misfit function,

- the need for model-based choices of residuals, observation sets, and bounds.

The case study is organized into four classes of models:

1. An ODE system with a cyclic equilibrium (Part 1): calibration of a single parameter in a two-dimensional system with a known limit cycle;

2. An ODE system with a cyclic equilibrium (Part 2): simultaneous calibration of two parameters in a similar system;

3. A difficult case: a sinusoidal system where different parameter values produce indistinguishable observations, illustrating limitations of systematic calibration;

4. A nonlinear example: a two-dimensional nonlinear system (Sel'kov model) with a limit cycle and no closed-form solution.

In each subsection we follow the same structure:

- *Model and observations:* definition of the ODE system and the way observational data are generated.

- *Residual definition:* construction of r_{Misfit} from samples and observations.

- *Numerical experiment:* a concrete experiment (or a family of experiments) based on the same example, varying the number and distribution of observations, the length of the integration window, and/or other design parameters in order to investigate calibration performance.

In all tests, the ultimate goal is to recover the chosen observational parameters from the synthetic observational data, which was generated by the model itself.

5.1.1 An ODE system with cyclic equilibrium – Part 1

Model and observations: We first consider a two-dimensional ODE system with a periodic equilibrium (limit cycle) and a single parameter to be calibrated. Let

$$
\begin{cases}
x' & = -y - \left(\dfrac{(t+c)^{-2}}{e^a + (t+c)^{-1}} \right) + \ln \left(e^a + (t+c)^{-1} \right) + y_0 \;, \\[3mm]
y' & = x - \left(\dfrac{(t+c)^{-2}}{e^a + (t+c)^{-1}} \right) + \ln \left(e^a + (t+c)^{-1} \right) - x_0 \;,
\end{cases}
\tag{5.3}
$$

with fixed constants $c > 0$, x_0, y_0 and parameter $a \in \mathbb{R}$. System (5.3) is well-defined for any $t > 0$, and the analytical solution of the system (5.6) is given by:

$$
\begin{cases}
x(t) & = \ln \left(e^a + (t+c)^{-1} \right) + \cos(t) + x_0 \;, \\[3mm]
y(t) & = \ln \left(e^a + (t+c)^{-1} \right) + \sin(t) + y_0 \;.
\end{cases}
\tag{5.4}
$$

When $x_0 = y_0 = 0$, and for $t >> 0$ (that is, t very large), we obtain:

$$
\ln \left(e^a + (t+c)^{-1} \right) \approx a \;,
$$

so that

$$
(x(t), y(t)) \approx (a + \cos(t), a + \sin(t)) \;,
\tag{5.5}
$$

which corresponds to the convergence for a periodic equilibrium (limit cycle).

We fix $c > 0$ and address the problem of calibrating the parameter a, which can assume any real value. We define an observational parameter a_{Obs} and generate synthetic observations from the asymptotic expression (5.5), that is,

$$
(x_{\mathrm{Obs}}(t), y_{\mathrm{Obs}}(t)) = (a_{\mathrm{Obs}} + \cos t, \; a_{\mathrm{Obs}} + \sin t) \;.
$$

To control the sampling along the limit cycle, we introduce the rescaled time variable $s = \pi t$, so that $(x(s), y(s))$ has period 2 in s instead of period 2π in t. All observation times are then chosen in a final time window corresponding to one or more periods after a sufficiently long integration time, ensuring that the numerical solution is close to the limit cycle.

To control the sampling along the limit cycle, we introduce the rescaled time variable $s = \pi t$, so that $(x(s), y(s))$ has period 2 in s instead of period 2π in t. All observation times are then chosen in a final time window corresponding to one or more periods after a *spinup* (a sufficiently long integration period used to stabilize the model, see Box 3.3), ensuring that the numerical solution is close to the limit cycle.

Misfit residual definition: We consider $N+1$ sample points along the last cycle (after the spinup). For each observation time $t \in T_{\text{Obs}}$ we compute both the model output $(x(t), y(t))$ and the observational state $(x_{\text{Obs}}(t), y_{\text{Obs}}(t))$. The residual vector is defined by collecting pointwise Euclidean distances,

$$r_{\text{Misfit}}(a) = (r_1(a), \ldots, r_{N+1}(a)) \ ,$$

with

$$r_i(a) = \|(x(t_i; a), y(t_i; a)) - (x_{\text{Obs}}(t_i), y_{\text{Obs}}(t_i))\| \ .$$

The misfit function is

$$f_{\text{Misfit}}(a) = \sum_{i=1}^{N+1} r_i(a)^2 \ .$$

Figure 5.2 illustrates the pointwise fitting on the limit cycle: the red curve is the observation set, the green curve is a model trajectory, and the points correspond to $N+1$ samples used in the residual. In the illustrated case, the time interval chosen for the model integration, $t \in [1, 398]$, was not long enough to be considered a spinup of the model. This situation highlights the need for caution when defining the spinup time for the model, as it may lead to accuracy loss when attempting to recover the observational parameters.

Numerical experiment (single-parameter limit-cycle calibration):

Experiment 1. We now design a specific experiment based on this example. The goal is to recover the parameter a_{Obs} from synthetic observations using DFO-LS and to study the influence of:

- the integration time (length of spinup),

- the number of samples $N+1$,

- the need for a heuristic step when defining the initial guess for the parameter search.

We proceed as follows:

1. Fix $c > 0$ (in this experiment, we set $c = 100$), choose a_{Obs}, and generate observations via (5.5).

2. Integrate (5.3) with an initial condition away from the limit cycle, for a sufficiently long time.

3. In a first set of runs, use a heuristic stage: evaluate f_{Misfit} on a grid of a values in a neighborhood of a_{Obs} and select the best value as initial guess for DFO-LS.

4. In a second set of runs, bypass the heuristic and start directly from a generic guess further away from a_{Obs}.

Figure 5.2 – Illustration of the pointwise fitting process on the limit cycle.
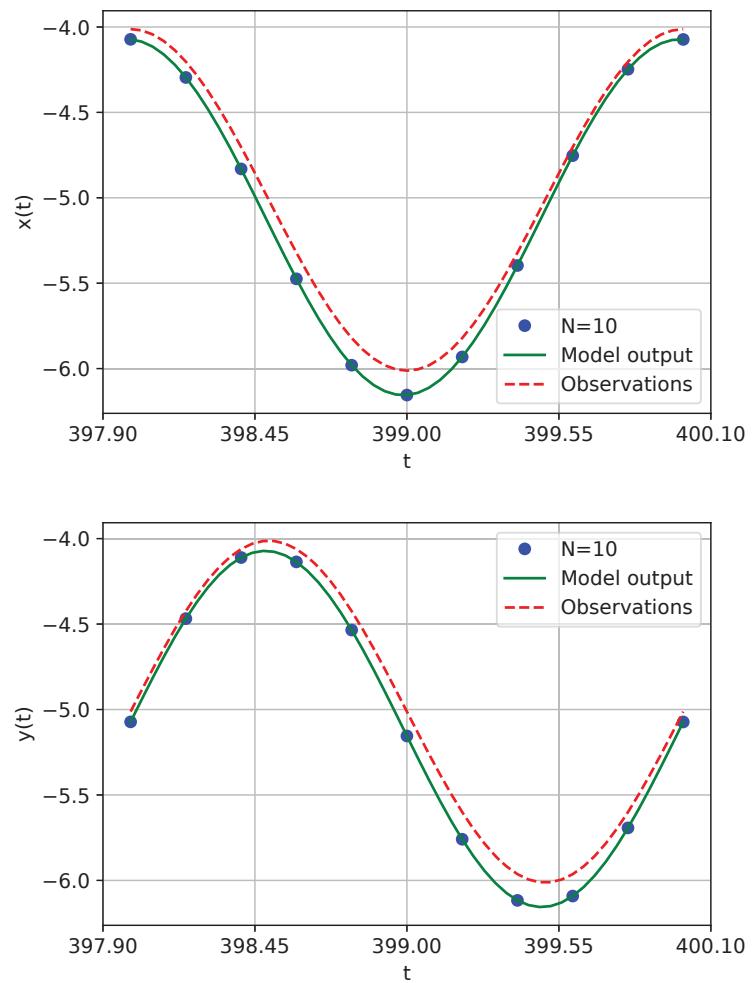
The red dashed line represents the observational data, while the green solid line represents the model output. Dots represent a set of N sampled points used in computing the misfit residual.

**Source:** the author (2026).

5. In both cases, define bounds as a fixed-size neighborhood of a_{Obs}. Note that the bounds are informed to the solver without any information about the observational parameter.

Figure 5.3 shows a typical heuristic evaluation for $a_{\text{Obs}} = 5.01$. The misfit displays a nearly linear trend around the observational parameter, and the best point found by the heuristic is close to a_{Obs}.

Figure 5.3 – Heuristic evaluation of $f_{\textbf{Misfit}}(a)$ around $a_{\textbf{Obs}} = 4.77$.
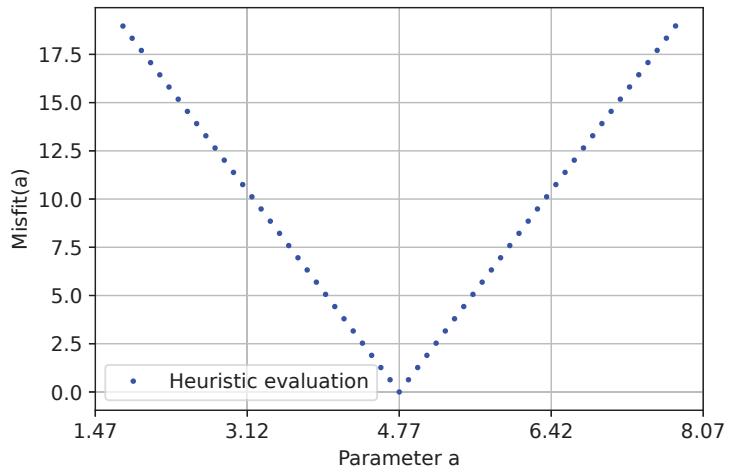
The seemingly linear behavior of the misfit function suggests that DFO-LS can converge to the observational parameter if given an initial guess from a relatively broad neighborhood of a_{Obs}, provided that the integration time is long enough.
Source: the author (2026).

Figure 5.4 shows typical fittings in x and y for the observational parameter $a_{\text{Obs}} = 5.1$ and two different numbers of observations $N + 1$. In all tested settings, the observational parameter is recovered with good accuracy when the integration time is sufficiently long.

Figures 5.5 and 5.6 summarize the dependence of the calibration accuracy on the number of observations and on the integration time. Due to the convergence to the limit cycle being fast, relatively short spinup time definitions were sufficient. For the experiments performed in this subsection, convergence to the limit cycle was relatively fast, and accurate recovery is possible over a broad range of N and integration windows.

Figure 5.4 – **Fitting of sampled model output values to the corresponding observations.**
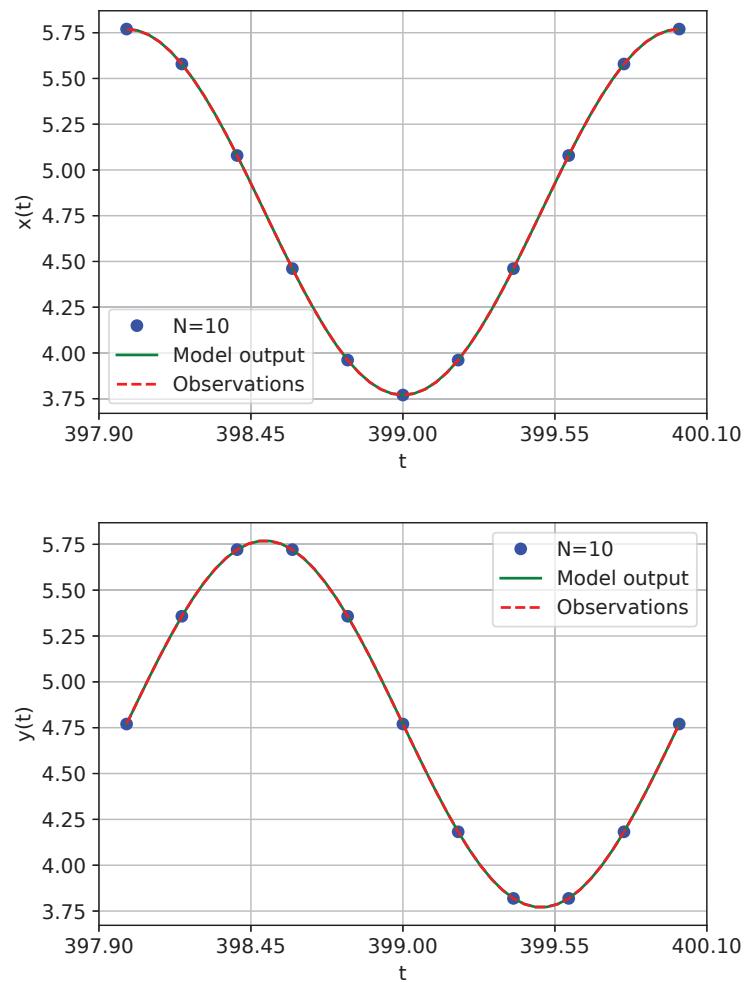
Considering the model defined by the equilibrium of equations (5.3), for $a_{\mathrm{Obs}} = 4.77$ and $N+1 = 11$, following the structure in Figure 5.2.

Source: the author (2026).

Figure 5.5 – Dependence on the number of observations $N + 1$ for the calibration accuracy of the parameter a.
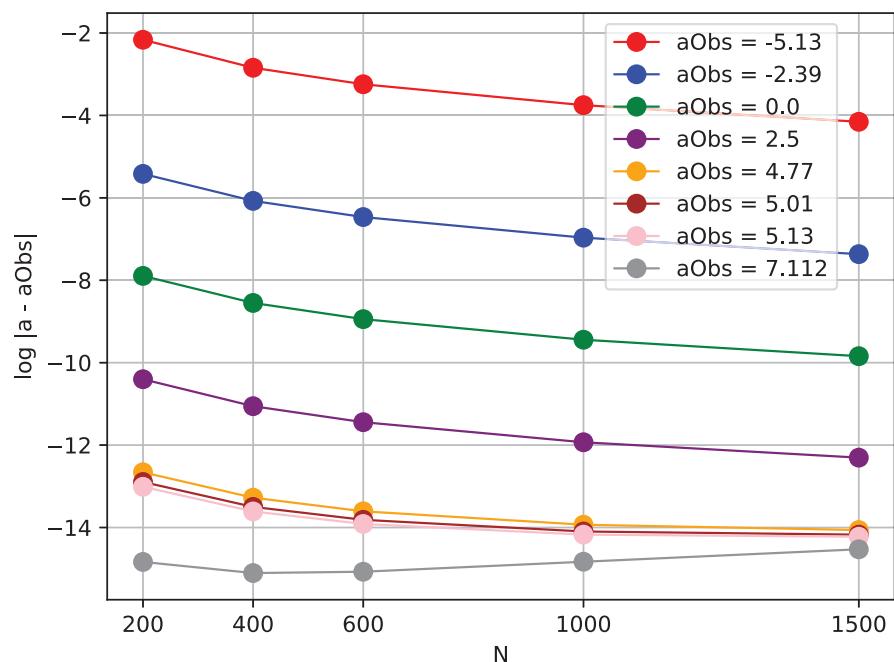
The dots represent the accuracy attained for the calibration experiment defined in Subsection 5.1.1 with the corresponding observational parameter a_{Obs}.

Source: the author (2026).

Figure 5.6 – Dependence of calibration accuracy for a on the integration time.
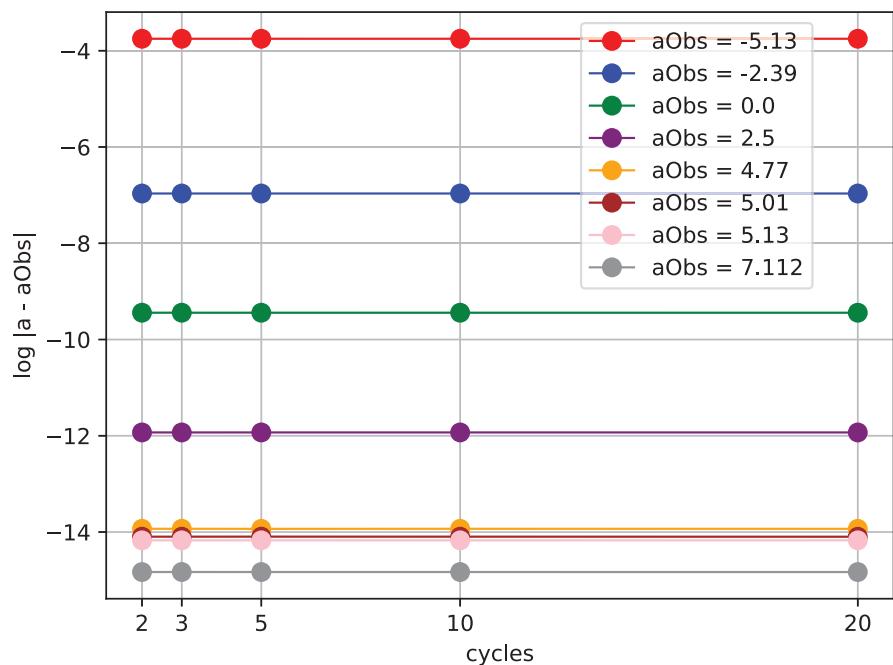
The length of spinup was considered as approximately $[1, T_{cycles} \cdot 2]$, where T_{cycles} is the number of cycles considered, each cycle with length equal to 2. The dots represent the accuracy attained for the calibration experiment defined in Subsection 5.1.1 with the corresponding observational parameter a_{Obs}.

Source: the author (2026).

5.1.2 An ODE system with cyclic equilibrium – Part 2

We now generalize the previous mathematical model in order to calibrate two parameters simultaneously.

Model and observations: Consider the ODE system:

$$\begin{cases} x' & = -y - \left(\dfrac{(t+c)^{-2}}{e^a + (t+c)^{-1}} \right) + \ln \left(e^b + (t+c)^{-1} \right) + y_0 \;, \\[4mm] y' & = x - \left(\dfrac{(t+c)^{-2}}{e^b + (t+c)^{-1}} \right) + \ln \left(e^a + (t+c)^{-1} \right) - x_0 \;, \end{cases} \tag{5.6}$$

with fixed $c > 0$, constants x_0, y_0 and parameters a, b. System (5.6) can be seen as a generalization of system (5.3) and is well-defined for any $t > 0$. The analytical solution of the system (5.6) is:

$$\begin{cases} x(t) & = \ln \left(e^a + (t+c)^{-1} \right) + \cos(t) + x_0 \;, \\[4mm] y(t) & = \ln \left(e^b + (t+c)^{-1} \right) + \sin(t) + y_0 \;. \end{cases} \tag{5.7}$$

For $x_0 = y_0 = 0$ and $t \gg 0$, we obtain:

$$\ln \left(e^a + (t+c)^{-1} \right) \approx a \;,$$
$$\text{and} \quad \ln \left(e^b + (t+c)^{-1} \right) \approx b \;,$$

so that

$$(x(t), y(t)) \approx (a + \cos(t), b + \sin(t)) \;, \tag{5.8}$$

which is again a limit cycle.

We fix $c > 0$ and address the problem of calibrating parameters $a, b > 0$. Given observational parameters $a_{\text{Obs}}, b_{\text{Obs}}$, we generate synthetic observations as:

$$(x_{\text{Obs}}(t), y_{\text{Obs}}(t)) = (a_{\text{Obs}} + \cos t, \; b_{\text{Obs}} + \sin t) \;,$$

at fixed time instants chosen as in Part 1 (subsection 5.1.1): after a spinup, over one or more cycles, with $s = \pi t$ for convenience.

Misfit residual definition: As in Part 1, we consider $N+1$ points along the last cycle and define the residual vector by pointwise distances between the model trajectory $(x(t), y(t))$ and the observational trajectory

$(x_{\mathrm{Obs}}(t), y_{\mathrm{Obs}}(t))$. The misfit function is

$$f_{\mathrm{Misfit}}(a, b) = \sum_{i=1}^{N+1} r_i(a, b)^2 \ ,$$

with r_i defined in the same way as before.

Numerical experiment (two-parameter limit-cycle calibration):

Experiment 2. In this experiment, we apply DFO-LS to simultaneously recover the parameters a_{Obs} and b_{Obs}, using a similar setup as in Part 1 except that here we set $c = 10^{-4}$ and now we consider a two-dimensional parameter space:

1. We set observational parameters $[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}]$ and generate observations from (5.8).

2. We integrate system (5.6) from an initial point away from the limit cycle and record samples after a spinup.

3. We perform a heuristic step where we evaluate $f_{\mathrm{Misfit}}(a, b)$ on a grid of parameter values in a rectangular neighborhood around $[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}]$ and select the best point as the initial guess for DFO-LS.

4. We set bounds as a smaller rectangle centered on the initial guess and then run DFO-LS.

Figure 5.7 shows contour plots of the heuristic evaluation of the misfit function for observational parameters $[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}] = [2.5, 3.5]$ and two choices for the number of observation, $N + 1$. In both cases the observational parameters appear as the global minimizer among the sampled points.

Figure 5.8 shows the resulting fitting of the predicted values to the observations for x and y, again using the sampling structure of Figure 5.2. The behavior is qualitatively similar to the one-parameter case, with successful recovery of $a_{\mathrm{Obs}}, b_{\mathrm{Obs}}$ over a range of settings.

5.1.3 A difficult case: when systematic calibration may fail

We now consider a model that illustrates intrinsic limitations of systematic calibration when the parameter is non-identifiable from the observations, that is, there is no way to ensure the recovering of observational parameters by fitting model outputs to observations.

Model and observations: Consider the ODE system:

$$\begin{cases} x' & = ay \ , \\ y' & = -ax \ . \end{cases} \tag{5.9}$$

Figure 5.7 – Heuristic contour plot of $f_{\mathrm{Misfit}}(a, b)$.
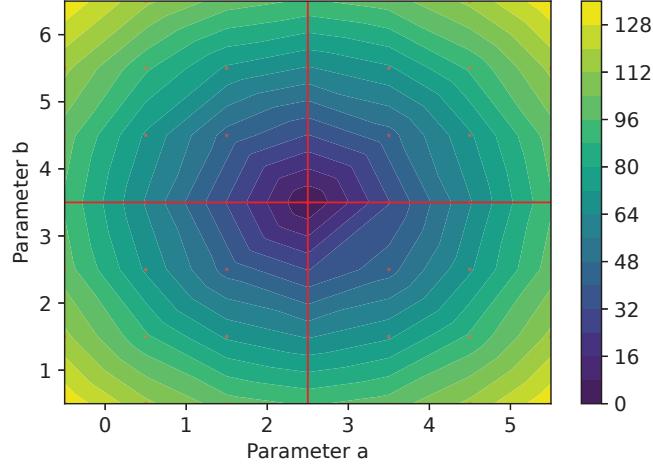
Observational parameters are $[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}] = [2.5, 3.5]$, considering $N = 10$ observations, and function evalua-tion at 49 parameter pairs. The highlighted lines intersect at the location of the observational parameters.
Source: the author (2026).

The non-trivial solution of system (5.9) is given by:

$$
\begin{cases}
x(t) & = \sin(at) \ , \\
y(t) & = \cos(at) \ .
\end{cases}
\tag{5.10}
$$

Let a_{Obs} be the observational parameter. If suitable bounds were not set for the search of parameter a, it may become impossible to recover a_{Obs} via calibration. In this example, for any $a_{\mathrm{Obs}} \in \mathbb{R}$, we see that:

$$
\begin{cases}
x(t) & = \sin(\bar{a}t) \ , \\
y(t) & = \cos(\bar{a}t) \ ,
\end{cases}
\tag{5.11}
$$

whenever $\bar{a} = a_{\mathrm{Obs}} + 2\pi \cdot k$, for any $k \in \mathbb{Z}$. Thus, in this case, it is important to ensure that the length of the search range for the parameter a is less than 2π.

To remove this ambiguity, we define a constrained optimization problem as

$$
\text{minimize} \quad f_{\mathrm{Misfit}}(a)
$$
$$
\text{subject to:} \quad a \ \in [0.001, 2\pi - 0.001]
$$

and set $a_{\mathrm{Obs}} = 1$ as the parameter to be recovered. Observations are defined as

$$
(x_{\mathrm{Obs}}(t), y_{\mathrm{Obs}}(t)) = (\sin(a_{\mathrm{Obs}}t), \cos(a_{\mathrm{Obs}}t)) \ ,
$$

at a predefined set of time instants.

Figure 5.8 – Fitting of the model output to the observational data.
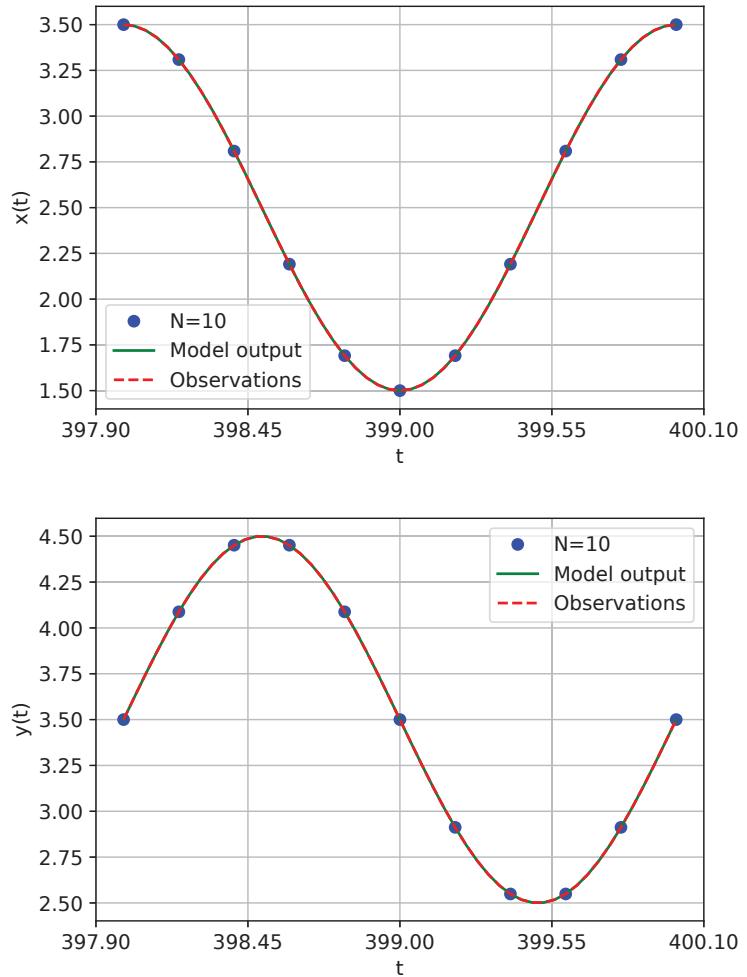
Considering the model defined by the equilibrium of equations (5.6), for $[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}] = [2.5, 3.5]$ and $N{+}1 = 11$ samples, following the structure of Figure 5.2.

Source: the author (2026).

Misfit residual definition: We consider two types of observations and corresponding residual entries:

1. **Type-1 (pointwise):** direct comparisons between $(x(t), y(t))$ and $(x_{\mathrm{Obs}}(t), y_{\mathrm{Obs}}(t))$ at selected time instants.

2. **Type-2 (aggregated):** groupings of type-1 observations into blocks and comparison of their averages.

Let m_1 be the number of type-1 residuals and m_2 the number of type-2 residuals, with $m = m_1 + m_2$.

We introduce a weighting constant $\alpha \in [0, 1]$ and define:

$$r_{\text{Misfit}}(a) = \left[\frac{\alpha^2}{m_1} \, r_1(a) \; , \; \frac{(1-\alpha)^2}{m_2} \, r_2(a) \right]$$

where $r_1(a) \in \mathbb{R}^{m_1}$ contains type-1 residuals, and $r_2(a) \in \mathbb{R}^{m_2}$ contains type-2 residuals (for example, the averages over fixed groups of type-1 misfits). The misfit function is again the squared norm of $r_{\text{Misfit}}(a)$.

The choice of (m_1, m_2, α) controls the dimension and structure of the residual and hence the cost and conditioning of the least-squares problem.

**Numerical experiment (non-identifiable sinusoidal system):**

Experiment 3. We consider the problem of recovering $a_{\text{Obs}} = 1$ with feasible region $\Omega = [0.001, 2\pi - 0.001]$. We test several configurations of (m_1, m_2, N, α), where $N + 1$ is the number of time samples used to define type-1 residuals:

$$m_1 = 5, \quad N = 20, \quad \alpha = 1 \,,$$

and variants with $m_2 > 0$ and $\alpha \in (0, 1)$.

Figures 5.9 and 5.10 show heuristic plots of $f_{\text{Misfit}}(a)$ in a neighborhood of a_{Obs} for two different residual settings. Although the global minimum is close to a_{Obs} and $f_{\text{Misfit}} \approx 0$ near it, the misfit function has many local minima. A zoom around a_{Obs} (Figure 5.11) shows the highly oscillatory structure that can trap DFO-LS in a local minima away from a_{Obs}.

Figure 5.9 – Heuristic evaluation of $f_{\text{Misfit}}(a)$.
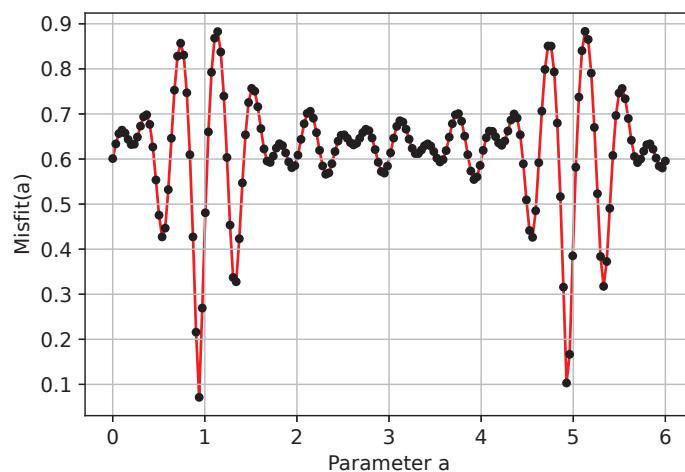
Evaluation at 180 points in a neighborhood of $a_{\text{Obs}} = 1$ (excluding a_{Obs} itself) for $m_1 = 5$, $N = 20$, $\alpha = 1$ (type-1 residuals only). Black dots indicate function evaluations, and the red line is an interpolation.
Source: the author (2026).

Figure 5.10 – Heuristic evaluation of $f_{\mathbf{Misfit}}(a)$.
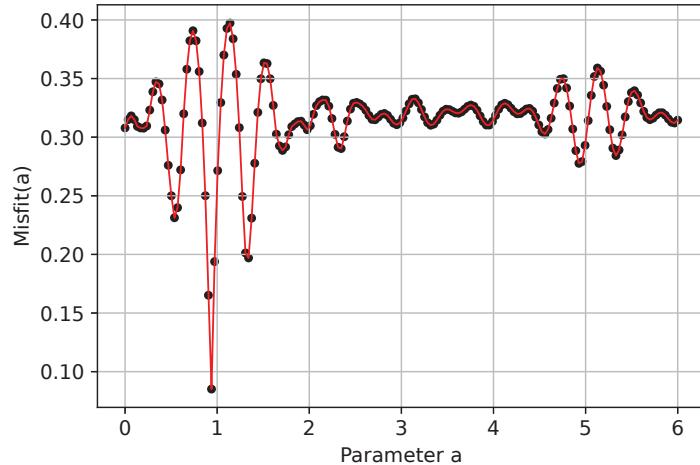
Evaluation at 180 points in a neighborhood of $a_{\mathrm{Obs}} = 1$ (excluding a_{Obs} itself) for $m_1 = 5$, $m_2 = 3$, $N = 20$, and $\alpha = 0.5$ (combining type-1 and type-2 residuals). Black dots indicate function evaluations, and the red line is an interpolation.

Source: the author (2026).

Figure 5.11 – Heuristic evaluation of $f_{\mathbf{Misfit}}(a)$.
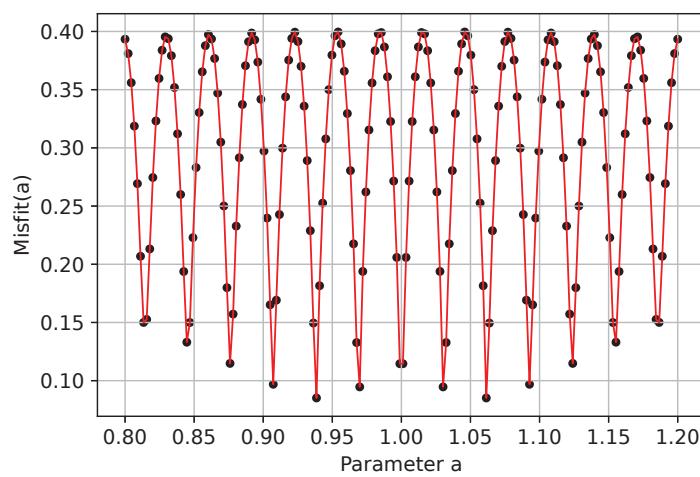
Evaluation at 180 points in a close neighborhood of $a_{\mathrm{Obs}} = 1$ (excluding a_{Obs} itself) for the same setting as Figure 5.10. Black dots indicate function evaluations, and the red line is an interpolation.

Source: the author (2026).

If we artificially restrict the search region to a very small interval around a_{Obs} (Figure 5.12), DFO-LS can typically recover a_{Obs}. However, this requires prior knowledge of the true parameter and therefore does not constitute a practical calibration strategy.

Figure 5.12 – Heuristic evaluation of $f_{\text{Misfit}}(a)$.
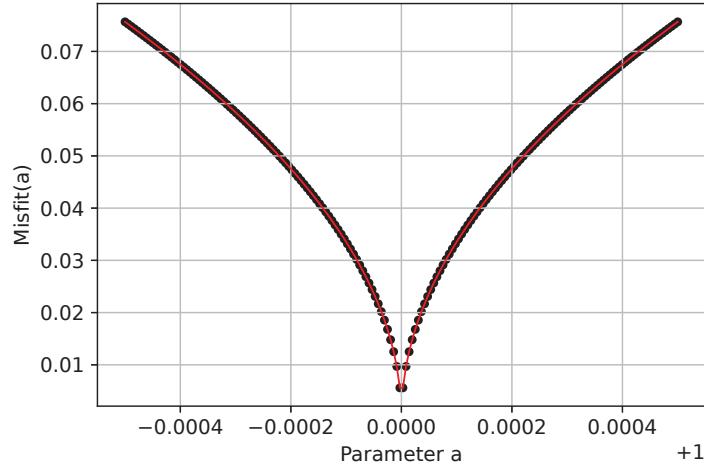

Evaluation at 180 points in a very small neighborhood of $a_{\text{Obs}} = 1$ (excluding a_{Obs} itself) for the same setting as Figure 5.10. Black dots indicate function evaluations, and the red line is an interpolation.
Source: the author (2026).

Last comments:

- For simplicity, we set the observational parameter as $a_{\text{Obs}} = 1$ in all these experiments, since for other observational parameters it is possible to follow the same procedure.

- As we will see in the following subsection, using heuristics to obtain a good initial guess is possible, which helps the DFO-LS solver find the solution more quickly. However, it is also possible to use an arbitrary initial guess, such as approximately the midpoint of the feasible interval, $a_0 = 3.14$.

- Regarding the type-2 of observations, we initially considered taking the value of the definite integral of the half-period ODE system as an observation and comparing it to the Riemann sum, obtained by manipulating the samples. However, this is another limitation for this calibration: When the estimate produced by the misfit function is coarse, comparing it to a precise theoretical reference does not benefit the calibration's accuracy: on the contrary, it can worsen it. In this sense, an important feature in defining observations is taking into account the model's ability to replicate them.

This example shows that even with a carefully constructed residual and a sophisticated optimizer, successful calibration may be impossible when the parameter is effectively non-identifiable from the available observations.

5.1.4 A nonlinear example: when the analytical solution is unknown

We now turn to a nonlinear system with a limit cycle and no closed-form solution, allowing us to test the calibration strategy in a setting closer to real biogeochemical models.

Model and feasible region: We consider the Sel'kov model (Strogatz [69], Example 7.3.3, p. 208):

$$\begin{cases} \dfrac{dx}{dt} & = -x + y \cdot (a + x^2) \, , \\[2mm] \dfrac{dy}{dt} & = b - y \cdot (a + x^2) \, . \end{cases} \tag{5.12}$$

where $a > 0$ and $b > 0$ are parameters. For suitable choices of (a, b), the system admits a stable fixed point or a stable limit cycle. In particular, for parameters in the region bounded by

$$b^2 = \frac{1}{2} \left(1 - 2a \pm \sqrt{1 - 8a} \right), \qquad a, b > 0 \, ,$$

which we define as the feasible region for the parameters in the following experiments (Figure 5.13), trajectories starting away from the equilibrium converge to a limit cycle, while trajectories starting sufficiently close to the equilibrium may converge to the fixed point.

Figure 5.13 – Convergence to a limit cycle.
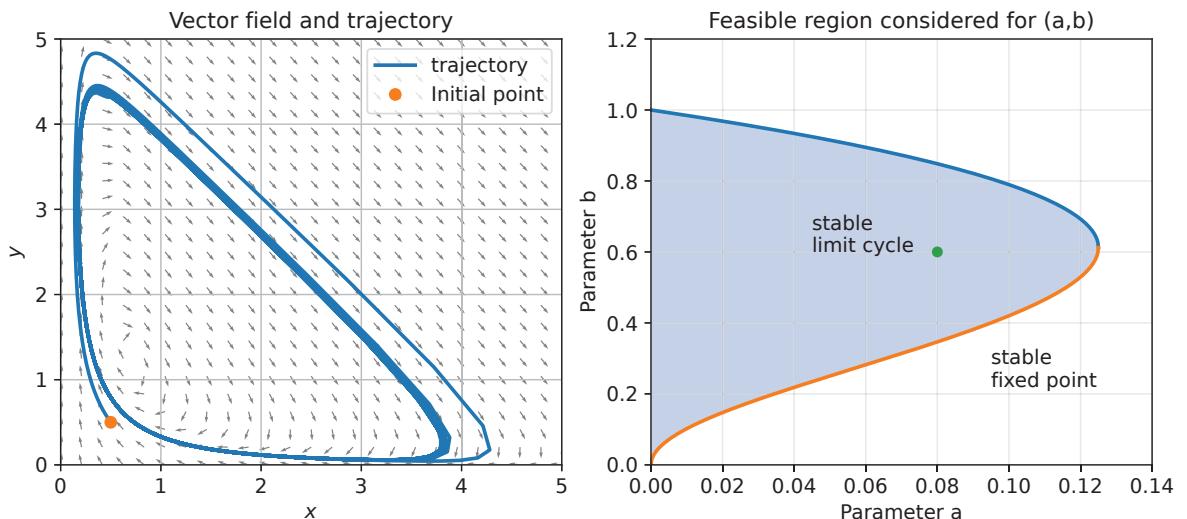
Right-hand panel shows a point $(\bar{a}, \bar{b})$ in the region defined by $b^2 = \frac{1}{2} \left(1 - 2a \pm \sqrt{1 - 8a} \right)$, $a, b > 0$. Left-hand panel shows the trajectories obtained by integrating the system (5.12) with parameters $(a, b) = (\bar{a}, \bar{b})$ and an initial point away from the equilibrium.
Source: the author (2026).

Producing the observations: Once we choose a pair of observational parameters inside the feasible region presented in Figure 5.13, the observations are constructed in two stages:

1. **Finding a point on the limit cycle:** we integrate (5.12) from a point slightly displaced from the fixed point and let the trajectory evolve for a long time, so that it converges to the limit cycle. The last point of this integration, denoted y_0, is taken as an approximation of a point on the observational limit cycle.

2. **Sampling and interpolation:** starting from y_0, we integrate (5.12) again over a time interval long enough to cover at least one full period of the limit cycle, and we save $N+1$ points along this integration. The resulting set of points constitutes the raw observational set. We then perform a linear interpolation to densify the observational curve so that the distance between consecutive points does not exceed a prescribed value Δc (Figure 5.14).

The parameter Δc is initially chosen empirically and can later be scaled relative to the diameter of the rectangle containing the limit cycle. In this way, we balance two competing goals: filling in gaps along the observational limit cycle and avoiding an excessive number of interpolation points, which would increase computational cost. Figure 5.14 compares the observational data, sampled from the approximated limit cycle obtained by integrating system (5.12) with observational parameters $[a_{\mathbf{Obs}}, b_{\mathbf{Obs}}] = [0.02, 0.6]$ and $N+1 = 401$ observations, and the interpolated dataset.

Figure 5.14 – Observational data sampled from the approximated limit cycle for $[a_{\mathbf{Obs}}, b_{\mathbf{Obs}}] = [0.001, 0.6]$, $N+1 = 401$ points (left panel), and the dataset obtained when adding interpolation points (right panel).
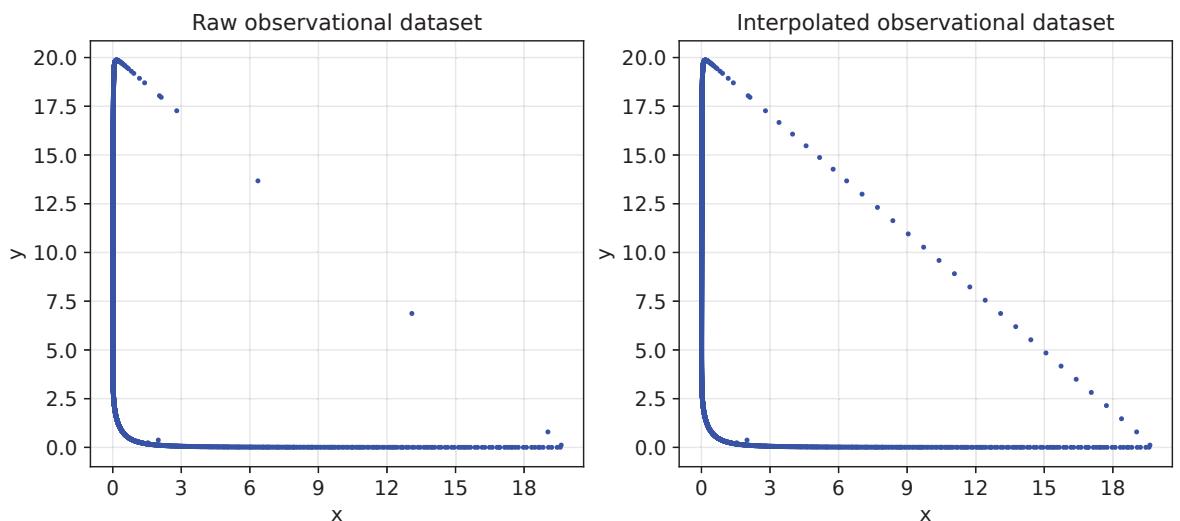
To implement the data interpolation, we first reorder the points obtained from the integration of the observational limit cycle. Then, the distance between consecutive points is calculated. If this distance is greater than Δc, points from the line segment connecting these two consecutive points are added to the observation set. This construction assumes that the observational limit cycle is (approximately) convex, ensuring that the piecewise linear interpolation is well-defined and does not introduce large geometric distortions.

In the following, we considered two options for setting the integration interval from which we sample the observations:

1. The last 100 integration times;

2. A time interval empirically expected to contain at least one, and at most two loops of the observational limit cycle.

In all the experiments presented here, the first option corresponded to a larger interval than the second one.

Optimization heuristic step: For the heuristic step, we define a simplified misfit function $f_H(a, b)$ comparing the distances between a small number of model output points and the observational dataset. The ODE system (5.12) is integrated from y_0 over a relatively short time interval, and $N + 1$ samples are taken at pre-defined time instants (which also need to be included when generating the observational set). The misfit $f_H(a, b)$ is taken as the least-squares sum of a four-dimensional version of the residual function, $r_H(a, b)$, defined as the first four entries of the residual function r_{Misfit}, which will be better described in the following.

We then build a coarse grid of parameter values (a, b) within the feasible region and evaluate f_H on this grid (Figure 5.15, left panel). The set of grid points that appear as local minimizers of f_H is retained for further refinement.

Initial guess and bounds: After evaluating f_H on the coarse grid, we select all local minimizers and, for each of them, we evaluate f_H at a finer set of points in a small neighborhood. The median of the resulting misfit values is used to define $\rho_{\max}$ (which we recall later), while the best local minimizer in this refined search is chosen as the initial guess for DFO-LS (Figure 5.15, right panel).

We then define personalized rectangular bounds around the initial guess, with width up to 20% of the total feasible range in each parameter. Assuming the initial guess is reasonable, this restriction helps DFO-LS focus on the most relevant region and reduces the number of iterations.

Misfit residual definition: We now define the residual function used by DFO-LS to calibrate (a, b). After the heuristic step, we construct the residual function $r_{\text{Misfit}}(a, b)$, which has five components:

- The first four components are the distance measures between the model output points and the observational dataset;

Figure 5.15 – Heuristic evaluation for recovering the parameter $a_{\mathbf{Obs}} = 0.08$, where the observational parameter $b_{\mathbf{Obs}} = 0.6$ was fixed.
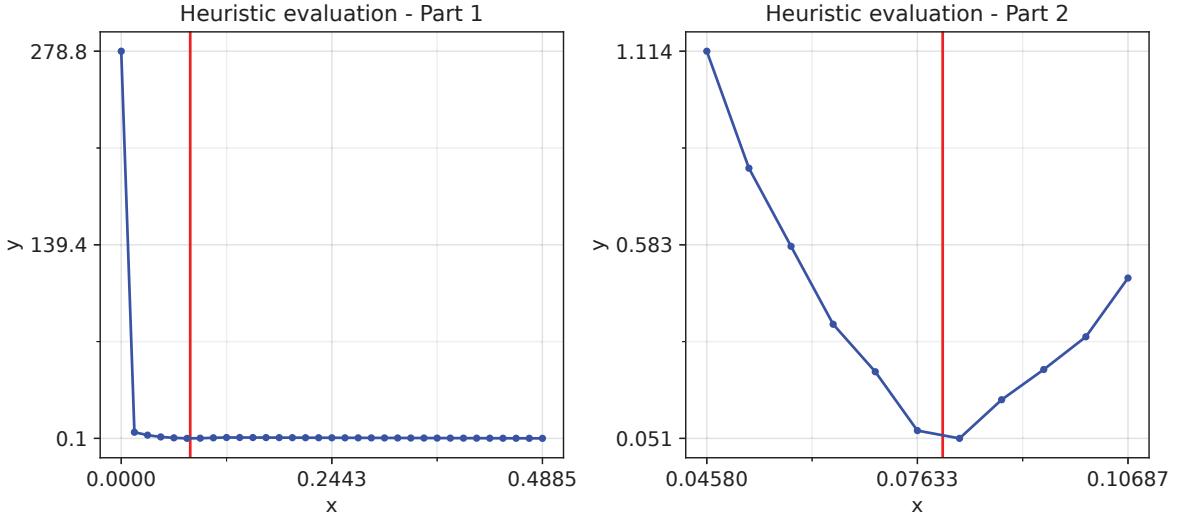
Left panel: First part of the heuristic evaluation, where a light version of the misfit function, $f_H(a)$, was evaluated over a discretization of the feasible interval for a into 33 points. Right panel: Heuristic refinement, with the evaluation of $f_H(a)$ at 11 points around each coarse local minimizer identified in the left panel, where a ceil is applied. In both plots, blue dots represent function evaluations, blue lines represent linear interpolation between these points, and the red vertical line identifies the observational parameter a_{Obs} on the x-axis.

Source: the author (2026).

- A fifth component is an external penalty enforcing the feasibility constraints for the parameters (a, b).

The observational limit cycle is divided into four quadrants (Figure 5.16). For each of the $N + 1$ points of the model output, we:

1. find the closest observational point;

2. compute the distance between this observation and the model output point;

3. and assign this distance to the residual component associated with the quadrant containing that observational point.

To promote continuity between quadrants, we also use transition regions (Figure 5.16): points near the boundaries contribute to more than one component via a smooth transition.

The fifth residual component is a penalty term, denoted $r_{\mathrm{penalty}}(a, b)$, that enforces the feasibility conditions

$$8a \leq 1, \qquad b^2 \leq \frac{1}{2}\Big(1 - 2a + \sqrt{1 - 8a}\Big), \qquad b^2 \geq \frac{1}{2}\Big(1 - 2a - \sqrt{1 - 8a}\Big).$$

94

Figure 5.16 – **Division of the observational limit cycle into four quadrants.**
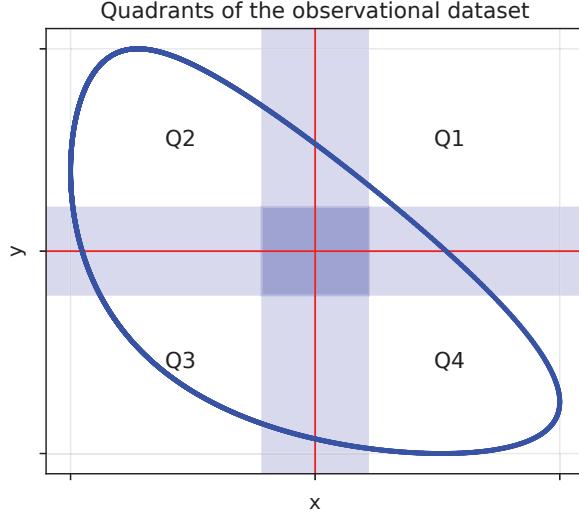
The red lines demarcate the limits of each quadrant. Blue rectangles identify transition regions between quadrants.
Source: the author (2026).

We define

$$
r_{\text{penalty}}(a, b) = \begin{cases} \rho_{\max}, & \text{if } v_1(a, b) > 0, \\[2mm] \min\left\{\rho_{\max}, p_{\text{scale}} \max\left\{0,\, v_2(a, b),\, v_3(a, b)\right\}\right\}, & \text{if } v_1(a, b) \leq 0, \end{cases}
$$

where

$$
\begin{aligned}
v_1(a, b) &= 8a - 1, \\
v_2(a, b) &= b^2 - \frac{1}{2}\left(1 - 2a + \sqrt{1 - 8a}\right), \\
v_3(a, b) &= -b^2 + \frac{1}{2}\left(1 - 2a - \sqrt{1 - 8a}\right),
\end{aligned}
$$

$\rho_{\max} > 0$ is a ceiling value for the residual components, and p_{scale} is a positive scaling constant (set to 1 in the experiments). The penalty ensures that points outside the feasible set incur a large residual.

The corresponding misfit function, $f_{\text{Misfit}}(a, b)$, is the squared norm of this five-dimensional residual vector.

Numerical experiments: Once the five-dimensional residual function and the initial guess are defined, we run DFO-LS to calibrate (a, b). The experiments are grouped into three tests, each corresponding to different calibration scenarios but all based on the same underlying example.

Experiment 4 (calibration of a for fixed b). We fix $b = b_{\text{Obs}} = 0.6$ and consider

$$[a_{\text{Obs}}, b_{\text{Obs}}] \in \{[0.001, 0.6], [0.02, 0.6], [0.04, 0.6], [0.06, 0.6], [0.08, 0.6], [0.1, 0.6], [0.12, 0.6]\} \ .$$

For each one of these seven cases, we run DFO-LS for $N + 1 \in \{26, 101, 401\}$ samples and compare the accuracy of the optimized parameter $\bar{a}$ in the recovery of a_{Obs}.

Figure 5.17 summarizes the results for two sampling strategies:

- Samples taken over the last 100 units of integration time;

- Samples taken over the last 1-2 periods of the limit cycle (that is, a shorter time window but still covering at least one full cycle).

Figure 5.17 – Experiment 4: accuracy in recovering a_{Obs}, with $b_{\text{Obs}} = 0.6$ fixed, depending on the number of samples $N + 1$ $N + 1$.
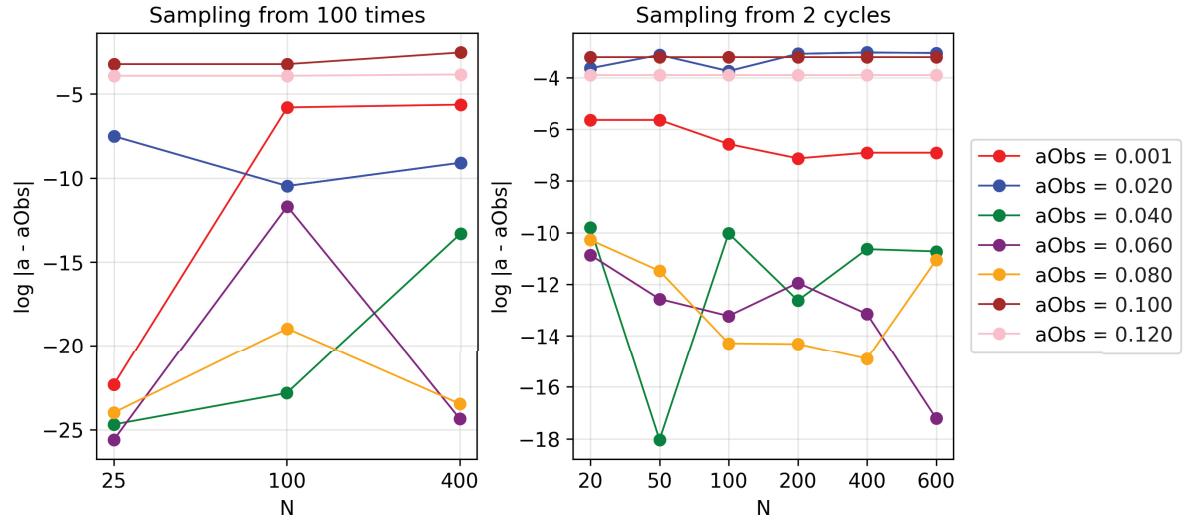
Results when samples were taken over the last 100 time units of the model integration (left), and results when samples were taken over approximately 1–2 periods of the limit cycle obtained by the model integration (right).

Source: the author.

Experiment 5 (calibration of b for fixed a). Here we fix $a = a_{\text{Obs}} = 0.04$ and vary b_{Obs}:

$$[a_{\text{Obs}}, b_{\text{Obs}}] \in \{[0.04, 0.3], [0.04, 0.4], [0.04, 0.6], [0.04, 0.7], [0.04, 0.9]\} \ .$$

Again, we consider $N + 1 \in \{26, 101, 401\}$ and compare sampling over the last 100 time units versus sampling over the last 1–2 periods. The results are shown in Figure 5.18.

Figure 5.18 – Experiment 5: accuracy in recovering $b_{\mathbf{Obs}}$, with $a_{\mathbf{Obs}} = 0.04$ fixed, depending on the number of samples $N + 1$ $N + 1$.
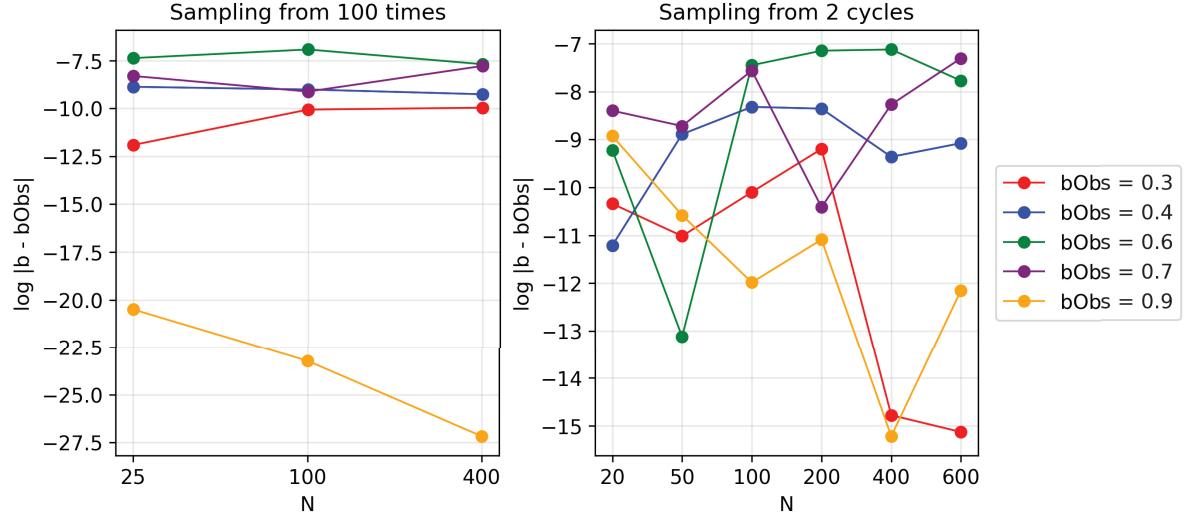
Results when samples were taken over the last 100 time units of the model integration (left), and results when samples were taken over approximately 1–2 periods of the limit cycle obtained by the model integration (right).

Source: the author.

Experiment 6 (simultaneous calibration of a and b). In the final set of experiments, we calibrate both parameters simultaneously for

$$[a_{\mathrm{Obs}}, b_{\mathrm{Obs}}] \in \{[0.02, 0.4], [0.02, 0.6], [0.02, 0.8], [0.06, 0.6], [0.12, 0.6]\} \ ,$$

and $N + 1 \in \{26, 101, 401\}$. Samples are taken over the last 100 time units. Figure 5.19 shows the calibration accuracy for different $(a_{\mathrm{Obs}}, b_{\mathrm{Obs}})$ and values of $N + 1$.

Last comments:

- To obtain the first 4 entries of the residual array, we divide the set of observations into quadrants. This strategy enables us to assign different weights to each quadrant, for example, proportionally to the number of observations located in each.

- For very small values of a_{Obs}, the Sel'kov system becomes stiff, and the misfit function may become ill-conditioned in a neighborhood of the observational parameters. This behavior becomes clear when attempting to recover $a_{\mathrm{Obs}} = 0.001$ while $b_{\mathrm{Obs}} = 0.6$ is fixed. In this case, the heuristic misfit $f_H(a)$ is poorly behaved near a_{Obs} (Figure 5.20), leading to a poor initial guess. As a consequence, the personalized bounds may exclude a_{Obs}, and even when the full feasible region is used, the DFO-LS

search tends to move away from the observational parameter a_{Obs}.

Figure 5.19 – Experiment 6: accuracy in simultaneously recovering $a_{\mathbf{Obs}}$ and $b_{\mathbf{Obs}}$ depending on the number of samples $N + 1$.
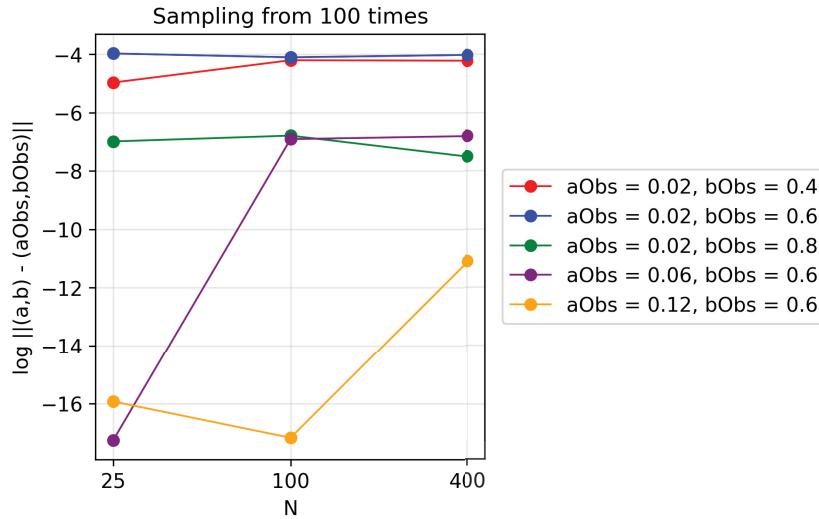
Samples were taken over the last 100 time units of the model integration.
Source: the author.

5.2 Results and Discussion

The experiments in this chapter highlight several modeling choices that strongly influence the quality and robustness of parameter calibration:

Heuristic steps: During the experiments in this chapter, we noted that not all calibration problems require a heuristic step. Well-behaved problems can even recover observational parameters using simpler strategies than those presented here. However, when dealing with more complex problems, performing a heuristic search to obtain a good initial guess generally reduces the computational cost of the calibration process, especially when the function used for the heuristic is computationally lighter than the function used for the calibration parameter search. Still, note that we cover examples of calibration with only one or two parameters: In problems with several parameters to be calibrated, it may be impractical to reproduce the heuristic step that considers a discretization of the feasible space of each variable, and so we may restrict the heuristic step to sample a set of points with less coverage of the search space.

Figure 5.20 – Heuristic evaluation for recovering the parameter $a_{\mathbf{Obs}} = 0.001$, where the observational parameter $b_{\mathbf{Obs}} = 0.6$ was fixed.
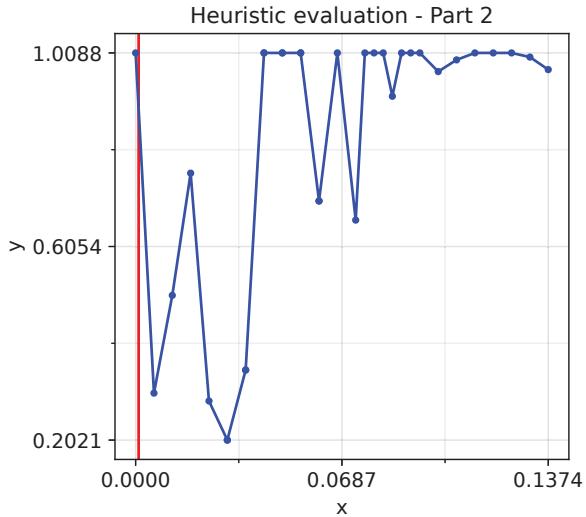
The function is ill-conditioned around a_{Obs}, leading to a bad choice of the initial guess. Points represent $f_H(a)$ evaluations with ceil, the red vertical line identifies the observational parameter a_{Obs} on the x-axis, which is a global minimizer for $f_H(a)$, not included in the evaluations. See the right panel of Figure 5.15.
Source: the author.

Misfit residual features: From the experiments in this chapter, we note that certain modeling and optimization choices can impact the quality of the results. One of them is the choice of an upper limit for the residual function entries (denoted as $\rho_{\max}$ in Subsection 5.1.4), which we will refer to as a *ceil*. This feature ensures that the search algorithm considers regions above the ceil as unattractive while mitigating scale variations that could compromise the accuracy of the results. By setting the ceil after the heuristic step, we ensure that this value is not too small, which would overly restrict the search region and make it difficult for DFO-LS to find a descent path to reach the minimizer. First, we have set the ceil as the maximum value of the misfit function obtained during the heuristic. This has allowed us to preserve all possible descent directions of the misfit function within the feasible set, a strategy that prioritizes exploration. Another choice was setting the ceil as the median of the heuristic misfit function values sampled in a neighborhood of local minimizers, in a strategy that prioritizes optimization accuracy. We considered the second strategy for the experiments in Subsection 5.1.4.

Effect of sample distribution and quadrant weights: In the Subsection 5.1.4, we represented the parameter space into quadrants, and considered transition regions between quadrants to ensure that the first four entries of the residual function effectively serve as coordinates. These transition regions aim to serve as a continuity factor between the first four entries of the residual array, achieved through a covering of the plane by four cards, and transition homeomorphisms between them.

When sampling more observations, we noticed that the accuracy of recovering observational parameters does not necessarily increase. Nevertheless, the overall accuracy remains very good. This phenomenon is due to the uneven distribution of points along the observational limit cycle obtained from the integration, although evenly distributed over time. Thus, in some regions of the limit cycle, the residual function compares the model output points only to the interpolated observational data. If there are model output points in a quadrant with these characteristics, there will be an additional error to the calibration. Even so, one precaution that should be taken is to avoid assigning weights too close to zero to quadrants containing few or no observations, as ignoring one of the quadrants leads to a calibration strategy similar to comparing model output and observational points one-on-one, which may only be effective in a small neighborhood of the optimal parameters.

In this context, an additional improvement is obtained by selecting the $N + 1$ samples from a shorter interval that empirically contains one to three periods of the limit cycle, rather than from a very long time window. In this case, the observational points tend to be more evenly distributed among quadrants, and interpolation becomes less critical.

Framework limitations and possible developments: There is a wide variety of optimization algorithms (for example, trust-region, direct search, model-based, evolutionary, surrogate, and Bayesian methods), and their performance depends strongly on characteristics of the problem, such as the amount of noise in the data, how smooth the function is, and the number of parameters [39]. For computationally expensive models, optimization methods that construct simpler probabilistic or regression-type approximations of the misfit function are particularly useful, as they can reduce the number of costly ODE integrations and handle noisy data in a natural manner [52, 74]. In problems involving the calibration of multiple parameters simultaneously, multi-objective versions of MADS and other derivative-free methods can be used to approximate Pareto fronts in a fully black-box setting [11].

From a modeling perspective, the framework proposed here can be extended in several directions. First, identifiability and sensitivity analyses (for example, variance-based methods) could be included as a preliminary step to help decide which parameters should be calibrated and how to design the residuals and observation sets. Second, the heuristic step, ceilings, and residual weights could be chosen adaptively, for example, using robust statistics of the misfit values or simple learning rules that update these choices during optimization [2, 74]. Third, the current form of the residuals (based on quadrants, penalties, and aggregated distances) could be altered to reflect better realistic data features of the problem approached. Finally, the full framework could be tested on larger conceptual models and real datasets, where model error, structural mismatch, and computational features are evaluated in detail.

In this sense, the examples presented in this chapter serve as a proof of concept, revealing both the strengths and limitations of systematic calibration with derivative-free methods, and pointing to a possible path for combining problem-specific residual design with more advanced and diverse optimization strategies.

Computational implementation: All models in this chapter were solved numerically using an implicit variable-step integration scheme based on the Backward Differentiation Formula (BDF) method. Relative and absolute tolerances were adaptively adjusted between 10^{-12} and 10^{-3} to ensure stable convergence and completion of the integration while keeping numerical errors as small as possible. As the examples show, the numerical integration settings can directly affect the smoothness and conditioning of the misfit function, and hence the performance of the calibration procedure. Further implementation details, including Python code and additional experiments, are provided in the *Appendix*.

6 A data-constrained model for the PEC biogeochemistry dynamics

In this chapter, we apply the parameter-calibration framework developed in *Chapter 5* to the conceptual model of the Paranaguá Estuarine Complex (PEC) developed in *Chapter 3*. To this end, we fit the model outputs to an observed dataset of nitrate (NO_3) and phytoplankton (measured in terms of Chlorophyll-A) over a one-year period. Finally, we use the resulting data-constrained model to test a few scenarios involving increased riverine nitrate loads. The Python code for these experiments is available in the *Appendix*.

6.1   General modeling setup

We consider the conceptual two-box model for the PEC described in *Chapter 3*. Specific details on our modeling approach are presented below. The data were identified from plots in [40] using *Web Plot Digitalizer* [66]. For the PEC dynamics using literature data, we initially applied linear interpolation to the observations, but the model–data fit was poor. We then hypothesized the presence of measurement noise and applied a smoothing strategy, in which each monthly observation was replaced by the average of that month and the two previous months. These data was presented in *Chapter 3*, as the upper box tracers' concentration in figures 3.12 and 3.13.

Forcing, initial and boundary conditions: Initial conditions for the model integration are obtained from linear interpolation of observational data reported in the literature [40], as presented in *Chapter 3*.

Target parameters: The parameters to be calibrated in this model are the daily maximum phytoplankton growth rate, $V_{\max}$, and the mortality rate λ.

Heuristic: For the heuristic, we considered a version of the misfit function in which the last year of the model integration over a period of 51 years is compared against 13 observations distributed over one year. For the benchmark tests, we implemented only one heuristic step, where we searched for a good initial guess for the local search on the best-fitting parameters. During the computational implementation, this stage was

parallelized.

Optimization approach: We used the DFO-LS solver for parameter calibration with an initial guess determined by the heuristic and bounds estimated from theoretical considerations in the literature [4, 10, 56]. The misfit residual array compares the last year of the model integration over 201 years to N+1 observations distributed across one year, with $N \in \{12, 120\}$. In all cases, we used the basic configuration of DFO-LS implementation [64], including bounds and the option `scaling_within_bounds`. The DFO-LS Python package also offers options to handle noise, which were not explored in this study.

6.2 Benchmarking and validating the approach

To validate the fitting capability of our framework, we conducted tests with the following objectives:

1. Ideally, to recover the observational parameters through the optimization strategy.

2. To reproduce the behavior of observational data using the calibrated model.

In this section, we will set chosen observational parameters and artificially generated observational data as outputs obtained from our model, based on the observational parameters and the previously considered forcing data. To obtain the observational data, we integrated our model with the observational parameters set over the 1000-year integration period, which is considered sufficient time for the model to reach an equilibrium state. Then, we sample the observations from the last year of the integration period at the same time instants considered in the definitions of each misfit function version.

For benchmark tests, the parameters $V_{\max}$ and λ were considered constant throughout the year, that is, invariant with respect to time. This simplification was related to the low level of complexity of the conceptual model, although it is not the only possibility. In the next section, we will see an example of calibration involving a time-varying parameter.

Considering the assumption that the behavior of the PEC phytoplankton population can be represented as a combination of the diatom species *S. costatum* and *A. glacialis* [62], we set bounds for the parameter space as:

$$0 \leq V_{\max} \leq \max\{V_{\max,1}, V_{\max,2}\}, \tag{6.1}$$

$$0 \leq \lambda \leq 1, \tag{6.2}$$

where $V_{\max,1} = 2.5$ and $V_{\max,2} = 1.5$ are set as the maximum daily growth parameter estimates for the *S. costatum* phytoplankton population [24], and *A. glacialis* phytoplankton population [56], respectively. As we did not perceive a significant enhancement in accuracy due to taking 120 samples instead of 12 in the

previous simulations, here we considered only the second option. We now present the case studies used as benchmark tests.

Case 1: Here, we choose the observational parameters:

$$V_{\max}^{\text{Obs}} = 1.0,$$
$$\lambda^{\text{Obs}} = 0.05.$$

That means the observations were generated under the assumptions of low daily growth and moderate daily mortality of the phytoplankton population. The artificial observations generated by the model outputs when considering these observational parameters and the fitting attained after the calibration process are presented in figures 6.1 and 6.2.

Figure 6.1 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 1 experiment.
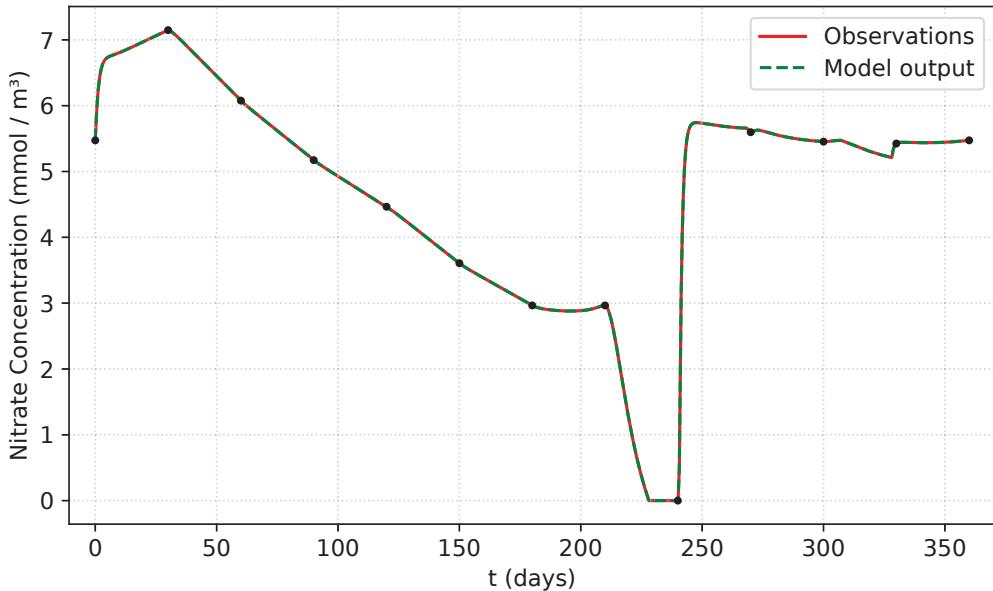
Observations were artificially generated by integrating the model with observational parameters $V_{\max}^{\text{Obs}} = 1.0$, and $\lambda^{\text{Obs}} = 0.05$. The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx V_{\max}^{\text{Obs}} + 8 \cdot 10^{-8}$, and $\bar{\lambda} \approx \lambda^{\text{Obs}} + 2 \cdot 10^{-9}$. The points represent the $N + 1$ fitting points used to calculate the misfit residual array; here, $N = 12$.

Source: the author.

Figure 6.2 – Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 1 experiment.
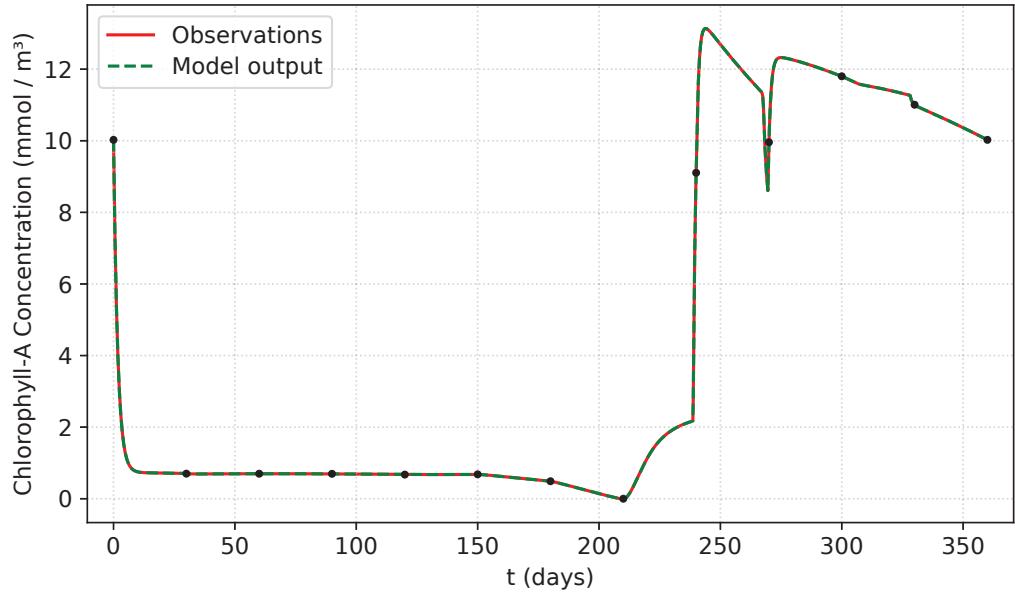
See the full description of the experiment settings in Figure 6.1.
Source: the author.

The parameter calibration is presented in figures 6.3 and 6.4, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 2.1700000000000004,$$

$$\lambda^{\text{Initial}} = 0.089,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 1.000000084741517,$$

$$\bar{\lambda} = 0.05000000269829062.$$

Figure 6.3 – Heuristic step on the calibration process of the parameters V_{max} and λ, in the Case 1 experiment.
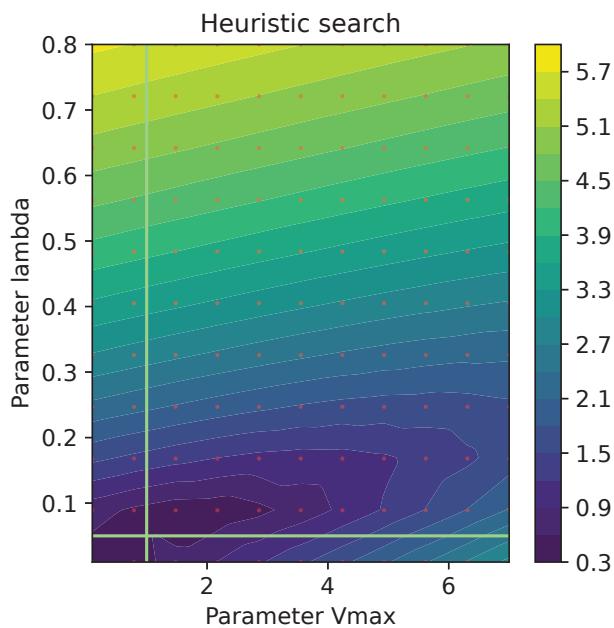
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the grid of points in the background identifies the parameters evaluated during the heuristic. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Figure 6.4 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 1 experiment.
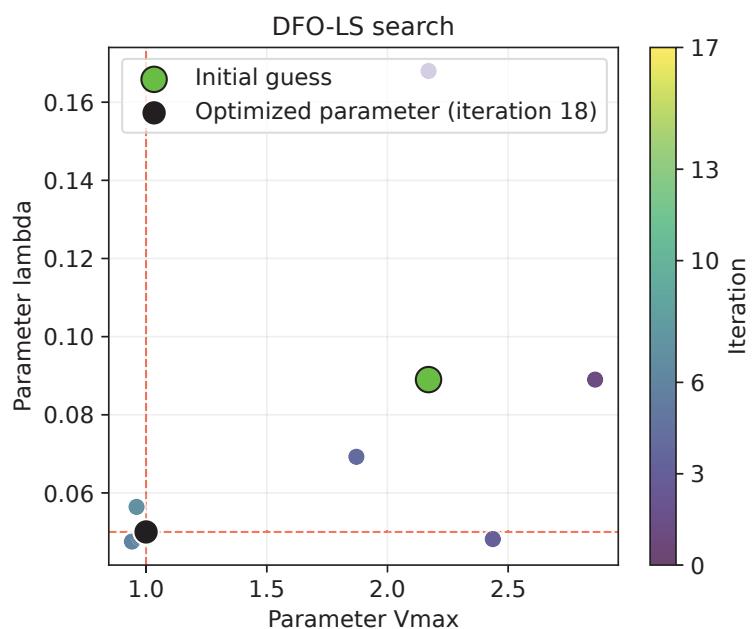
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.3). The sidebar assigns a color scale to the iterations, allowing identification of the convergence pattern. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Case 2: Here, we choose the observational parameters:

$$V_{\max}^{\text{Obs}} = 1.4,$$

$$\lambda^{\text{Obs}} = 0.05.$$

That means the observations were generated under the assumptions of moderate daily growth and moderate daily mortality of the phytoplankton population. The artificial observations generated by the model outputs when considering these observational parameters and the fitting attained after the calibration process are presented in figures 6.5 and 6.6.

Figure 6.5 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 2 experiment.
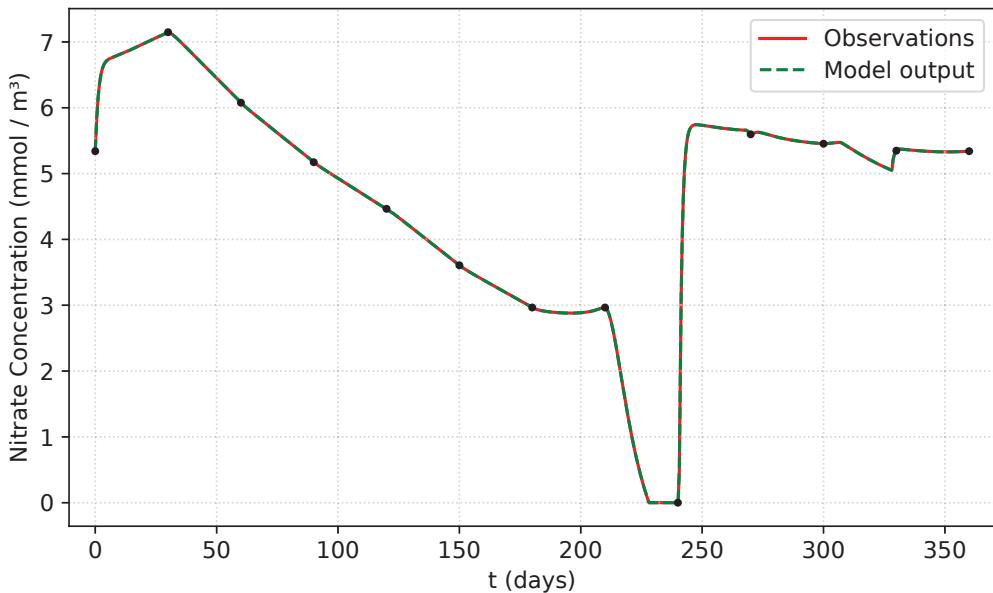
Observations were artificially generated by integrating the model with observational parameters $V_{\max}^{\text{Obs}} = 1.4$, and $\lambda^{\text{Obs}} = 0.05$. The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx V_{\max}^{\text{Obs}} + 9 \cdot 10^{-8}$, and $\bar{\lambda} \approx \lambda^{\text{Obs}} - 1 \cdot 10^{-9}$. The points represent the $N + 1$ fitting points used to calculate the misfit residual array; here, $N = 12$.

Source: the author.

Figure 6.6 – **Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 2 experiment.**
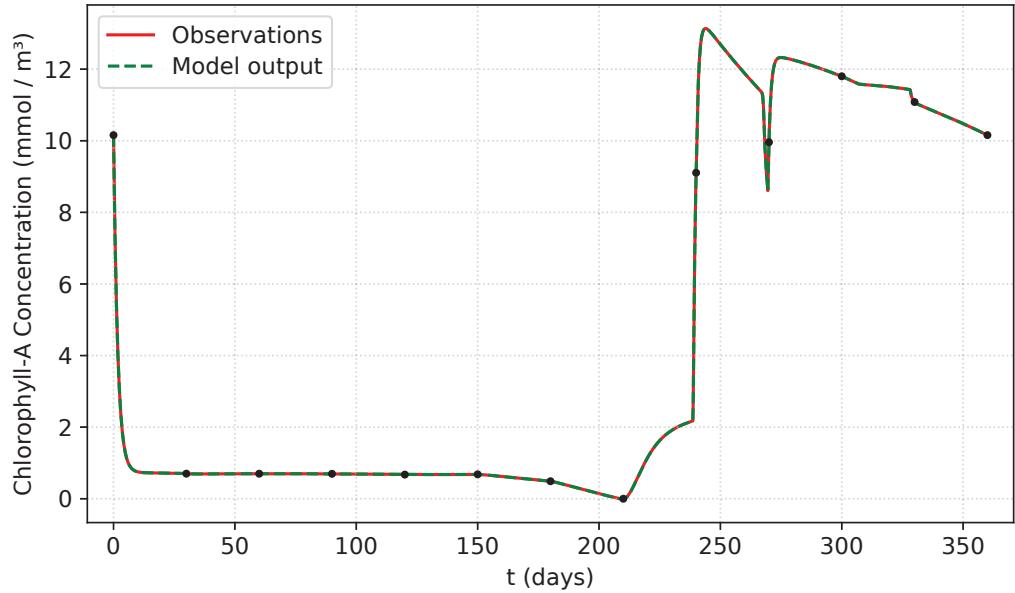
See the full description of the experiment settings in Figure 6.5.
Source: the author.

The parameter calibration is presented in figures 6.7 and 6.8, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 2.1700000000000004,$$

$$\lambda^{\text{Initial}} = 0.089,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 1.4000000989350427,$$

$$\bar{\lambda} = 0.04999999880062813.$$

Figure 6.7 – Heuristic step on the calibration process of the parameters $V_{\max}$ and λ, in the Case 2 experiment.
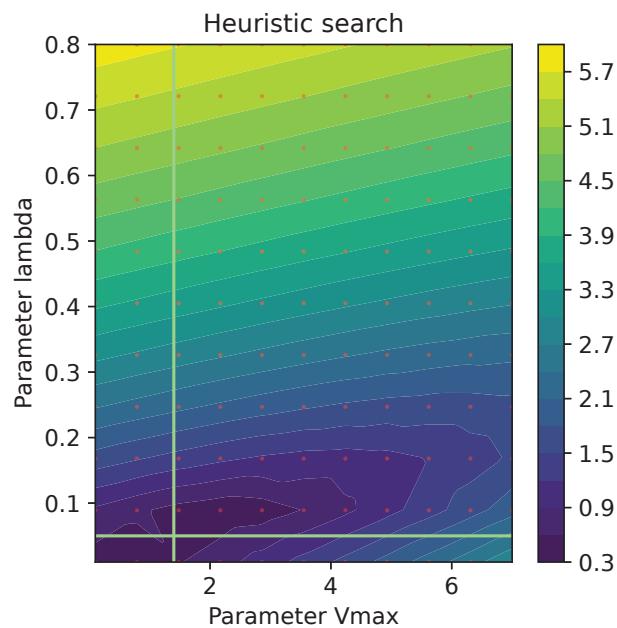
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the background grid of points identifies the parameters evaluated during the heuristic. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Figure 6.8 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 2 experiment.
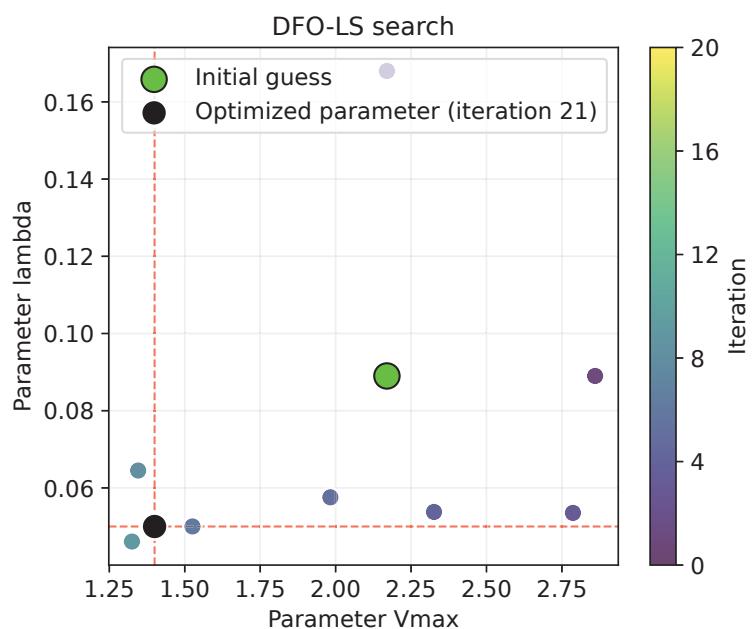
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.7). The sidebar assigns a color scale to the iterations, allowing identification of the convergence pattern. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Case 3: Here, we choose the observational parameters:

$$V_{\max}^{\text{Obs}} = 1.4,$$

$$\lambda^{\text{Obs}} = 0.3.$$

That means the observations were generated under the assumptions of moderate daily growth and high daily mortality of the phytoplankton population. The artificial observations generated by the model outputs when considering these observational parameters and the fitting attained after the calibration process are presented in figures 6.9 and 6.10.

Figure 6.9 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 3 experiment.
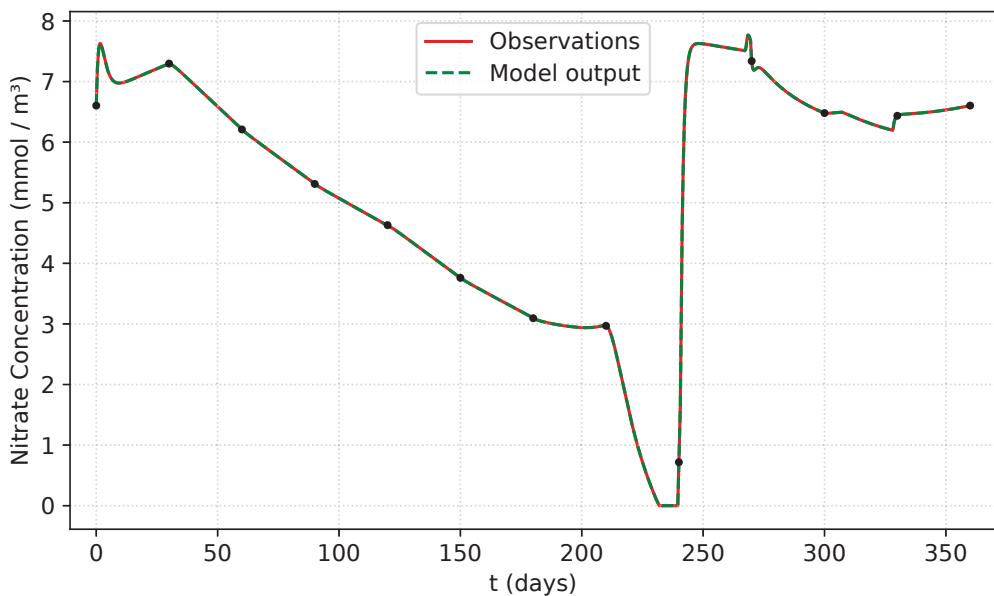
Observations were artificially generated by integrating the model with observational parameters $V_{\max}^{\text{Obs}} = 1.4$, and $\lambda^{\text{Obs}} = 0.3$. The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx V_{\max}^{\text{Obs}} + 3 \cdot 10^{-8}$, and $\bar{\lambda} \approx \lambda^{\text{Obs}} + 8 \cdot 10^{-10}$. The points represent the $N + 1$ fitting points used to calculate the misfit residual array; here, $N = 12$.

Source: the author.

Figure 6.10 – Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 3 experiment.
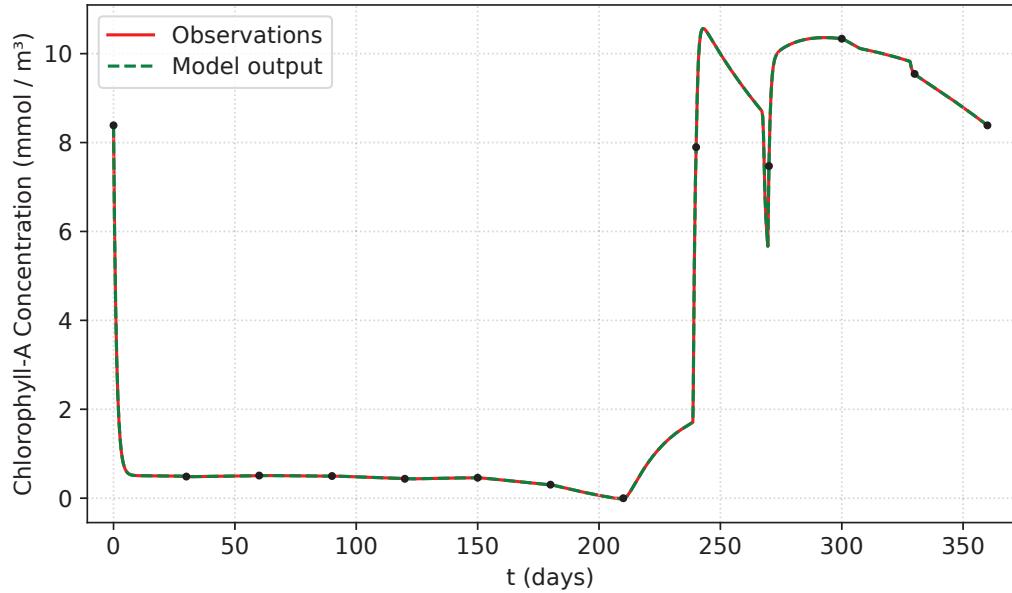
See the full description of the experiment settings in Figure 6.9.
Source: the author.

The parameter calibration is presented in figures 6.11 and 6.12, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 2.1700000000000004,$$

$$\lambda^{\text{Initial}} = 0.326,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 1.400000032860641,$$

$$\bar{\lambda} = 0.3000000008494046.$$

Figure 6.11 – Heuristic step on the calibration process of the parameters V_{max} and λ, in the Case 3 experiment.
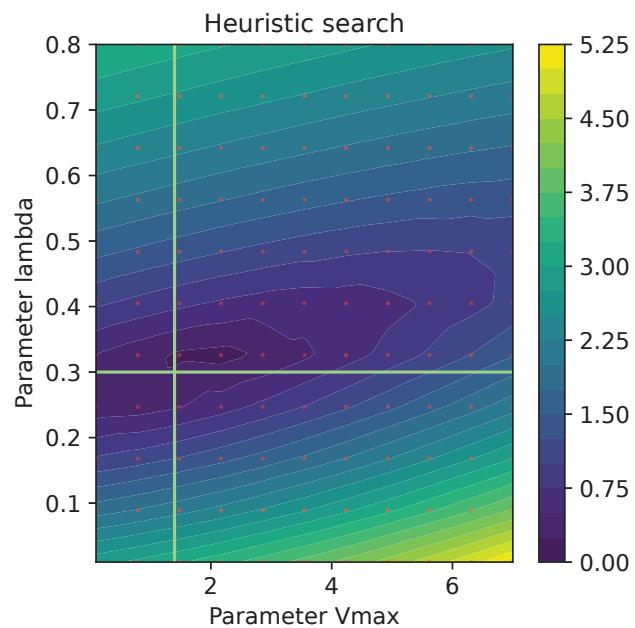
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the background grid of points identifies the parameters evaluated during the heuristic. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Figure 6.12 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 3 experiment.
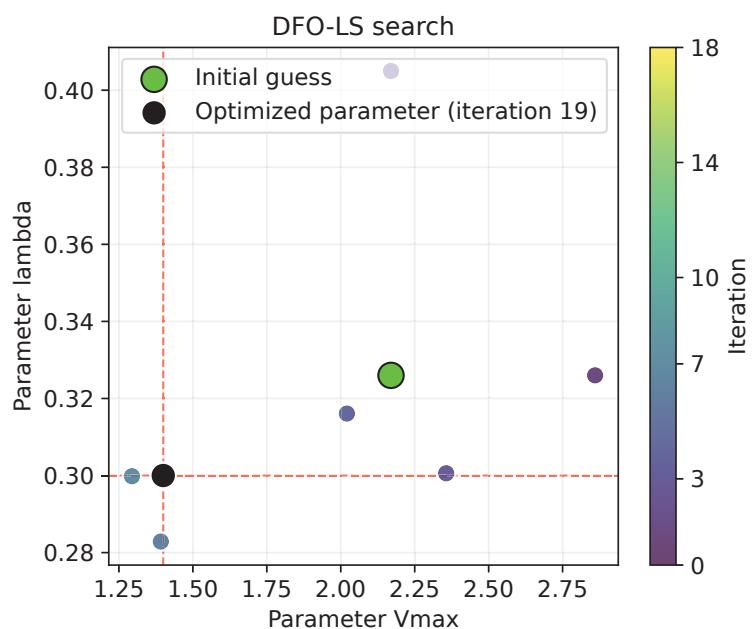
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.11). The sidebar assigns a color scale to the iterations, allowing identification of the convergence pattern. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

**Case 4:** Here, we choose the observational parameters:

$$V_{\max}^{\mathrm{Obs}} = 2.0,$$

$$\lambda^{\mathrm{Obs}} = 0.05.$$

That means the observations were generated under the assumptions of high daily growth and moderate daily mortality of the phytoplankton population. The artificial observations generated by the model outputs when considering these observational parameters and the fitting attained after the calibration process are presented in figures 6.13 and 6.14.

Figure 6.13 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 4 experiment.
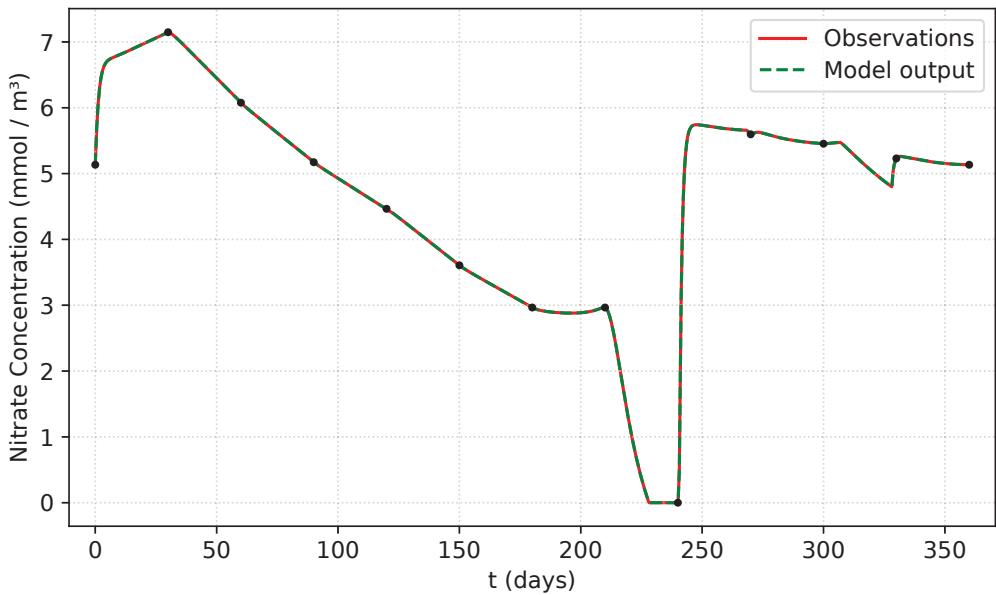
Observations were artificially generated by integrating the model with observational parameters $V_{\max}^{\mathrm{Obs}} = 2.0$, and $\lambda^{\mathrm{Obs}} = 0.05$. The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx V_{\max}^{\mathrm{Obs}} + 1 \cdot 10^{-6}$, and $\bar{\lambda} \approx \lambda^{\mathrm{Obs}} + 9 \cdot 10^{-9}$. The points represent the $N+1$ fitting points considered for the calculation of the misfit residual array, here we consider $N = 12$.

Source: the author.

Figure 6.14 – Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 4 experiment.
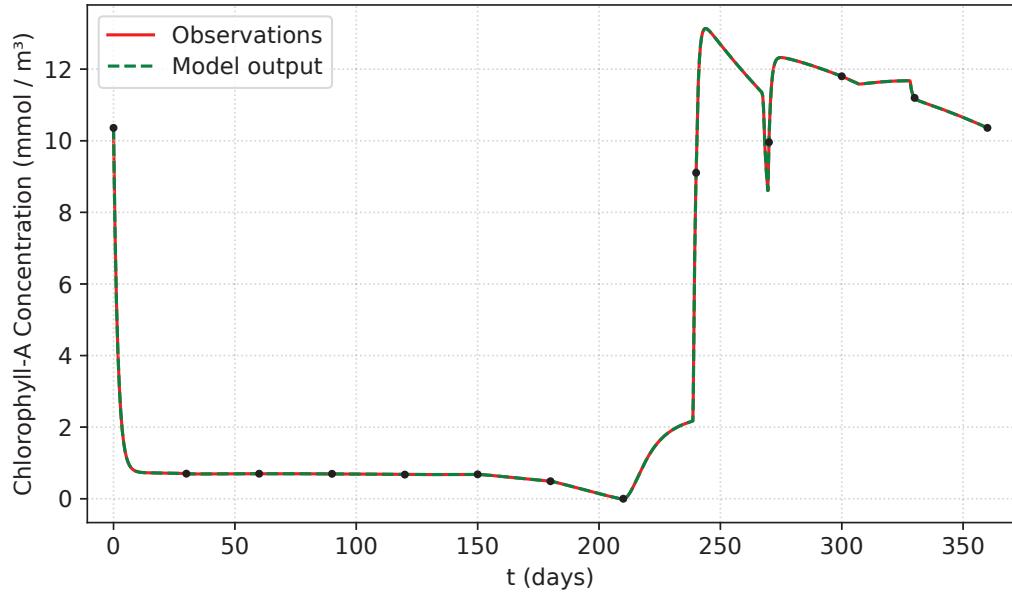
See the full description of the experiment settings in Figure 6.13.
Source: the author.

The parameter calibration is presented in figures 6.15 and 6.16, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 2.8600000000000003,$$

$$\lambda^{\text{Initial}} = 0.089,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 2.0000014823868333,$$

$$\bar{\lambda} = 0.05000000913800312.$$

Figure 6.15 – Heuristic step on the calibration process of the parameters V_{max} and λ, in the Case 4 experiment.
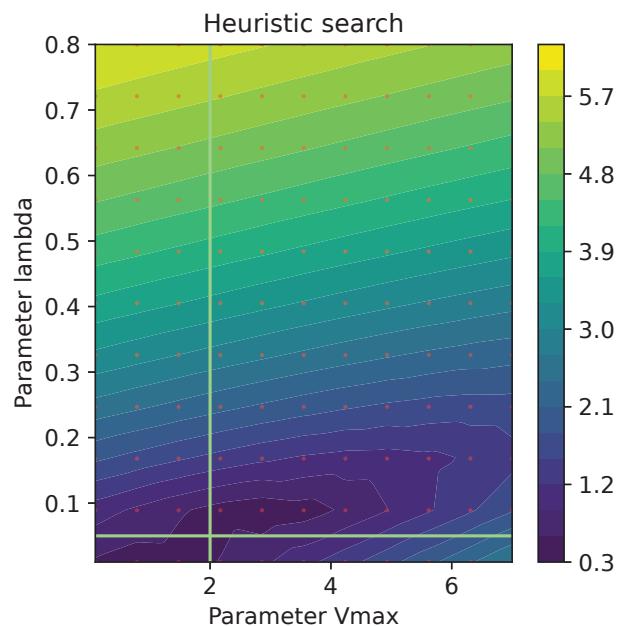
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the background grid of points identifies the parameters evaluated during the heuristic. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Figure 6.16 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 4 experiment.
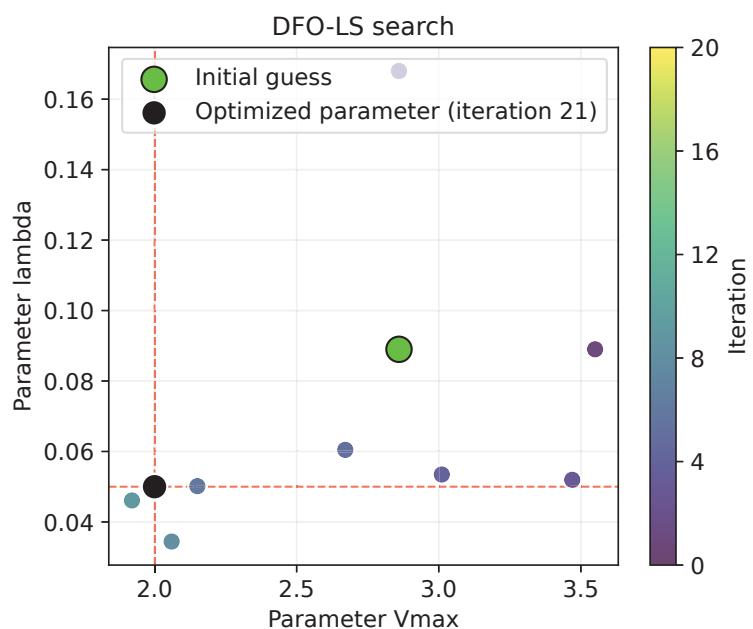
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.15). The sidebar assigns a color scale to the iterations, allowing identification of the convergence pattern. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Case 5: Here, we choose the observational parameters:

$$V_{\max}^{\mathrm{Obs}} = 2.0,$$

$$\lambda^{\mathrm{Obs}} = 0.3.$$

That means the observations were generated under the assumptions of high daily growth and high daily mortality of the phytoplankton population. The artificial observations generated by the model outputs when considering these observational parameters and the fitting attained after the calibration process are presented in figures 6.17 and 6.18.

Figure 6.17 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 5 experiment.
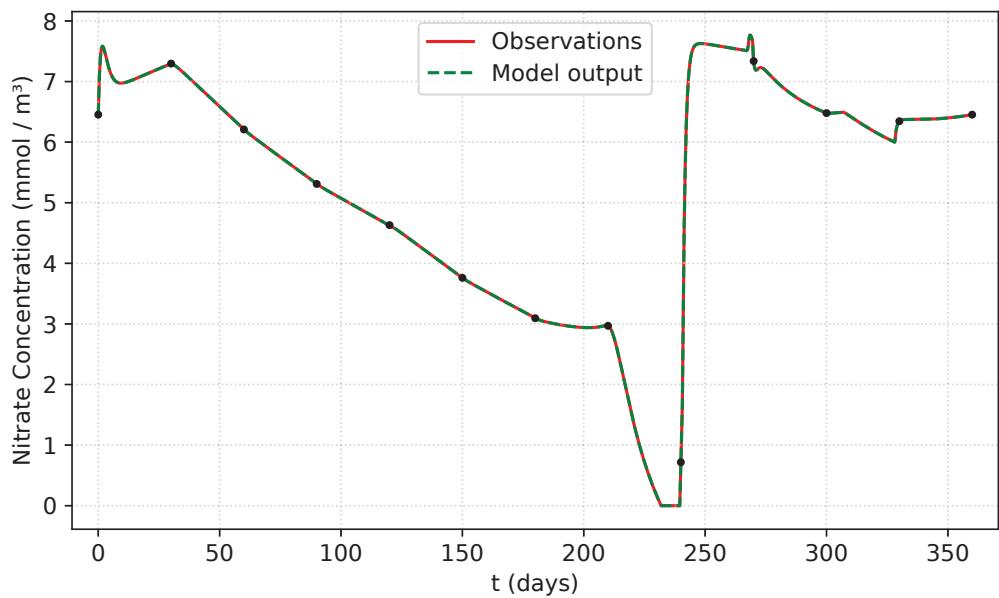
Observations were artificially generated by integrating the model with observational parameters $V_{\max}^{\mathrm{Obs}} = 2.0$, and $\lambda^{\mathrm{Obs}} = 0.3$. The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx V_{\max}^{\mathrm{Obs}} + 1 \cdot 10^{-6}$, and $\bar{\lambda} \approx \lambda^{\mathrm{Obs}} - 1 \cdot 10^{-8}$. The points represent the $N+1$ fitting points considered for the calculation of the misfit residual array, here we consider $N = 12$.

Source: the author.

Figure 6.18 – Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 5 experiment.
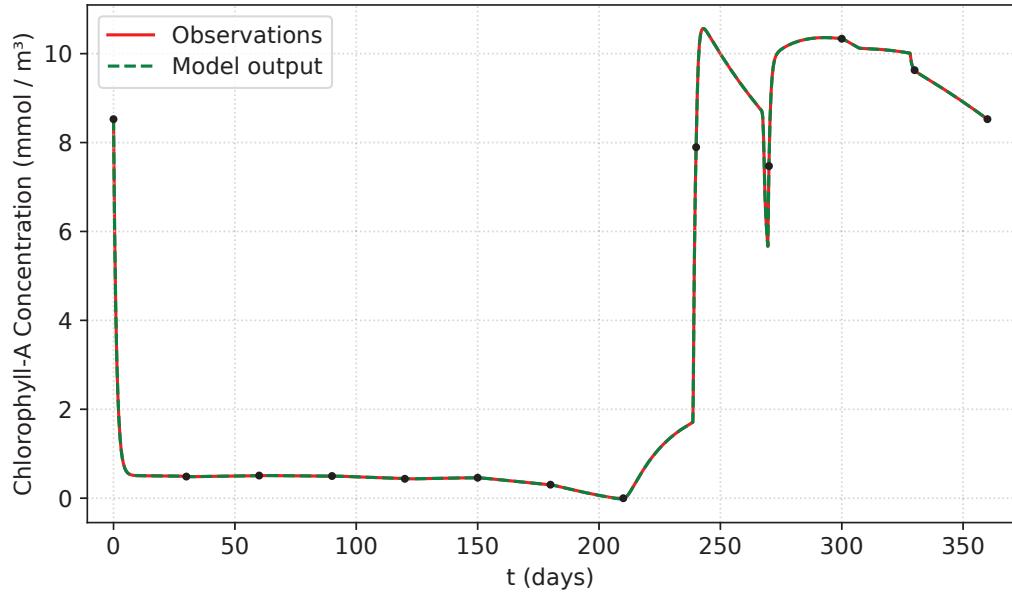
See the full description of the experiment settings in Figure 6.17.
Source: the author.

The parameter calibration is presented in figures 6.19 and 6.20, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 2.8600000000000003,$$

$$\lambda^{\text{Initial}} = 0.326,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 2.000001216474024,$$

$$\bar{\lambda} = 0.2999999872607704.$$

Figure 6.19 – Heuristic step on the calibration process of the parameters $V_{\max}$ and λ, in the Case 5 experiment.
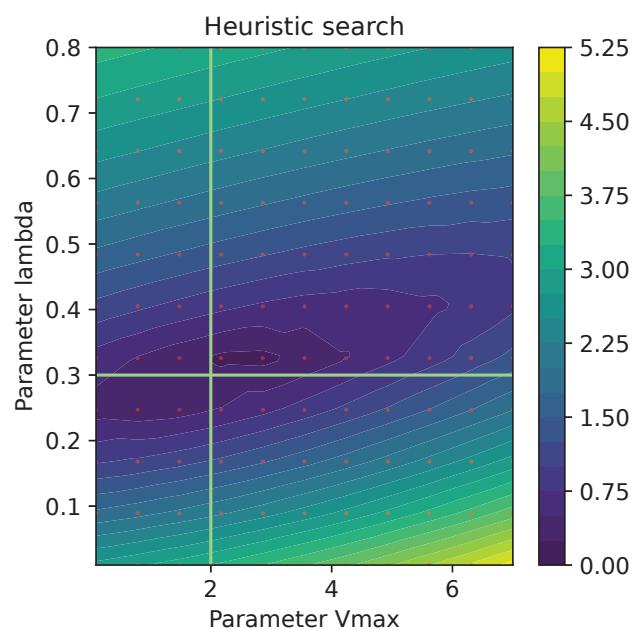
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the background grid of points identifies the parameters evaluated during the heuristic. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

Figure 6.20 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 5 experiment.
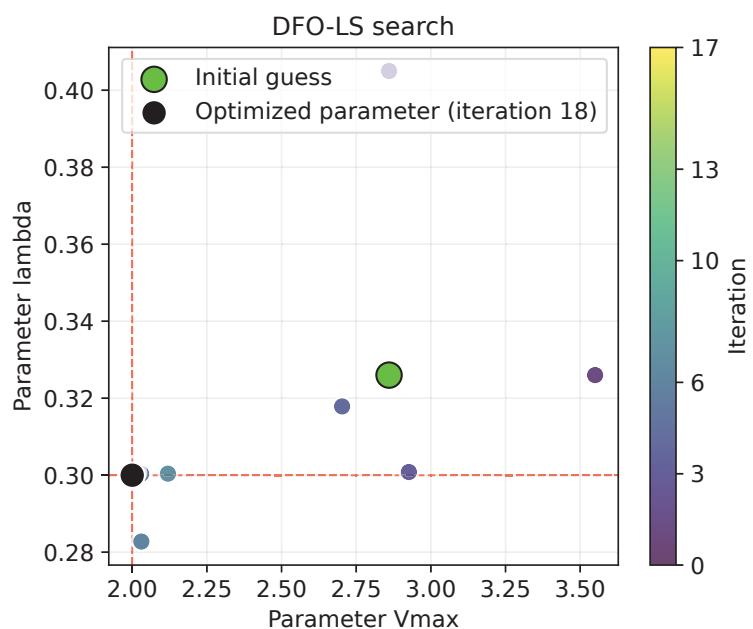
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.19). The sidebar assigns a color scale to the iterations, allowing identification of the convergence pattern. The intersection of the highlighted horizontal and vertical lines marks the location of the observational parameters.

Source: the author.

6.3 Fitting the model to observations

Here we present two experiments on fitting the model to the data available, without knowing the values of observational parameters. In our first experiment, we calibrate both parameters $V_{\max}$ and λ simultaneously. In our second experiment, we set time-dependent values for $V_{\max}$ and calibrate only the parameter λ.

Case 6: Calibration of two parameters simultaneously. The fitting attained after the calibration process is presented in figures 6.21 and 6.22.

Figure 6.21 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 6 experiment.
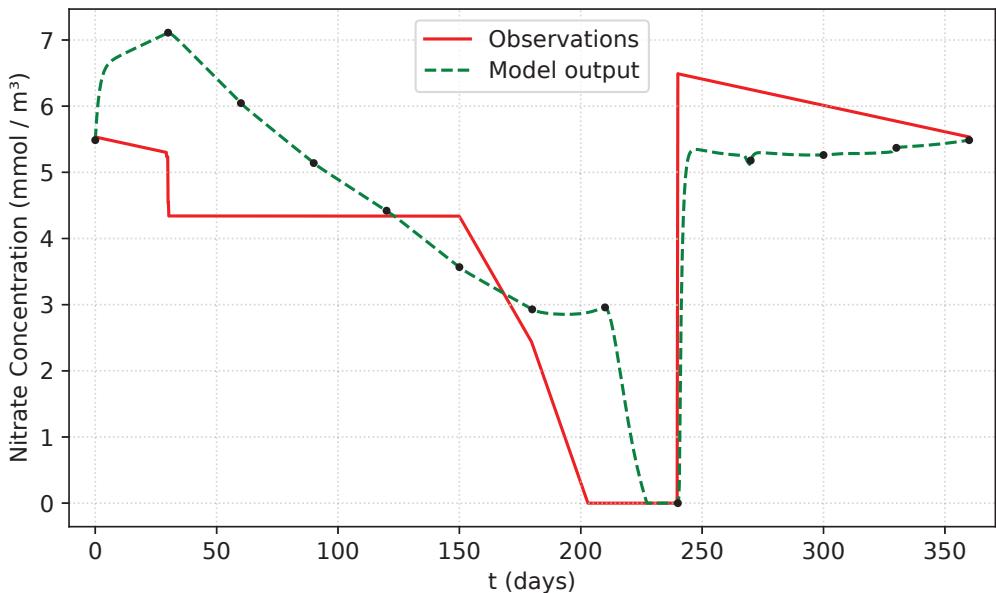
Observations were generated by interpolation of tracers' concentration data (figures 3.12 and 3.13). The model output plot was obtained by integrating the model described by equations (3.8) - (3.10) with the calibrated parameters $\bar{V}_{\max} \approx 0.24$, and $\bar{\lambda} \approx 0.01$. The points represent the $N + 1$ fitting points used to calculate the misfit residual array; here, $N = 12$.

Source: the author.

Figure 6.22 – **Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 6 experiment.**
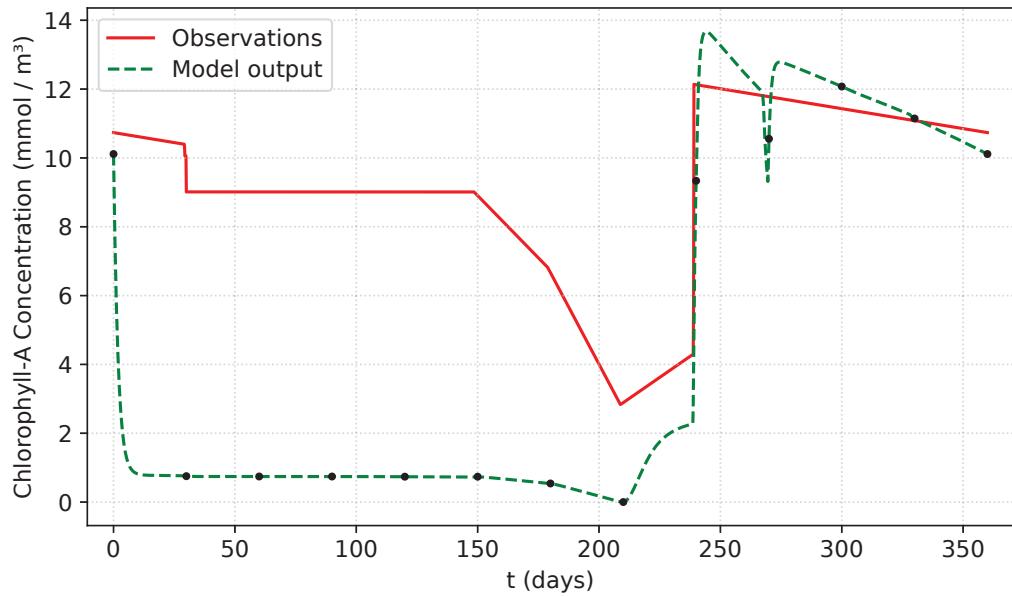
See the full description of the experiment settings in Figure 6.21.
Source: the author.

The parameter calibration is presented in Figures 6.23 and 6.24, as the one-step heuristic search for an adequate initial guess, which was found as:

$$V_{\max}^{\text{Initial}} = 0.7899999999999999,$$

$$\lambda^{\text{Initial}} = 0.010000000000000009,$$

and the DFO-LS search, which converged to the optimized parameters:

$$\bar{V}_{\max} = 0.24001317934207675,$$

$$\bar{\lambda} = 0.01.$$

This represents both a very low daily growth of the phytoplankton population and a very low daily mortality rate. This may result from a correlation between the two parameters or even from the water circulation flux of the PEC. To evaluate the first of these possibilities, we carried out the following experiment.

Figure 6.23 – Heuristic step on the calibration process of the parameters V_{max} and λ, in the Case 6 experiment.
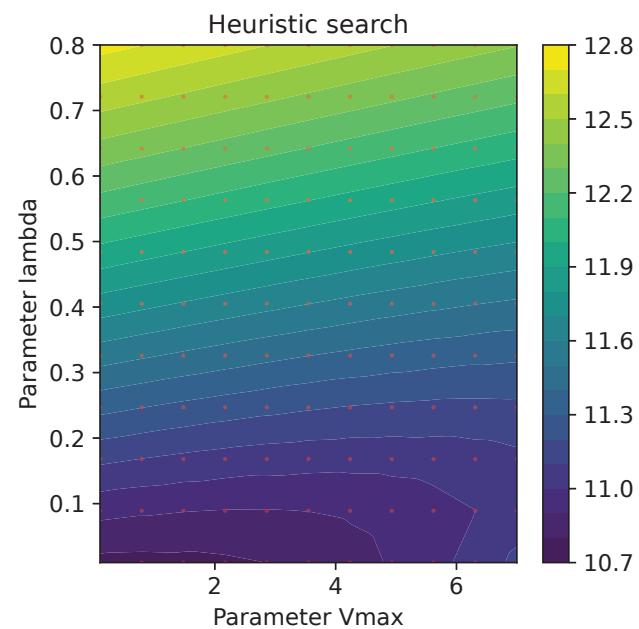
The figure shows a contour plot representing the heuristic step used to determine the initial guess. The sidebar associates colors with the orders of magnitude of the heuristic misfit function evaluations, and the background grid of points identifies the parameters evaluated during the heuristic.

Source: the author.

Figure 6.24 – DFO-LS search on the calibration process of the parameters $V_{\max}$ and λ, in the Case 6 experiment.
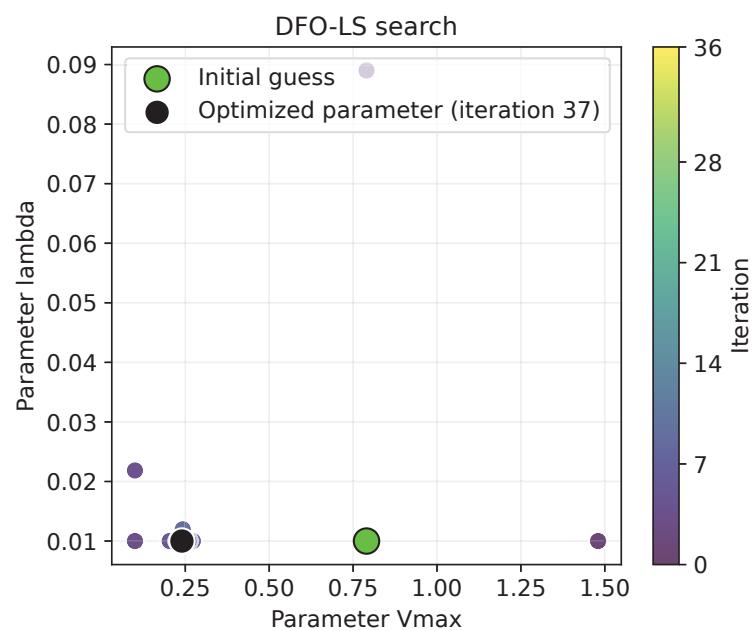
Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic step (Figure 6.23). The sidebar assigns a color scale to the iterations, allowing identification of their convergence pattern.

Source: the author.

Case 7: Calibration of one parameter only. In this experiment, we calibrated only the parameter λ, corresponding to the average daily mortality rate of the entire phytoplankton population in the PEC, as a constant. To do so, we set the parameter $V_{\max}$, the maximum daily growth rate of the phytoplankton population in PEC, as a function of time, $V_{\max}(t)$. This setting was based on sampled data found in the literature, in a manner analogous to the temperature and salinity forcing described previously. More specifically, we assumed the hypothesis that the phytoplankton population in the PEC is dominated by the diatom species *Skeletonema costatum* and *Asterionellopsis glacialis*, which is supported by the data presented in [62]. For both of these species, laboratory data on daily growth rates are available [24, 56], which were considered in setting $V_{\max}(t)$. We consider the equation:

$$C_{\mathrm{PHY}}(t) = C_{\mathrm{PHY},1}(t) + C_{\mathrm{PHY},2}(t) \ , \tag{6.3}$$

where $C_{\mathrm{PHY},1}(t)$ and $C_{\mathrm{PHY},2}(t)$ corresponds to the Clorophill-A concentration associated to *S. costatum* and *A. glacialis*, respectively, in the upper region of PEC at a given time instant t. Considering the data represented in Figure 6.25, we defined a weighting function $w_{\mathrm{PHY}}(t)$ such that:

$$C_{\mathrm{PHY},1}(t) = w_{\mathrm{PHY}} \cdot C_{\mathrm{PHY}}(t) \ , \tag{6.4}$$

$$\text{and} \quad C_{\mathrm{PHY},2}(t) = (1 - w_{\mathrm{PHY}}(t)) \cdot C_{\mathrm{PHY}}(t) \ , \tag{6.5}$$

where $0 \leq w_{\mathrm{PHY}}(t) \leq 1$, at any time instant t. In Figure 6.25, $w_{\mathrm{PHY}}(t)$ is represented as the blue line plot.

Figure 6.25 – Percentage distribution of diatom cell counts in the PEC, when considering only the species *S. costatum* and *A. glacialis*.
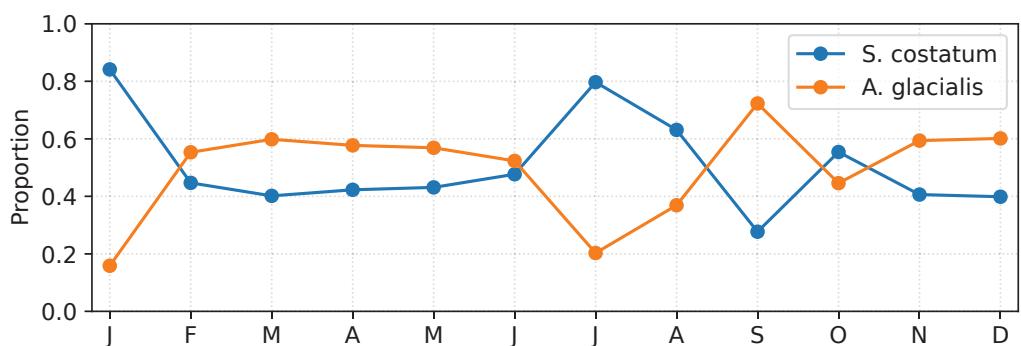
The points correspond to monthly averages, and the line represents daily values computed by linear interpolation of the monthly data.
Source: the author.
Data: Rezende; Brandini (2022).

Figure 6.26 – Daily estimates on the parameter $V_{\mathrm{max}}(t)$.
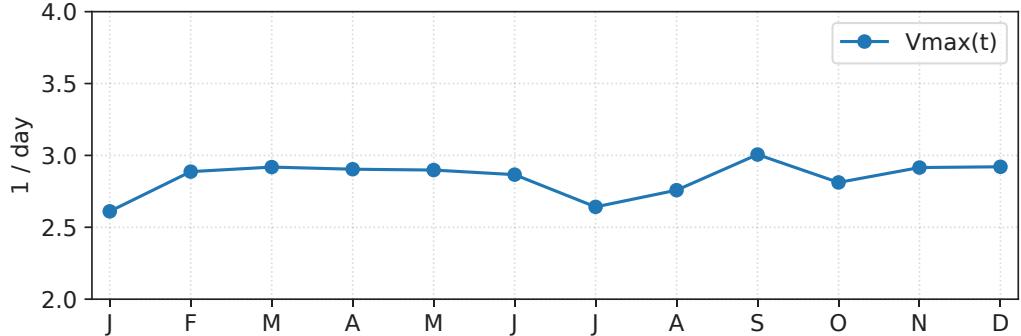
Based on data from Figure 6.25.
Source: the author.

We can summarize the assumptions presented in equations (6.3) - (6.5) by setting $V_{\mathrm{max}}(t)$ (Figure 6.26) as:

$$V_{\mathrm{max}}(t) = w_{\mathrm{PHY}}(t) \cdot V_{\mathrm{max},1} + (1 - w_{\mathrm{PHY}}(t))\ V_{\mathrm{max},2}\ , \tag{6.6}$$

recalling that $V_{\mathrm{max},1}$ and $V_{\mathrm{max},2}$ are constant maximum daily growth rates for the *S. costatum* and *A. glacialis* phytoplankton populations, respectively, and considering a slightly modified version of our model, described by the equations (6.7) - (6.9):

$$\frac{dC_{\mathrm{N}}(t)}{dt} = \frac{0.588\ C_{\mathrm{N}}^{\mathrm{river}}(t) \cdot Q_{\mathrm{river}}(t)}{\mathrm{Vol}_{\mathrm{box}}} + \frac{C_{\mathrm{N}}^{\mathrm{low}}(t) \cdot Q_{\mathrm{ocean}}(t)}{\mathrm{Vol}_{\mathrm{box}}} + r \cdot \lambda \cdot C_{\mathrm{PHY}}(t)$$

$$- \alpha(T(t))\ \beta(S_{\mathrm{up}}(t))\ \frac{V_{\mathrm{max}}(t) \cdot C_{\mathrm{PHY}}(t)\ C_{\mathrm{N}}(t)}{C_{\mathrm{N}}(t) + K} - \frac{C_{\mathrm{N}}(t) \cdot Q_{\mathrm{ebm}}(t)}{\mathrm{Vol}_{\mathrm{box}}}\ . \tag{6.7}$$

$$\frac{dC_{\mathrm{PHY}}(t)}{dt} = C_{\mathrm{PHY}}(t) \cdot \left(\alpha(T(t))\ \beta(S_{\mathrm{up}}(t))\ \frac{V_{\mathrm{max}}(t) \cdot C_{\mathrm{N}}(t)}{C_{\mathrm{N}}(t) + K} - \lambda - \frac{Q_{\mathrm{ebm}}(t)}{\mathrm{Vol}_{\mathrm{box}}} \right) +$$

$$+ \frac{C_{\mathrm{PHY}}^{\mathrm{low}}(t) \cdot Q_{\mathrm{ocean}}(t)}{\mathrm{Vol}_{\mathrm{box}}}\ . \tag{6.8}$$

$$\frac{dC_{\mathrm{N}}(t)}{dt} = 0\ ,\ \frac{dC_{\mathrm{PHY}}(t)}{dt} = 0\ . \tag{6.9}$$

The value obtained by calibrating λ was $\bar{\lambda} = 0.01$. This information may be representative of the average mortality rate of the phytoplankton population in the PEC. However, since the lower limit established for the parameter is reached, it also reinforces the possibility that the estuary's circulation flow is so intense that phytoplankton mortality does not have time to make a relevant impact on phytoplankton concentration in the PEC. For the calibrated value of λ, the fitting obtained from the comparison between observational data and data produced by the model, in relation to concentrations of biochemical tracers in the upper part of the PEC, is presented in figures 6.27 and 6.28.

Figure 6.27 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period, for the Case 7 experiment.
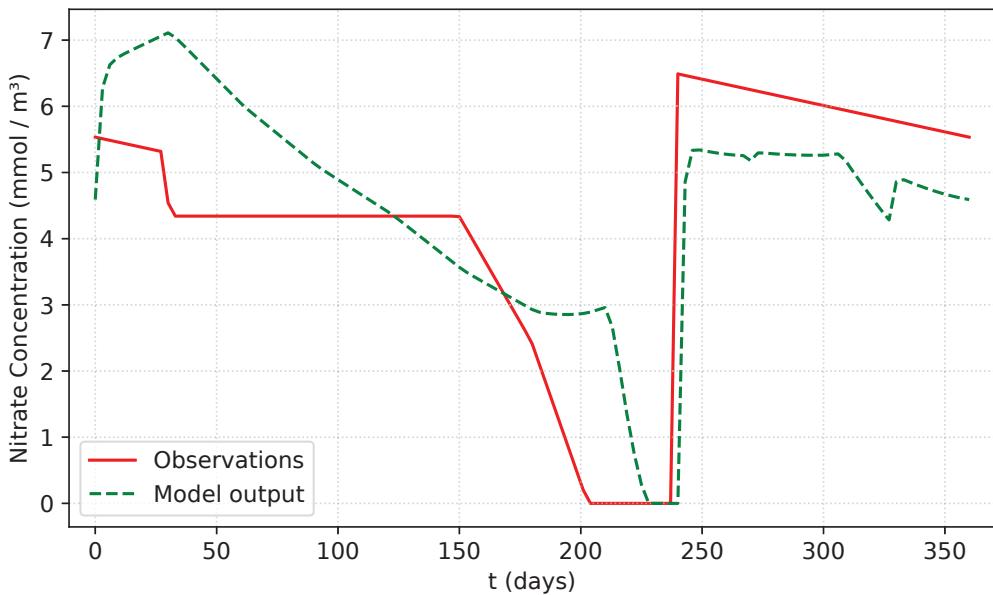
Observations were generated by interpolation of tracers' concentration data (figures 3.12 and 3.13). The model output plot was obtained by integrating the model described by equations (6.7) - (6.9) with the calibrated parameter $\bar{\lambda} = 0.01$. Here we consider $N = 120$ fitting points.
Source: the author.

Unlike the previous experiments in this chapter, in an attempt to obtain maximum precision in this calibration, we considered $N = 120$ observations and two heuristic steps to find a good initial guess, in addition to imposing a ceiling for the misfit function that was calibrated (figures 6.29 and 6.30). The use of increased complexity in the calibration strategy did not guarantee a significant performance improvement, suggesting that in this case the simpler version of the framework, with only one heuristic step, without limiting the ceiling of the misfit function and considering only $N = 12$ observations, was much more advantageous to use, presenting lower computational cost, greater ease of implementation, and similar performance to the more complex version.

Figure 6.28 – Fitting of average phytoplankton concentrations in the upper part of the PEC over a one-year period, for the Case 7 experiment.
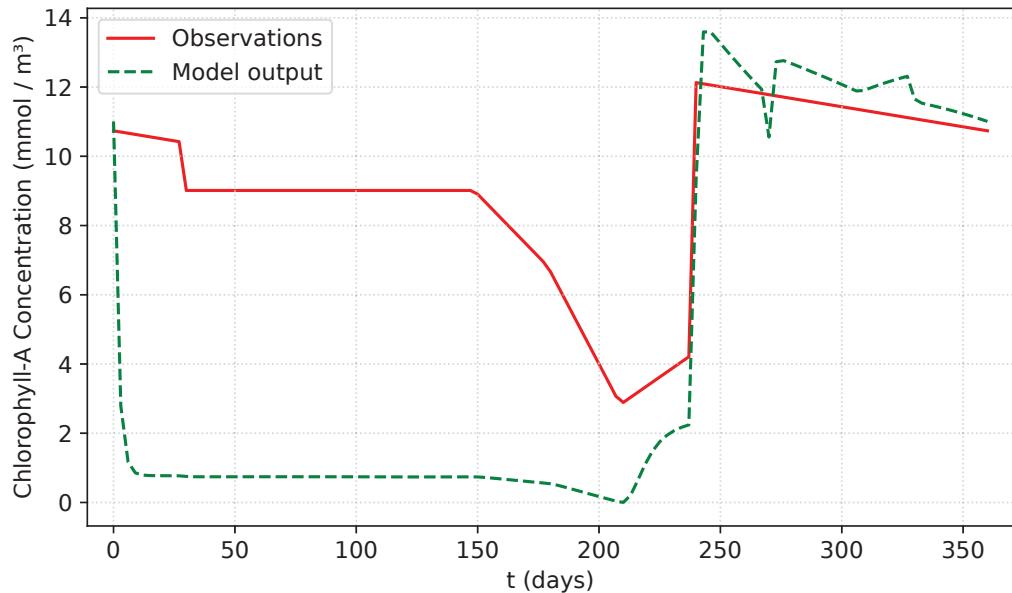
See the full description of the experiment settings in Figure 6.27.

Source: the author.

Figure 6.29 – Heuristic steps on the calibration process of the parameter λ, in the Case 7 experiment.
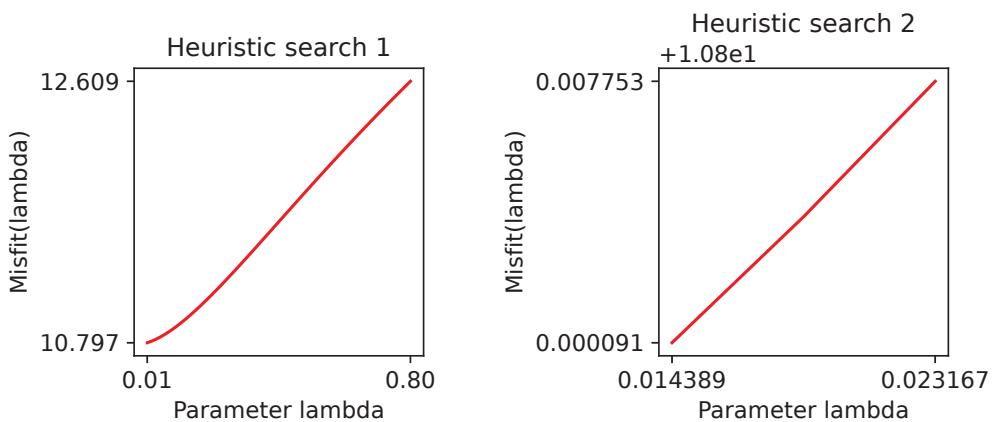
The left-hand and the right-hand plots represent the first and second heuristic steps used to determine the initial guess. The vertical axis shows the values attained by the evaluations of the respective heuristic misfit functions.

Source: the author.

Figure 6.30 – Optimization step on the calibration process of the parameter λ, in the Case 7 experiment.
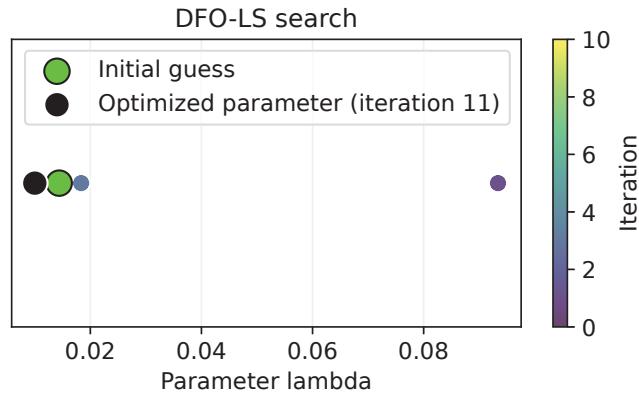
The figure shows the iterations of the parameter search to optimize the fit between model output and observational data for the average tracer concentrations in the upper region of PEC. Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic steps (Figure 6.29). The sidebar assigns a color scale to the iterations, allowing identification of their convergence pattern.

Source: the author.

The data produced by the model, when using the calibrated parameter, presents behavior similar to that of the observational data. This result reinforces the argument that even a conceptual model, when well-calibrated and based on appropriate assumptions, can reproduce the modeled data with some degree of accuracy. It would be valuable to explore the ideal horizon for the integration period used in the heuristic misfit and the residual misfit definitions in each case, as using a smaller interval reduces computational cost, although in some cases it may also impair calibration accuracy. In the benchmark experiments, we observed that the target parameters were accurately recovered, while the fit between the outputs produced by the optimized parameters and those produced by the observational parameters evidenced the applicability of the calibration framework.

## 6.4 Estimating the nitrate input from rivers, rain, and human activities

Initially, we have set the nitrate concentration related to the flow $Q_{\mathrm{river}}(t)$ to be 1.47 times the nitrate concentration in the Nhundiaquara River, $C_{\mathrm{N}}^{\mathrm{river}}(t)$, which was an assumption partially supported by the literature. However, the model's output did not seem to fit the PEC data adequately. To solve this problem, we considered a change in the model and calibrated a positive constant w_{river} that multiplies $C_{N}^{\mathrm{river}}(t)$. We considered only the observational data for phytoplankton (Figure 3.13), and denoted the concentration

observations for phytoplankton in the upper part of PEC at a time instant t by $C_{\text{PHY}}^{\text{up}}(t)$. The mathematical model considered for this experiment is defined by equations (6.10) - (6.11).

$$
\frac{dC_{\text{N}}(t)}{dt} \; = \; \frac{w_{\text{river}} \; C_{\text{N}}^{\text{river}}(t) \cdot Q_{\text{river}}(t)}{\text{Vol}_{\text{box}}} + \frac{C_{\text{N}}^{\text{low}}(t) \cdot Q_{\text{ocean}}(t)}{\text{Vol}_{\text{box}}} + r \cdot \lambda \cdot C_{\text{PHY}}^{\text{up}}(t)
$$

$$
- \, \alpha(T(t)) \; \beta(S_{\text{up}}(t)) \; \frac{V_{\text{max}}(t) \cdot C_{\text{PHY}}^{\text{up}}(t) \; C_{\text{N}}(t)}{C_{\text{N}}(t) + K} - \frac{C_{\text{N}}(t) \cdot Q_{\text{ebm}}(t)}{\text{Vol}_{\text{box}}} \; . \tag{6.10}
$$

$$
\frac{dC_{\text{N}}(t)}{dt} \; = \; 0 \; . \tag{6.11}
$$

During calibration, we noticed a downward trend in the optimized value for parameter w_{river}, but since our idea was not to reduce it too much in order to keep conservative changes, we considered search bounds of 40% to 60% of the previously set value for w_{river}, which was 1.47. The fitting attained after the calibration process is presented in Figure 6.31.

We set the parameters $V_{\text{max}}(t)$, based on the **Case 7** data, and $\lambda = 0.05$, constant. The parameter calibration is presented in figures 6.32 and 6.33, as the two-step heuristic search for an adequate initial guess, which was found as the lowest permitted by the searching bounds. The DFO-LS search converged this same value:

$$
\bar{w}_{\text{river}} = 0.588.
$$

Since we considered the fit obtained to be satisfactory, we considered the result of this calibration for our model presented in *Chapter 3*. This modeling strategy proved useful for extrapolating data in the model based on correlated data available.

Figure 6.31 – Fitting of average nitrate concentrations in the upper part of the PEC over a one-year period.
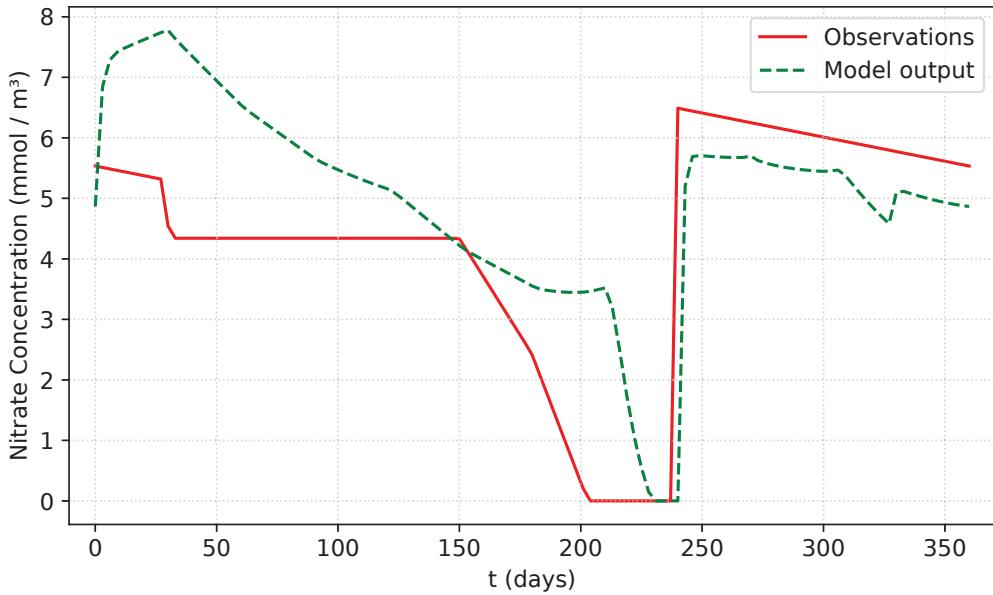
Fitting for the experiment on calibrating the nitrate input on PEC from riverine, pluvial, and human sources. Observations were generated by interpolation of tracers' concentration data (figures 3.12 and 3.13). The model output plot was obtained by integrating the model described by equations (6.10) - (6.11) with the calibrated parameter $\bar{w}_{\text{river}} = 0.588$. Here we consider $N = 120$ fitting points.

Source: the author.

Figure 6.32 – Calibration process of the parameter w_{river}.
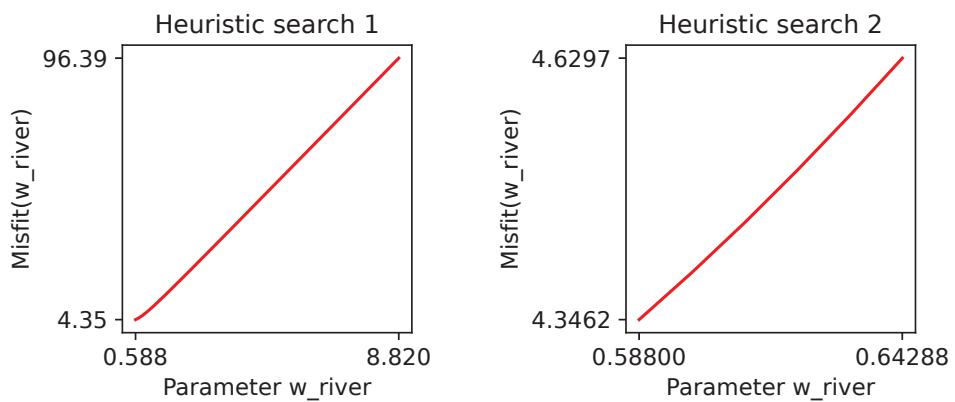
Heuristics for the experiment to calibrate the nitrate input to the PEC from riverine, pluvial, and human sources. The left-hand and the right-hand plots represent the first and second heuristic steps used to determine the initial guess. The vertical axis shows the values attained by the evaluations of the respective heuristic misfit functions.

Source: the author.

Figure 6.33 – Optimization step on the calibration process of the parameter w_{river}.
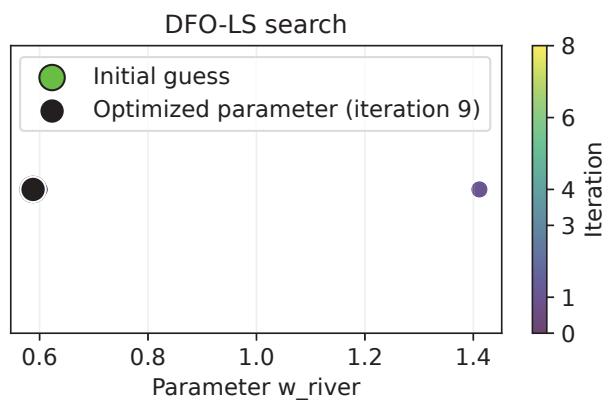
Parameter search for the experiment on calibrating the nitrate input on PEC from riverine, pluvial, and human sources. The figure shows the iterations of the parameter search to optimize the fit between model output and observational data for the average tracer concentrations in the upper region of PEC. Each point corresponds to the parameters obtained in an iteration of the DFO-LS optimization algorithm, being the initial guess obtained from the heuristic steps (Figure 6.32). The sidebar assigns a color scale to the iterations, allowing identification of their convergence pattern.

Source: the author.

7 Conclusions

Summary

This thesis developed a bespoke conceptual biogeochemical model for the PEC, and introduced a data-constrained framework for calibrating its parameters based on a derivative-free optimization method. We explored the tracer-conservation problem and the formulation of a least-squares problem to accommodate the problem features in different scenarios. The framework was exercised in synthetic (twin) experiments and with one year of NO_3 and phytoplankton observations from the PEC, followed by a scenario analysis of increased riverine nitrate load, exemplifying how to apply the same framework for a marine region when data is available.

Implications, Relevance, and Limitations

This work delivers a reproducible formulation of parameter calibration for marine biogeochemical tracers as a weighted least-squares problem, as well as a demonstration on how to use a modern derivative-free optimizer for calibrating a conceptual estuarine model using both synthetic and real data, and a small, documented pipeline – from problem specification to optimization runs and diagnostics – that can be adapted to other systems.

Within the Brazilian context, the study is, to our knowledge, among the first to combine data-constrained calibration of an estuarine biogeochemical model with state-of-the-art derivative-free optimization, with several possibilities for future development. Even though a conceptual model was used, the exercise opens a path for more ambitious national applications as observational datasets and computational resources expand.

The work also clarifies where the approach falls short. The biogeochemical state can be reproduced reasonably well even when individual parameters are not precisely recovered (equifinality), calling for performance metrics and uncertainty analyses that prioritize state trajectories. Temporal and spatial data sparsity – and measurement noise – limit identifiability; simple smoothing helped, but more explicit noise modeling (for example, within DFO-LS) remains to be explored. By design, the conceptual model does not capture all seasonal and spatial heterogeneity of the PEC present in the bibliography, which caps attainable fit and

136

suggests structural extensions (for example, seasonal terms in $V_{\max}$ and λ). Finally, time/CPU constraints limited restarts and methodological variants, which matter in nonconvex settings. Practically, this study lays groundwork for conceptual modeling studies that seek to integrate local data, marine biogeochemical formulations, and modern optimization in a reproducible way.

Future avenues for this work

Although this thesis focused on experimenting with conceptual models, the methodology developed here can be applied, with minor adaptations, to complex, multidimensional settings. Below we exemplify two possible avenues for further work.

A natural next step is to extend the basic framework developed here to a medium-complexity model of the PEC by coupling a (potentially more complex) biogeochemical module to a circulation model that better represents the region's geography, introducing seasonal structure in $V_{\max}$ and λ (harmonics or splines), add additional tracers when available (PO_4, O_2), and exploring the optimization options available, such as noise-aware DFO-LS with multi-start restarts and CMA-ES. Scenario analysis can be expanded (for example, joint changes in riverine NO_3 loads and freshwater discharge) with uncertainty quantification via ensembles of calibrated parameter sets. The anticipated availability of updated data for the PEC in the near future will enable insights into contemporary environmental issues and the impacts of human activities. In fact, historical reconstructions show that human interference on PEC has profoundly altered biogeochemical functioning over the last decades [76]. Together with evidence of eutrophication and shifts in organic matter sources [43, 44], this scenario underscores the need for predictive tools that can explore the effects of various stressors. Incorporating new observational datasets into the current modeling framework is a future aim that will enable more robust scenario testing, such as assessing thresholds for harmful algal blooms [4, 61] or evaluating strategies for mitigating anthropogenic nutrient loads. The combination of updated empirical data with model-based experiments represents a promising approach to enhancing environmental management and informing policy interventions in estuarine environments.

BIBLIOGRAPHY

[1] ALARIE, Stéphane *et al.* Two decades of blackbox optimization applications. **EURO Journal on Computational Optimization**, v. 9, p. 100011, 2021. ISSN 2192-4406. DOI: `https://doi.org/10.1016/j.ejco.2021.100011`. Available from: `https://www.sciencedirect.com/science/article/pii/S2192440621001386`.

[2] ANDRÉS-THIÓ, NICOLAU; MUÑOZ, MARIO ANDRÉS; SMITH-MILES, KATE. METHODOLOGY AND CHALLENGES OF SURROGATE MODELLING METHODS FOR MULTI-FIDELITY EXPENSIVE BLACK-BOX PROBLEMS. **The ANZIAM Journal**, v. 66, n. 1, p. 35–61, 2024. DOI: `10.1017/S1446181124000087`.

[3] BARROSO, S. H.; BECKER, H.; MELO, B. V. M. M. Influence of river discharge on phytoplankton structure and nutrient concentrations in four tropical semiarid estuaries. **Brazilian Journal of Oceanography**, v. 64, p. 37–48, 2016.

[4] BRANDINI, Frederico. Ecological studies in the Bay of Paranaguá: I. horizontal distribution and seasonal dynamics of the phytoplankton. **Boletim do Instituto Oceanográfico**, v. 33, p. 139–147, Dec. 1984. DOI: `10.1590/S0373-55241985000200004`.

[5] CANUEL, Elizabeth A.; HARDISON, Amber K. Sources, Ages, and Alteration of Organic Matter in Estuaries. **Annual Review of Marine Science**, Annual Reviews, v. 8, Volume 8, 2016, p. 409–434, 2016. ISSN 1941-0611. DOI: `https://doi.org/10.1146/annurev-marine-122414-034058`. Available from: `https://www.annualreviews.org/content/journals/10.1146/annurev-marine-122414-034058`.

[6] CARMO CALIJURI, M. do; CUNHA, D.G.F. **Engenharia ambiental: conceitos, tecnologia e gestão**. [*S. l.*]: Elsevier, 2013. ISBN 9788535259544. Available from: `https://books.google.com.br/books?id=pJ_KNAEACAAJ`.

[7] CARTIS, Coralia; ROBERTS, Lindon. A derivative-free Gauss–Newton method. **Mathematical Programming Computation**, v. 11, n. 4, p. 631–674, 2019. ISSN 1867-2957. DOI: `10.1007/s12532-019-00161-7`. Available from: `https://doi.org/10.1007/s12532-019-00161-7`.

[8] CARTIS, Coralia *et al.* **Improving the Flexibility and Robustness of Model-Based Derivative-Free Optimization Solvers**. [*S. l.: s. n.*], 2018. Available from: `https://arxiv.org/abs/1804.00154`.

[9] CATTANI, P. E. **Estimativa do assoreamento sedimentar ao longo da Baía de Paranaguá-PR**. 2009. s. 100. MA thesis – Centro de Estudos do Mar, Setor de Ciências da Terra. Universidade Federal do Paraná.

[10] CLOERN, James. Empirical Model of Skeletonema costatum Photosynthetic Rate, with Applications in the San Francisco Bay Estuary. **Advances in Water Resources - ADV WATER RESOUR**, v. 1, p. 267–274, Sept. 1978. DOI: `10.1016/0309-1708(78)90040-4`.

[11] CUSTÓDIO, A. L. *et al.* Direct Multisearch for Multiobjective Optimization. **SIAM Journal on Optimization**, v. 21, n. 3, p. 1109–1140, 2011. DOI: `10.1137/10079731X`. Available from: `https://doi.org/10.1137/10079731X`.

[12] DAN, Solomon Felix; LIU, Su-Mei; YANG, Bin. Geochemical fractionation, potential bioavailability and ecological risk of phosphorus in surface sediments of the Cross River estuary system and adjacent shelf, South East Nigeria (West Africa). **Journal of Marine Systems**, v. 201, p. 103244, 2020. ISSN 0924-7963. DOI: `https://doi.org/10.1016/j.jmarsys.2019.103244`. Available from: `https://www.sciencedirect.com/science/article/pii/S0924796319303811`.

[13] DAVIDSON, K. *et al.* An investigation of non-steady-state algal growth. I. An experimental model ecosystem. **Journal of Plankton Research**, v. 21, n. 5, p. 811–837, May 1999. ISSN 0142-7873. DOI: `10.1093/plankt/21.5.811`. Available from: `https://academic.oup.com/plankt/article-pdf/21/5/811/9913120/210811.pdf`.

[14] EVANS, Geoffrey T.; PARSLOW, John S. A Model of Annual Plankton Cycles. **Biological Oceanography**, v. 3, n. 3, p. 327–347, 1985. DOI: `https://doi.org/10.1080/01965581.1985.10749478`.

[15] GARRISON, T. **Fundamentos de oceanografia**. [*S. l.*]: Cengage Learning, 2010. ISBN 9788522106776. Available from: `https://books.google.com.br/books?id=19OHPgAACAAJ`.

[16] GIANESELLA, S.M.F.; SALDANHA-CORRÊA, F. M.P. Oceanos e áreas costeiras. *In:* ENGENHARIA ambiental: conceitos, tecnologia e gestão. Rio de Janeiro: Elsevier, 2013. p. 789.

[17] GRUBER, N.; DONEY, S. C. Modeling of ocean biogeochemistry and ecology. v. 5, p. 547–560, 2019.

[18] HANSEN, Nikolaus; OSTERMEIER, Andreas. Completely Derandomized Self-Adaptation in Evolution Strategies. **Evolutionary Computation**, v. 9, p. 159–195, June 2001. DOI: `10.1162/106365601750190398`.

[19] HANSEN, Nikolaus *et al.* Real-Parameter Black-Box Optimization Benchmarking 2009: Noisy Functions Definitions, Jan. 2009.

[20] HARRIS, C. Coupled Atmosphere-Ocean Modelling. *In:* NEW Frontiers In Operational Oceanography. [*S. l.*]: GODAE OceanView, 2018. chap. 16, p. 445–463. Retrieved from Florida State University Repository. Available from: `http://purl.flvc.org/fsu/fd/FSU_libsubv1_scholarship_submission_1536243251_5afbe0aa`.

[21] INPE/LABREN. **Atlas Brasileiro de Energia Solar (2ª edição – 2017): Base de dados**. [*S. l.: s. n.*], 2017. Catálogo BIG/INPE. Médias mensais e anuais do total diário de irradiação. Available from: `https://data.inpe.br/geonetwork/srv/api/records/282a3d34-5a7a-46c8-b10b-333a2c813ed9`.

[22] IOCCG. **Guide to the Implementation of Ocean Color Remote Sensing**. [*S. l.*], 2014.

[23] KELLER, David P. *et al.* Southern Ocean biological impacts on global ocean oxygen. **Geophysical Research Letters**, v. 43, n. 12, p. 6469–6477, 2016. DOI: `https://doi.org/10.1002/2016GL069630`. Available from: `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL069630`.

[24] KHAN, Saleha *et al.* Physiological observations on a diatom Skeletonema costatum (Greville) Cleve. **Bangladesh Journal of Fisheries Research**, v. 2, n. 2, p. 109–118, Jan. 1998.

[25] KHATIWALA, Samar. A computational framework for simulation of biogeochemical tracers in the ocean. **Glob. Biogeochem. Cycles**, v. 21, Sept. 2007. DOI: 10.1029/2007GB002923.

[26] KHATIWALA, Samar. Fast spin up of Ocean biogeochemical models using matrix-free Newton–Krylov. **Ocean Modelling**, v. 23, n. 3, p. 121–129, 2008. ISSN 1463-5003. DOI: https://doi.org/10.1016/j.ocemod.2008.05.002. Available from: https://www.sciencedirect.com/science/article/pii/S1463500308000589.

[27] KRIEST, I. Calibration of a simple and a complex model of global marine biogeochemistry. **Biogeosciences**, v. 14, n. 21, p. 4965–4984, 2017. DOI: 10.5194/bg-14-4965-2017. Available from: https://bg.copernicus.org/articles/14/4965/2017/.

[28] KRIEST, I.; KHATIWALA, S.; OSCHLIES, A. Towards an assessment of simple global marine biogeochemical models of different complexity. **Progress in Oceanography**, v. 86, n. 3, p. 337–360, 2010. ISSN 0079-6611. DOI: https://doi.org/10.1016/j.pocean.2010.05.002. Available from: https://www.sciencedirect.com/science/article/pii/S0079661110000844.

[29] KRIEST, I.; OSCHLIES, A. Swept under the carpet: organic matter burial decreases global ocean biogeochemical model sensitivity to remineralization length scale. **Biogeosciences**, v. 10, n. 12, p. 8401–8422, 2013. DOI: 10.5194/bg-10-8401-2013. Available from: https://bg.copernicus.org/articles/10/8401/2013/.

[30] KRIEST, I.; OSCHLIES, A. MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes. **Geoscientific Model Development**, v. 8, n. 9, p. 2929–2957, 2015. DOI: 10.5194/gmd-8-2929-2015. Available from: https://gmd.copernicus.org/articles/8/2929/2015/.

[31] KRIEST, I.; OSCHLIES, A.; KHATIWALA, S. Sensitivity analysis of simple global marine biogeochemical models. **Global Biogeochemical Cycles**, v. 26, n. 2, 2012. DOI: https://doi.org/10.1029/2011GB004072. Available from: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GB004072.

[32] KRIEST, I. *et al.* Calibrating a global three-dimensional biogeochemical ocean model (MOPS-1.0). **Geoscientific Model Development**, v. 10, n. 1, p. 127–154, 2017. DOI: 10.5194/gmd-10-127-2017. Available from: https://gmd.copernicus.org/articles/10/127/2017/.

[33] KRIEST, I. *et al.* One size fits all? Calibrating an ocean biogeochemistry model for different circulations. **Biogeosciences**, v. 17, n. 12, p. 3057–3082, 2020. DOI: 10.5194/bg-17-3057-2020. Available from: https://bg.copernicus.org/articles/17/3057/2020/.

[34] KVALE, K. F. *et al.* Evaluation of the transport matrix method for simulation of ocean biogeochemical tracers. **Geoscientific Model Development**, v. 10, n. 6, p. 2425–2445, 2017. DOI: 10.5194/gmd-10-2425-2017. Available from: https://gmd.copernicus.org/articles/10/2425/2017/.

[35] KWIATKOWSKI, L. *et al.* iMarNet: an ocean biogeochemistry model intercomparison project within a common physical ocean modelling framework. **Biogeosciences**, v. 11, n. 24, p. 7291–7304, 2014. DOI: `10.5194/bg-11-7291-2014`. Available from: `https://bg.copernicus.org/articles/11/7291/2014/`.

[36] LALLI, C.; PARSONS, T.R. **Biological Oceanography: An Introduction**. [*S. l.*]: Butterworth-Heinemann, 1997. [Open University oceanography series]. ISBN 9780080527994. Available from: `https://books.google.com.br/books?id=c6J5hlcjFaAC`.

[37] LAMOUR, M. R.; ANGULO, R. J.; SOARES, C. R. Bathymetrical Evolution of Critical Shoaling Sectors on Galheta Channel, Navigable Access to Paranaguá Bay, Brazil. **Journal of Coastal Research**, Coastal Education & Research Foundation, Inc., v. 23, n. 1, p. 49–58, 2007. ISSN 07490208, 15515036. Available from: `http://www.jstor.org/stable/4300405`. Visited on: 7 Nov. 2025.

[38] LANA, Paulo *et al.* The Subtropical Estuarine Complex of Paranagua Bay, Brazil. *In:* [*s. l.*]: Springer, Jan. 2001. v. 144, p. 131–145. ISBN 978-3-642-08657-1. DOI: `10.1007/978-3-662-04482-7_11`.

[39] LARSON, Jeffrey; MENICKELLY, Matt; WILD, Stefan M. Derivative-Free Optimization Methods. **Acta Numerica**, v. 28, p. 287–404, 2019. DOI: `10.1017/S0962492919000060`.

[40] MACHADO, Eunice *et al.* Temporal and spatial dynamics of nutrients and particulate suspended matter in Paranaguá Bay, PR, Brazil. **Nerítica**, v. 11, p. 17–36, Jan. 1997. DOI: `10.5380/rn.v11i1-2.85136`.

[41] MARONE, E. *et al.* Land-ocean fluxes in the Paranaguá Bay estuarine system, southern Brazil. **Brazilian Journal of Oceanography**, v. 53, n. 3–4, p. 169–181, 2005.

[42] MARTIN, Adrian P. *et al.* When to add a new process to a model – and when not: A marine biogeochemical perspective. **Ecological Modelling**, v. 498, p. 110870, 2024. ISSN 0304-3800. DOI: `https://doi.org/10.1016/j.ecolmodel.2024.110870`. Available from: `https://www.sciencedirect.com/science/article/pii/S0304380024002588`.

[43] MARTINS, César *et al.* Coupling spectroscopic and chromatographic techniques for evaluation of the depositional history of hydrocarbons in a subtropical estuary. **Environmental pollution (Barking, Essex : 1987)**, v. 205, p. 403–414, July 2015. DOI: `10.1016/j.envpol.2015.07.016`.

[44] MARTINS, César C. *et al.* Anthropogenic organic matter inputs indicated by sedimentary fecal steroids in a large South American tropical estuary (Paranaguá estuarine system, Brazil). **Marine Pollution Bulletin**, v. 60, n. 11, p. 2137–2143, 2010. ISSN 0025-326X. DOI: `https://doi.org/10.1016/j.marpolbul.2010.07.027`. Available from: `https://www.sciencedirect.com/science/article/pii/S0025326X10003334`.

[45] MELO VIRÍSSIMO, Francisco de; MARTIN, Adrian P.; HENSON, Stephanie A. Influence of Seasonal Variability in Flux Attenuation on Global Organic Carbon Fluxes and Nutrient Distributions. **Global Biogeochemical Cycles**, v. 36, n. 2, e2021gb007101, 2022. e2021GB007101 2021GB007101. DOI: `https://doi.org/10.1029/2021GB007101`. Available from: `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GB007101`.

[46] MELO VIRÍSSIMO, Francisco de; STAINFORTH, David A. Micro- and Macroparametric Uncertainty in Climate Change Prediction: A Large Ensemble Perspective. **Bulletin of the American**

Meteorological Society, American Meteorological Society, Boston MA, USA, v. 106, n. 7, e1319–e1341, 2025. DOI: `10.1175/BAMS-D-24-0064.1`. Available from: `https://journals.ametsoc.org/view/journals/bams/106/7/BAMS-D-24-0064.1.xml`.

[47] MELO VIRÍSSIMO, Francisco de *et al.* Seasonality in Carbon Flux Attenuation Explains Spatial Variability in Transfer Efficiency. **Geophysical Research Letters**, v. 51, n. 4, e2023gl107050, 2024. e2023GL107050 2023GL107050. DOI: `https://doi.org/10.1029/2023GL107050`. Available from: `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL107050`.

[48] MENDEN-DEUER, Susanne; LESSARD, Evelyn. Menden-Deuer S, Lessard EJ.. Carbon to volume relationships for dinoflagellates, diatoms, and other protest plankton. Limnol Oceanogr 45: 569-579. **Limnology and oceanography**, v. 45, p. 569–579, May 2000. DOI: `10.4319/lo.2000.45.3.0569`.

[49] MIDDELBURG, Jack. **Marine Carbon Biogeochemistry: A Primer for Earth System Scientists**. [*S. l.*]: SpringerBriefs in Earth System Sciences, Jan. 2019. ISBN 978-3-030-10822-9. DOI: `10.1007/978-3-030-10822-9`.

[50] MIZERKOWSKI, Byanka Damian *et al.* Sources, loads and dispersion of dissolved inorganic nutrients in Paranaguá Bay. **Ocean Dynamics**, v. 62, n. 10, p. 1409–1424, 2012. DOI: `10.1007/s10236-012-0569-x`. Available from: `https://doi.org/10.1007/s10236-012-0569-x`.

[51] MUELLER, Juliane *et al.* CH4 parameter estimation in CLM4. 5bgc using surrogate global optimization. **Geoscientific Model Development Discussions**, v. 8, Jan. 2015. DOI: `10.5194/gmdd-8-141-2015`.

[52] MÜLLER, Juliane; DAY, Marcus. Surrogate Optimization of Computationally Expensive Black-Box Problems with Hidden Constraints. **INFORMS Journal on Computing**, v. 31, n. 4, p. 689–702, 2019. DOI: `10.1287/ijoc.2018.0864`. Available from: `https://doi.org/10.1287/ijoc.2018.0864`.

[53] NIEMEYER, D.; KRIEST, I.; OSCHLIES, A. The effect of marine aggregate parameterisations on nutrients and oxygen minimum zones in a global biogeochemical model. **Biogeosciences**, v. 16, n. 15, p. 3095–3111, 2019. DOI: `10.5194/bg-16-3095-2019`. Available from: `https://bg.copernicus.org/articles/16/3095/2019/`.

[54] NOCEDAL, Jorge; WRIGHT, Stephen. **Numerical Optimization**. [*S. l.: s. n.*], Jan. 2006. ISBN 978-0-387-30303-1. DOI: `10.1007/978-0-387-40065-5`.

[55] OLIVER, S. *et al.* A derivative-free optimisation method for global ocean biogeochemical models. **Geoscientific Model Development**, v. 15, n. 9, p. 3537–3554, 2022. DOI: `10.5194/gmd-15-3537-2022`. Available from: `https://gmd.copernicus.org/articles/15/3537/2022/`.

[56] OLMSTEAD, Jason. **A Pilot Study on Growth Rates of *Asterionellopsis glacialis*, *Thalassiosira* sp., and *Chaetoceros* sp. for Potential Biofuel Production**. Friday Harbor, WA, May 2011. [Friday Harbor Laboratories Student Research Papers]. ZooBot Research Apprenticeship BIOL 479, Spring 2011. Available from: `https://digital.lib.washington.edu/researchworks/handle/1773/26608`. Visited on: 12 Aug. 2025.

[57] OSCHLIES, A.; LENZ, J. Biogeochemical modeling of the ocean: A review. **Reviews of Geophysics**, v. 48, p. 1–63, 2010.

[58] PARSONS, T. R.; MAITA, Y.; LALLI, C. M. **A Manual of Chemical and Biological Methods for Seawater Analysis**. Oxford: Pergamon Press, 1984.

[59] PASSOS, Ana Carolina dos *et al.* Fishes of Paranaguá Estuarine Complex, South West Atlantic. **Biota Neotropica**, v. 12, n. 3, Sept. 2012. Available from: `https://www.biotaneotropica.org.br/BN/article/view/1002`.

[60] PÉREZ-FERNÁNDEZ, Begoña; VIÑAS, Lucía; BESADA, Victoria. Concentrations of organic and inorganic pollutants in four Iberian estuaries, North Eastern Atlantic. Study of benchmark values estimation. **Marine Chemistry**, v. 224, p. 103828, 2020. ISSN 0304-4203. DOI: `https://doi.org/10.1016/j.marchem.2020.103828`. Available from: `https://www.sciencedirect.com/science/article/pii/S0304420320300827`.

[61] PROCOPIAK, Leticia; FERNANDES, Luciano; MOREIRA-FILHO, Hermes. Diatomáceas (Bacillariophyta) marinhas e estuarinas do Paraná, Sul do Brasil: lista de espécies com ênfase em espécies nocivas. **Biota Neotropica**, v. 6, Jan. 2006. DOI: `10.1590/S1676-06032006000300013`.

[62] REZENDE, Kátia; BRANDINI, Frederico. Variação sazonal do fitoplâncton na zona de arrebentação da praia de Pontal do Sul (Paranaguá - Paraná)Variação sazonal do fitoplâncton na zona de arrebentação da praia de Pontal do Sul (Paranaguá - Paraná). **Revista Nerítica**, v. 11, p. 49–62, Mar. 2022. DOI: `10.5380/rn.v11i1-2.85138`.

[63] RIOS, Luis; SAHINIDIS, Nikolaos. Derivative-free optimization: A review of algorithms and comparison of software implementations. **Journal of Global Optimization**, v. 56, Nov. 2009. DOI: `10.1007/s10898-012-9951-y`.

[64] ROBERTS, Lindon. **DFO-LS: Derivative-Free Optimizer for Least-Squares Minimization**. [*S. l.*]: Numerical Algorithms Group (NAG), 2025. `https://numericalalgorithmsgroup.github.io/dfols/build/html/index.html`.

[65] RODRIGUES, Dulce; MENDIONDO, Eduardo. Bacias hidrográficas: caracterização e manejo sustentável. *In:* [*s. l.*]: Elsevier, Mar. 2013. p. 47–74. ISBN 978-85-352-5954-4.

[66] ROHATGI, Ankit. **WebPlotDigitizer**. [*S. l.: s. n.*], 2025. `https://automeris.io/WebPlotDigitizer`. Version 5.2, accessed 2025-07-10.

[67] SARMIENTO, J. L.; GRUBER, N. **Ocean Biogeochemical Dynamics**. [*S. l.*]: Princeton University Press, 2006.

[68] SOUZA, M. M. **Modelagem computacional do Complexo Estuarino de Paranaguá sob a influência de ondas, marés e descarga fluvial**. 2015. s. 41. MA thesis – Centro de Estudos do Mar, Setor de Ciências da Terra. Universidade Federal do Paraná.

[69] STROGATZ, Steven H. **Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering**. 2nd. Boulder, CO: CRC Press, 2018. ISBN 978-0-8133-4910-7.

[70] TEAM, MITgcm Development. **MITgcm: A Community Ocean Model**. [*S. l.: s. n.*], 2023. Available from: `https://mitgcm.readthedocs.io/en/latest/`.

[71] TERRA (IAT), Instituto Água e. **Sistemas de informações hidrológicas**. [*S. l.: s. n.*], 2020. Available from: `http://www.sih-web.aguasparana.pr.gov.br/sih-web/gerarRelatorioVazoesFluviometricasMensais.do?action=carregarInterfaceInicial`.

[72] UNESCO. **Protocols for the Joint Global Ocean Flux Study (JGOFS)**. [*S. l.*], 1994. [IOC Manuals and Guides, 29].

[73] VERRI, Giorgia *et al.* A box model to represent estuarine dynamics in mesoscale resolution ocean models. **Ocean Modelling**, v. 148, p. 101587, 2020. ISSN 1463-5003. DOI: `https://doi.org/10.1016/j.ocemod.2020.101587`. Available from: `https://www.sciencedirect.com/science/article/pii/S1463500319300642`.

[74] VU, Ky Khac *et al.* Surrogate-based methods for black-box optimization. **International Transactions in Operational Research**, v. 24, n. 3, p. 393–424, 2017. DOI: `https://doi.org/10.1111/itor.12292`. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1111/itor.12292`.

[75] WILD, Stefan M; SARICH, Jason; SCHUNCK, Nicolas. Derivative-free optimization for parameter estimation in computational nuclear physics. **Journal of Physics G: Nuclear and Particle Physics**, IOP Publishing, v. 42, n. 3, p. 034031, Feb. 2015. DOI: `10.1088/0954-3899/42/3/034031`. Available from: `https://doi.org/10.1088/0954-3899/42/3/034031`.

[76] WILHELM, Marines Maria. **Panorama histórico das mudanças ambientais no complexo estuarino de Paranaguá, PR, Brasil: uma abordagem geoquímica para interpretar as mudanças ambientais recentes**. 2023. PhD thesis – Universidade Federal do Paraná, Curitiba, PR. Tese (Doutorado) – Programa de Pós-Graduação em Sistemas Costeiros e Oceânicos. Available from: `https://hdl.handle.net/1884/87412`.

[77] WILSON, Jamie D. *et al.* The biological carbon pump in CMIP6 models: 21st century trends and uncertainties. **Proceedings of the National Academy of Sciences**, v. 119, n. 29, e2204369119, 2022. DOI: `10.1073/pnas.2204369119`. Available from: `https://www.pnas.org/doi/pdf/10.1073/pnas.2204369119`.

[78] ZEHR, Jonathan P.; WARD, Bess B. Nitrogen Cycling in the Ocean: New Perspectives on Processes and Paradigms. **Applied and Environmental Microbiology**, v. 68, n. 3, p. 1015–1024, 2002. DOI: `10.1128/AEM.68.3.1015-1024.2002`. Available from: `https://journals.asm.org/doi/pdf/10.1128/aem.68.3.1015-1024.2002`.

[79] ZEM, R. C. **Hidrodinâmica e fluxo de material particulado em suspensão na região estuarina da Baia de Paranaguá – PR**. 2008. s. 155. MA thesis – Universidade Federal Fluminense.

APPENDIX – Supplementary Materials for Computational Implementation and Experiments

This appendix complements the references made throughout the thesis, gathering all supplementary materials related to the computational experiments conducted here. It includes: (i) the organization of the source code, (ii) implementation details of the mathematical models, (iii) instructions for executing the experiments, (iv) the DFO-LS optimization method used for parameter calibration, and (v) practical examples of applying the DFO-LS method to the models.

A.1 Repository structure and code organization

To ensure reproducibility and transparency, all source code used in this thesis is publicly available. The repositories are logically organized as follows:

- /**Optimization**/ – Functions associated with the DFO-LS method presented in *Chapter 4*.

- /**Experiments**/ – Scripts used in the computational experiments discussed in *Chapter 5*.

- /**PECmodel**/ – Implementation of the mathematical models described in *Chapter 2*, including differential equations and auxiliary functions, files related to the PEC model introduced in *Chapter 3*, and calibration experiments for the estuary model used in *Chapter 6*.

All repositories are available at:

$$\text{https://github.com/leticiabecher/ThesisExperiments}$$

A.2 Computational environment and execution procedures

The experiments described in chapters 2 and 5 were conducted under the following computational environment:

- **Python**: version 3.10 or newer.

- **Libraries**: NumPy, SciPy, Matplotlib, Pandas.

- **Operating system**: Linux or Windows with Python support.

To run any experiment, clone the repository and execute:

```
python experiment_name.py
```

Each folder includes further explanations inside its respective `README.md` file.

A.3 Implementation of the mathematical models

The mathematical models described in the main chapters were fully implemented in Python. The differential equations from *Chapter 2* were solved using SciPy's `solve_ivp` integrator.

Example implementation snippet:

```python
def model_equations(t, y, params):
    k1, k2, vmax = params
    dydt = [
        k1 * y[0] - k2 * y[1],
        vmax * y[1] / (0.1 + y[1])
    ]
    return dydt
```

A.4 The DFO-LS optimization method

The DFO-LS method described in *Chapter 4* is a derivative-free optimization algorithm for least-squares problems. The adapted pseudocode is shown in Algorithm 1.

Algorithm 1 DFO-LS Method adapted from [55]

1: **Input:** number of parameters n, initial point θ_0, initial trust region radius Δ_0, parameter bounds, and maximum number of function evaluations.
2: Construct the initial interpolation set Y_0 by evaluating the misfit function at $n + 1$ points.
3: **for** each iteration **do**
4:    Build a local quadratic model of the misfit.
5:    Solve the trust-region subproblem.
6:    Evaluate the candidate point.
7:    Update the interpolation set.
8:    Adjust the trust-region radius.
9: **end for**
10: **Output:** estimated calibrated parameters.

Examples and explanations of the DFO-LS Python solver usage, including basic and advanced settings, can be found in [64].

A.5 Calibration framework tests

In *Chapter 5*, several calibration examples are presented. The complete scripts are available at:

https://github.com/leticiabecher/ThesisExperiments/tree/main/Framework_testing

A.6 PEC model experiments

The PEC model introduced in *Chapter 3* was calibrated using the scripts contained in the PEC directory. The materials include:

- differential equations of the PEC model;

- parameter sets used in calibration experiments;

- optimization scripts;

- result validation codes and visualizations.

Full implementations are located at:

    https://github.com/leticiabecher/ThesisExperiments/tree/main/PEC_experiments

A.7 Estuary calibration experiments

The experiments described in *Chapter 6* use real environmental data and apply the DFO-LS method for parameter estimation.

The supplementary files include:

- data preprocessing routines;

- the estuary dynamical model;

- the misfit function for calibration;

- experiment execution scripts and visualization tools.

Complete code available at:

    https://github.com/leticiabecher/ThesisExperiments/tree/main/PEC_experiments

This appendix supports and documents all computational methods and experiments performed in the thesis, ensuring full reproducibility of the presented results.