Development of a diffusion kernel density estimator and application on marine carbon-13 isotope data

M.Sc. M.Sc. Maria-Theresia Pelz
aus
Kiel

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel
eingereicht im Jahr 2023

Kiel Computer Science Series (KCSS) 2023/4 dated 2023-06-06

ISSN 2193-6781 (print version)
ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via https://www.informatik.uni-kiel.de/kcss

The author can be contacted via https://www.linkedin.com/in/maria-theresia-pelz-07511327a

Published by the Department of Computer Science, Kiel University

Algorithmic Optimal Control – CO_2-Uptake of the Ocean

Please cite as:

▻ Maria-Theresia Pelz. *Development of a diffusion kernel density estimator and application on marine carbon-13 isotope data* Number 2023/4 in Kiel Computer Science Series. Department of Computer Science, 2023. Dissertation, Faculty of Engineering, Kiel University.

```
@book{PelzDissertation,
  author    = {Maria-Theresia Pelz},
  title     = {Development of a diffusion kernel density estimator
               and application on marine carbon-13 isotope data},
  publisher = {Department of Computer Science, Kiel University},
  year      = {2023},
  number    = {2023/4},
  doi       = {10.21941/kcss/2023/4},
  series    = {Kiel Computer Science Series},
  note      = {Dissertation, Faculty of Engineering,
               Kiel University.}
}
```

About this Series

The Kiel Computer Science Series (KCSS) covers dissertations, habilitation theses, lecture notes, textbooks, surveys, collections, handbooks, etc. written at the Department of Computer Science at Kiel University. It was initiated in 2011 to support authors in the dissemination of their work in electronic and printed form, without restricting their rights to their work. The series provides a unified appearance and aims at high-quality typography. The KCSS is an open access series; all series titles are electronically available free of charge at the department's website. In addition, authors are encouraged to make printed copies available at a reasonable price, typically with a print-on-demand service.

Please visit `http://www.informatik.uni-kiel.de/kcss` for more information, for instructions how to publish in the KCSS, and for access to all existing publications.

1. Gutachter:	Prof. Dr. Thomas Slawig Christian-Albrechts-Universität Kiel
2. Gutachter:	Prof. Dr. Andreas Oschlies GEOMAR – Helmholtz-Zentrum für Ozeanforschung Kiel

Datum der mündlichen Prüfung: 16. Mai 2023

Zusammenfassung

Ziel meiner Arbeit war die Entwicklung eines Kerndichteschätzers mit guter Auflösung typischer Strukturen in Wahrscheinlichkeitsdichten von Ozeandaten am Beispiel organischer Kohlenstoffisotopendaten ($\delta^{13}C_{POC}$) innerhalb des neu entstehenden Felds Marine Data Science. Klassische Datenwissenschaften, ein allgemeines Verständnis der Ozeanforschung, Kommunikationsfähigkeit und Selbstsicherheit sind grundlegende Anforderungen an Wissenschaffende der Marine Data Science.

Organische Kohlenstoffisotopendaten waren zu Beginn meiner Arbeit mit etwa 500 Datenpunkten global verfügbar. Ich habe den vorhandenen Datensatz in einer ersten Version auf zunächst 4732 Datenpunkte erweitert, in einer zweiten auf 6952. Beide sind bei PANGAEA veröffentlicht zusammen mit Metainformationen wie Messort, -zeit und -methode und Interpolationen. Eine Beschreibung der zeitliche und geographische Verteilung der ersten Version habe ich bei Earth System Science Data veröffentlicht.

Die Entwicklung des Kerndichteschätzers habe ich auf die existierende Idee, ihn als Lösung der Diffusionsgleiung zu berechnen, aufgebaut. Mein Algorithmus nutzt finite Differenzen im Ort und equidistante Zeitschritte mit einem impliziten Euler-Verfahren und approximiert den optimalen Glättungsparameter durch zwei Pilot-Schritte. Im Vergleich zu anderen bekannten Kerndichteschätzern erzeugt mein Algorithmus verlässliche Approximationen von multimodalen und Rand-nahen Verteilungen auf künstlichen und realen Ozeandaten und ist robust gegenüber Rauschen. Meine Implementierung ist als Python-Paket auf Zenodo veröffentlicht, ihre Beschreibung bei Geoscientific Model Development eingereicht.

Ich konnte in meiner Arbeit zeigen, dass mein Kerndichteschätzer zuverlässig Ozeandaten auswertet und damit eine Grundlage zur verbesserten Kalibrierung von Erdsystemmodellen legt. Gleichzeitig konnte ich zur Definition und Etablierung des feldes Marine Data Science beitragen.

Summary

My work developed a kernel density estimator that well resolves typical structures of probability densities, which was demonstrated on a newly compiled marine data set of organic carbon-13 isotope ratios ($\delta^{13}C_{POC}$). All work was conducted within the emerging field of marine data science. I identified classical data science, a general understanding of ocean science, communication skills, and confidence as requirements for marine data scientists.

In the beginning of my work, the existing $\delta^{13}C_{POC}$ data consisted of about 500 data points in the global ocean. I expanded the existing data set to 4732 data points in a first version, and additionally to 6952 in a second. Both are published at PANGAEA along with meta information such as measurement location, time, and method, and interpolations. I have published a description of the temporal and geographic distribution of the first version at Earth System Science Data.

I designed the development of the kernel density estimator algorithm on the existing concept of computing it as a solution of the diffusion equation. My algorithm uses finite differences in space and equidistant time steps with an implicit Euler method, and approximates the optimal smoothing parameter by two pilot steps. Compared to other well-known kernel density estimators, my algorithm produces reliable approximations of multimodal and boundary-close distributions on artificial and real marine data and is robust to noise. My implementation is published as a Python package on Zenodo, its description is submitted to Geoscientific Model Development.

I was able to show that my kernel density estimator reliably evaluates ocean data and thus lays a foundation for calibrating Earth system models. At the same time, I was able to contribute to the definition and establishment of the field of Marine Data Science.

Acknowledgements

I want to thank my supervisors **Thomas Slawig**, **Andreas Oschlies**, **Markus Schartau** and **Christopher Somes** for teaching and mentoring me throughout my research and development as a marine data scientist.

I want to thank my coauthors **Carola Trahms** and **Vanessa Lampe** for their priceless input and encouragement.

I want to thank my colleagues **Jane Eitzen**, **Matthias Vetter**, **Monika Peschke**, **Hela Mertens**, **Avan Antja** and **Peer Kröger** for always offering me help of any kind.

I want to thank the **Kiel University**, the **GEOMAR** – Helmholtz Centre for Ocean Research, my graduate school **MarDATA** – Helmholtz School for Marine Data Science and especially my coordinator **Enno Prigge** and my fellow student representative **Martin Prinzler** for creating such an inspiring and lively research environment.

I want to thank my family **Mechthild Verwega**, **Leon Domurath**, **Sheila Klink**, **Heinz Pelz**, **Herr Lucky** and **Princess Milka** for always keeping my back.

I want to thank **Tabea Löblein** and **Benedict Philippi** for proofreading this work.

I want to thank all my friends, family and colleagues, who I did not manage to name explicitly, for believing in me and supporting me.

Contents

List of Figures

List of Tables

List of Algorithms

Part I

Introduction

Chapter 1

Being a researcher in the emerging interdisciplinary field of marine data science

An ocean of data – The slogan of my graduate school refers to the increasing amount of available marine data and the huge opportunity to draw knowledge from these. Currently, the evaluation of marine data relies on the proper selection and application of a data evaluation technique within the specific marine research domain. This can result in keeping outdated measures in the analysis or the application of unsuited tools, since the selection of such tools is generally difficult. The development of data analysis methods is the main task of data science and usually conducted without a specific target application. This makes data science tools well applicable to a broad variety of data without specific restrictions or requirements. But domain data generally comes with specific features and domain research with specific requirements for what to draw from the data. Marine data science is an emerging interdisciplinary research field that approaches the development and application of data science tools directly targeted towards the evaluation of marine data. Marine data scientists require a fundamental education in well established data science fields such as mathematics, computer science or engineering as well as a broad overview of marine science knowledge and typical data characteristics [VTA+21]. By this, marine data scientists can tackle current marine research questions by evaluating marine data with best suited data science methods and furthermore (re)design methods directly targeted to best support marine research.

The aim of my doctoral research in marine data science was to develop

a statistical tool for the application on data with typical features of marine data. My research required in-depth knowledge about non-parametric statistics, analysis, numerics and computer science on one hand and a good understanding of marine biogeochemical modeling on the other. Initial in my work, I had to overcome typical challenges of interdisciplinary research: get to know different domain specific languages, identify different domain specific research steps and goals, and learn about non-transferability of concepts such as optimization procedures for models [OCK+22].

My first important marine data science step was the collection of data. Mathematical methods are developed and designed for general and non-specific data $X \in \mathbb{R}^n$, $n \in \mathbb{N}$ in marine science the application on *real* (measurement or simulation) data is commonly standard. For this, I manually collected, pre-processed and double-checked thousands of data points from ocean measurements. The sources were mainly data sets from other marine researchers [e.g. GF94; TGH+19; LPC+19] and a data platform [Alf]. I published these data on PANGAEA, an Earth science data publishing platform [VST+21; PST+22] and a description of the resprective data in an Earth science data journal [VSS+21].

For the development of my statistical tool, I used my previously collected data (among others) to directly test its performance on real marine data. I used different marine biogeochemical data sets as well as direct communication with marine researchers to identify typical marine data characteristics and requirements for an analytic tool. I published the software at a general-purpose open repository [PS23]. Furthermore, I submitted a paper describing the algorithm to a geoscientific modeling journal (Part II, Chap. 4).

Finally, I decided to additionally support the establishment of this emerging research field of marine data science. For this, I became the student representative of my local graduate students. This included to represent their interests in the steering board. We discussed the general structure of the education of young marine data scientists, decided which research projects should be funded and how to best support the growths of marine data science. Furthermore, I hosted a workshop about researchers' demands and desires towards marine data science and published the outcome in the Frontiers journal together with my fellow marine data scientist Carola Trahms [VTA+21]. Finally, the two of us brought marine

data science to an international young marine researchers conference as an individual session.

Developing own significant research while connecting and communicating with researchers from both – marine and data sciences – is key in successfully becoming a marine data scientist.

Chapter 2

Opportunities and challenges in modelling the carbon cycle

Carbon is a crucial element for life on Earth. It is part of all living organisms, fluxes throughout the entire Earth system, supplies energy and determines climate conditions. Since the beginning of the industrial era the global carbon cycle is strongly perturbed by human activities, such as mining, burning of fossil fuels, deforestation and factory farming. By this, anthropogenic CO_2 emissions form one of the main driving forces of current and future climate change [IPC13].

The ocean acts as an important role in buffering the rise of atmospheric CO_2 due to fossil fuel emissions. Gaseous atmospheric carbon dissolves at the surface of the Earth's oceans, some of which gets incorporated into biomass by photosynthesizing phytoplankton and passes through the marine food web. As parts of feces and dead biomass, carbon sinks as particulate organic matter down towards the ocean sediments. The majority of this carbon gets remineralized. Other parts can be buried and enter an enormous storage pool for up to thousands to million of years. About 60% of the anthropogenic CO_2 emissions have already been compensated by such natural sinks, still leaving the atmosphere anthropogenically enriched by about 880 Gt CO_2 since 1750 [IPC14]. Observing specific carbon isotopes can help to identify sources of carbon [RW86]. The particulate organic carbon-13 isotope is generally denoted as a ratio as

$$\delta^{13}\mathrm{C} = \left(\frac{\frac{^{13}\mathrm{C}}{^{12}\mathrm{C}}}{R_{std}} - 1 \right), \tag{2.0.1}$$

where $R_{std} = 0.0112372$ is a standard ratio and ^{12}C and ^{13}C are the absolute

concentrations of the respectove carbon isotopes [Hay04].

Earth system models can serve to understand and predict global carbon dynamics. They simulate the element cycling [ISS+13; HRA+14] and thus the ocean's carbon uptake capacity [BWR+17; NMK+16]. Also, modeling of specific carbon isotopes has become of particular interest [TB08; SGC+13; SS16]. Earth system models can assist the understanding of past and present and predict future conditions. Furthermore, they serve as a test laboratory for mitigation strategies to tackle the current threats of climate change. Hence, the reliability of these models is crucial, and thus their calibration by comparison of simulation to corresponding observational data is a common strategy.

Calibrating models is important to improve their robustness, since these are simplified representations of the real world. The comparison of simulation to field data can give an impression into how well the model reproduces real world processes and their associated data patterns. Important processes, interactions, and exchanges between the atmosphere, hydrosphere, biosphere, and chemosphere must be appropriately resolved. Consequently, a tuning of model parameters can increase the ability of the model to reproduce data patterns observed in-situ.

Typical calibration methods require equally sized data of observations and corresponding model outputs and reduce the data down to a comparable amount. Field data measurements are unevenly distributed as well in time as in space, and furthermore usually sparse. In comparison to this, model data is available in every single grid cell and time step for each simulated tracer. A typical calibration method is the root mean squared error, which requires field and model data to be equally sized. The reduction of data to comparable subsets is commonly achieved by incorporating only data points from grid cells, where both data types are available. By this, many likely useful information gets lost and spatial biases of the model are highly resolved in the error calculation. A different approach that considers more information is to calculate a mean or median of the data before comparison allowing to use all available data. Nevertheless, for multimodal data this approach is highly unsuited, because it completely disregards many possibly occurring data structures.

The ability to explore data distributions as a continuous functions independent of the amount of available data provides an opportunity to

compare unevenly sized data sets. Such a tool can be used to explore field and simulation data individually and afterwards construct a cost function comparing these two results. Before such exploration, data can be selected from meaningful locations and times, e.g. biomes and seasons, to stay within a comparable data subsample. Still, this approach allows to take all available data into account and disregards spatial or temporal biases within a reasonable range.

Chapter 3

Data exploration by nonparametric density estimation

Probability density functions (PDFs) are simple tools to explore data structures more independent of the amount of available data points than other traditional approaches, while still resolving important data distribution features. PDFs are constructed from data values and neglect meta information as space and time. Mathematically, they are integrable non-negative functions $f : \mathcal{A} \to [0, \infty]$ from a probability space $(\Omega, \mathcal{A}, P)$ into the non-negative real numbers and allow to directly relate the probability of the occurrence of a data value $x \in \mathbb{R}$ within a specific range $[a, b] \subseteq \mathbb{R}$ via the relationship

$$P(a < x < b) = \int_a^b f(x)\, dx \text{ for all } a < b \in \mathbb{R} \tag{3.0.1}$$

[Sil86]. This connection is possible since the PDF is the derivative of the distribution of the data [Irl10].

Statistical approaches towards the estimation of PDFs are differentiated in parametric and non-parametric ones. The traditional parametric approach for estimation of an unknown density assumes that this belongs to a class of known distributions. In this case only the distribution parameters – like mean and variance – are estimated. Many real world data sets do not belong to known distributions and hence such an estimate is designed to fail. Non-parametric statistics target the estimation of the PDF without the introduction of assumptions and construct the estimate directly from the data [Tsy09] and therefore are used in this work.

A kernel density estimator (KDE) is a common non-parametric method for the estimation of PDF. It equips every data point with a so called kernel

(function). The kernels are usually downscaled versions of well-known PDFs, ensuring the final estimate being a PDF itself by inheriting all properties of its kernels. The final KDE is the weighed sum of all kernels and by this takes information of every single data point into account and treats all of them equally. Consequently, every point's information equally contributes to the resulting estimate. A smoothing parameter determines the smoothness of the KDE by defining the width of the individual kernels. Generally, a kernel function $K : \mathbb{R} \to \mathbb{R}_{>0}$ shall satisfy

$$\sup_{y \in \mathbb{R}} |K(y)| < \infty \wedge \int |K(y)|\, dy < \infty \wedge \lim_{y \to \infty} |yK(y)| = 0 \wedge \int K(y)\, dy = 1.$$

The KDE of the density f of a data set $X := (X_i)_{i=1}^N \subseteq \mathbb{R}$ is defined as

$$\hat{f} : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geqslant 0}, (x;t) \mapsto \frac{1}{n\sqrt{t}} \sum_{i=1}^{N} K\left(\frac{x - X_i}{\sqrt{t}}\right). \tag{3.0.2}$$

The parameter $\sqrt{t} \in \mathbb{R}_{>0}$ denotes the smoothing parameter of the KDE [Par62].

KDEs promise to be useful for the evaluation of Earth system data, but available implementations have shortcomings in the application on marine biogeochemical data. The most commonly used is the Gaussian KDE. This is built from Gaussian kernel functions and known to work fast and be insensitive to noise. Unfortunately, it tends to oversmoothing [BMP77]. This makes the Gaussian KDE perform poorly on multimodal data and boundary close values [MR94]. For example, in a model assessment, important data features can be unresolved and create a false impression of the model's performance.

A modern approach to construct a KDE is done by solving the partial differential equation describing the diffusion heat process [CM00]. This approach proofed to work better on multimodal and boundary close data [BGK10], and further studies already showed robust results of this KDE [BGK10; DCR11]. The time parameter of this differential equation equals the squared bandwidth parameter $t \in \mathbb{R}_{>0}$ [CM00]. Together with a parameter function, Neumann boundary conditions and the δ-distribution of the input data as initial value [BGK10], this approach can

be summarized in the initial value problem

$$\frac{\partial}{\partial t}u\left(x;t\right)=\frac{1}{2}\frac{d^2}{dx^2}\left(\frac{u\left(x;t\right)}{p\left(x\right)}\right) \qquad ,x\in\Omega,t\in\mathbb{R}_{>0} \tag{3.0.3}$$

$$\frac{\partial}{\partial x}\left(\frac{u\left(x;t\right)}{p\left(x\right)}\right)=0 \qquad ,x\in\partial\Omega,t\in\mathbb{R}_{>0} \tag{3.0.4}$$

$$u\left(x;0\right)=\frac{1}{N}\sum_{j=1}^{N}\delta\left(x-X_j\right) \qquad ,x\in\Omega \tag{3.0.5}$$

with $\Omega\subseteq\mathbb{R}$ a domain, $X\in\Omega^N$ the input data and $p\in C^2\left(\Omega,\mathbb{R}_{>0}\right)$ such that $\left\|p''\right\|_\infty<\infty$. The solution of this system $u\in C^{2,1}\left(\Omega\times\mathbb{R}_{>0},\mathbb{R}_{\geqslant 0}\right)$ is then called the *diffusion KDE*.

The parameter function p acts inversely to a classical diffusion coefficient. By this, it provides adaptive smoothing to the diffKDE. The smaller the values of p are, the higher is the diffusion impact and vice versa. Selecting p as a simple KDE itself lets the estimation best improve from this property [BGK10].

An optimal choice for the smoothing parameter is generally seen to be a minimizer of the asymptotic mean integrated squared error [Par62]. For the diffusion KDE, this means that the final iteration time in the solution of Eq. 3.0.3 is optimally chosen to be

$$T^*=\left(\frac{\mathbf{E}\left(\sqrt{p\left(X\right)}\right)}{2N\sqrt{\pi}\left\|\left(\frac{f}{p}\right)''\right\|_{L^2}^2}\right)^{\frac{2}{5}} \tag{3.0.6}$$

[BGK10]. The calculation of this is depending on the parameter function p as well as on the true PDF f. I propose here to apply a simple KDE to pre-calculate a first estimate for f to be used in Eq. 3.0.6. Using simple first KDE calculations in support of the calculation of the diffusion KDE is generally called *pilot estimation steps* [Abr82].

Designing a diffusion KDE algorithm and software for the exploration of marine biogeochemical data can fundamentally increase the knowledge that can be extracted from these data. The KDE can resolve the distribution from multimodal and boundary-close data even from noisy data sets.

To be appropriate for such task, the diffusion KDE requires a suitable approximation of the final iteration time T^* and the two pilot estimates for p and f. Choosing a proper discretization and programming language are furthermore important for the applicability in marine research. Finally, testing the diffusion KDEs efficiency and performance on real marine data ensures its usefulness in the calibration of Earth system models.

Chapter 4

Data collection and experimental setup

The aim of my work was to develop a kernel density estimator (KDE) for the exploration of marine data. I collected and compiled a new $\delta^{13}C_{POC}$ data set as an example application for the KDE. The carbon isotope ratio $\delta^{13}C_{POC}$ provides useful information into carbon origins and traces through the Earth system. All of this work was elabrorated in the framework of the emerging field of marine data science. My specific marine data science research is located in between the fields of statistics, probability theory, numerics, optimization, analysis and computer science, marine biogeochemical modeling, marine biology and marine chemistry.

4.1 Drawing the perspective of marine data science

Conducting research in an emerging field provides scientists the opportunity to actively shape the future direction of their field. I aim to strengthen the general definition and visibility of marine data science by my own research as well as by meta research about this field. For this, I teamed up with my fellow marine data scientist Carola Trahms and organized a workshop with marine and data scientists and marine data science graduate students. Our aim was to obtain general information from all of them about

- expected benefits for own research field in educating and employing marine data scientists

▻ the difference between a support scientist and a marine data scientist

▻ skills and knowledge a marine data scientist should possess

▻ career opportunities for marine data scientists

We published the outcome of the workshop [VTA+21] and set up a collaborative team of scientists of all domains and career stages listed in Tab. 4.1. Carola Trahms and I were the co-leaders of the study. Together we set the goals for the content and distributed the tasks among the coauthors. The resulting input contributions are summarized in Tab. 4.2

Table 4.1. Authors team of the marine data science publication

	Marine scientists	Data scientists	Marine data scientists
PhD students			Carola Trahms Maria-Theresia Pelz Martin Prinzler
Coordinators	Avan Antja Enno Prigge		
Researchers	Christopher Somes Markus Schartau		
Professors	Arne Biastoch	Thomas Slawig Thorsten Dickhaus Matthias Renz	

Table 4.2. Authors contributions to the marine data science publication

contribution	researcher
marine data	Schartau, Somes
marine science knowledge	Antja, Biastoch
data science knowledge	Dickhaus, Renz, Slawig, Prinzler
soft & interface skills	Antja, Prigge
training	Prigge
introduction, summary, conclusion	Trahms, Pelz

4.2 Collection of the $\delta^{13}C_{POC}$ data

The first step of this work was to set up a data base of marine field data. I chose $\delta^{13}C_{POC}$ data that shall later be incorporated in model calibration [SS16]. I published the new data set together with one of my supervisors and other marine researchers at the world data center for for Earth and environmental science PANGAEA [Alf]. Currently, there are two versions of the data available: my first originally published data base referred to as first version [VST+21] and a second updated data base extended by additional data referred to as second version [PST+22]. Additionally, I published a data description paper in the journal Earth System Science Data (ESSD) explaining in detail the set up of first data base version and showing its main data characteristics [VSS+21].

I built the first version on an existing data base by [GF94]. I collected data from [Alf] and added data provided by [LPC+19] and [TGH+19]. The data base is set up in a CSV file containing all relevant meta information. Together with Christopher Somes, we additionally provided NetCDF files of the data interpolated onto two different global grids. A coarse one for model calibration on a UViC Earth System Model grid [SS16] and a fine one for comparison with other gridded data [GWP+18].

The second data base version was set up with support of my supervisor Christopher Somes. It contains extensions by data from [CH20; ZZC+14; WCW+99; EPH+19; GEH+21]. This version is also provided as a spreadsheet file and as a NetCDF file interpolated into a global grid [GWP+18]. The second data base version is also published at the data platform PANGAEA [PST+22].

4.3 Development of a diffusion-based kernel density estimator

The idea to calculate the classical Gaussian KDE by solving the diffusion heat partial differential equation was first proposed by Chaudhuri and Marron [CM00] and its benefits were mainly investigated by Botev et al. [BGK10]. The latter stated that this approach is well suited for multimodal and boundary close data. This inspired me to choose the diffusion ap-

proach to build a software package for the approximation of probability density functions designed for typical data features of marine biogeochemical data. These generally include multiple modes from different biogeochemical influences, boundary close data as in size analyses and also noise from measurement errors or numerical simulations.

I decided to design a new algorithm of the diffusion KDE that best estimates densities of marine data. I will refer to this new diffusion KDE by diffKDE from here on. The algorithm is mainly based on a new bandwidth approximation of Eq. 3.0.6, which builds on two pilot estimation steps for the unknown functions f and p in Eq. 3.0.6. Using simple KDEs as pilot estimates was already a common strategy to increase the accuracy of the final estimate [Abr82; She04]. Previously, pilot estimates were often chosen as Gaussian KDEs [BGK10]. Both of my pilot estimates are derived by solving the diffusion equation. Their final iteration times are data based approaches by [Sil86]. The pilot estimates are directly plugged into Eq. 3.0.6 for calculation of the final iteration time of the diffKDE. One of them additionally acts as a parameter function in Eq. 3.0.3 inversely to a diffusion coefficient as proposed by [BGK10].

An additional feature of my implementation is the provision of a family of estimates at different smoothing intensities and letting the user choose their own preference of smoothing intensity. This is an implicit feature of the temporal solution of the diffusion equation. It solves the possibility of no single optimal smoothing parameter being available [Sco12] by following the idea to provide a series of estimates [BMP77; She04].

I chose the software language Python 3 [VD09] using SciPy [GVB+22] and Numpy [HMW+20]. Visualizations are realized with Matplotlib [Hun07] and calculations based on the Python Math package [Van20]. I chose Python, because it is an open source and a popular programming language in scientific research and data exploration.

My first approach was a finite element discretization build on the software framework FEniCS [ABH+15]. This approach delivers fast and reliable results, but is depending on running the software in an own environment [Inc20] on Unix platforms and even more difficult approaches like docker [Mer14] on Windows machines.

To make my software valuable for a broad community, I decided to reimplement the software based on finite elements. This new discretization

follows concepts by [Sla15] and was already used in a similar approach to implement a diffusion based kernel density estimator for linear networks implemented in R by [MBN16].

I published my software package at Zenoodo [PS23] and submitted a description paper of the new algorithm and implementation to Geoscientific Model Development (GMD) [PSS+23].

Chapter 5

List of publications

Please note that my first three publications were published under the name Verwega. I – Maria-Theresia Pelz – am first author of these.

- **Maria-Theresia Verwega**, Christopher J Somes, Markus Schartau, Robyn Elizabeth Tuerena, Anne Lorrain, Andreas Oschlies, and Thomas Slawig. "Description of a global marine particulate organic carbon-13 isotope data set". In: Earth System Science Data 13.10 (Oct. 2021), pp. 4861–4880. https://doi.org/10.5194/essd-13-4861-2021

- **Maria-Theresia Verwega**, Carola Trahms, Avan N Antia, Thorsten Dickhaus, Enno Prigge, Martin H U Prinzler, Matthias Renz, Markus Schartau, Thomas Slawig, Christopher J Somes, and Arne Biastoch: "Perspectives on marine data science as a blueprint for emerging data science disciplines". In: Frontiers in Marine Science 8 (Dec. 2021). https://doi.org/10.3389/fmars.2021.678404

- **Maria-Theresia Verwega**, Christopher J Somes, Robyn E Tuerena, and Anne Lorrain. A global marine particulate organic carbon-13 isotope data product. data set. 2021. https://doi.org/10.1594/PANGAEA.929931

- **Maria-Theresia Pelz**, Christopher J Somes, Robyn E Tuerena, Anne Lorrain, Hilary G Close, Lillian C Henderson, Katie St John Glew, Boris Espinasse, and Clive N Trueman. A global marine particulate organic carbon-13 isotope data product (version2). en. 2022. https://doi.org/10.1594/PANGAEA.946915

- **Maria-Theresia Pelz**, Markus Schartau, Christopher J Somes, Vanessa Lampe, and Thomas Slawig.: A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation

for the analysis of data in geoscience and ecological research, Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2023-17, in review, 2023.

▻ **Maria-Theresia Pelz** and Thomas Slawig: (2023). "Diffusion-based kernel density estimator (diffKDE) (Version 1)". Zenodo. https://doi.org/10.5281/zenodo.7594915

Part II

Main research and publications

Chapter 1

Perspectives on Marine Data Science as a Blueprint for Emerging Data Science Disciplines

First author paper published in Frontiers in Marine Science at the 2nd December 2021.

frontiers in Marine Science

PERSPECTIVE
published: 02 December 2021
doi: 10.3389/fmars.2021.678404

Perspectives on Marine Data Science as a Blueprint for Emerging Data Science Disciplines

Maria-Theresia Verwega[1,2†], Carola Trahms[1,2†], Avan N. Antia[2], Thorsten Dickhaus[3], Enno Prigge[1], Martin H. U. Prinzler[3,4], Matthias Renz[2], Markus Schartau[1], Thomas Slawig[2], Christopher J. Somes[1] and Arne Biastoch[1,2]*

[1] GEOMAR - Helmholtz Centre for Ocean Research, Kiel, Germany, [2] Kiel University, Kiel, Germany, [3] University of Bremen, Bremen, Germany, [4] AWI - Alfred Wegner Institute for Polar and Ocean Research, Bremerhaven, Germany

OPEN ACCESS

Edited by:
Viola Liebich, Bremen Society for Natural Sciences, Germany

Reviewed by:
Robert William Schlegel, Institut de la Mer de Villefranche (IMEV), France
Maike Sonnewald, Princeton University, United States

***Correspondence:**
Carola Trahms
ctrahms@geomar.de

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

Received: *09 March 2021*
Accepted: *11 November 2021*
Published: *02 December 2021*

Citation:
Verwega M-T, Trahms C, Antia AN, Dickhaus T, Prigge E, Prinzler MHU, Renz M, Schartau M, Slawig T, Somes CJ and Biastoch A (2021) Perspectives on Marine Data Science as a Blueprint for Emerging Data Science Disciplines. Front. Mar. Sci. 8:678404. doi: 10.3389/fmars.2021.678404

Earth System Sciences have been generating increasingly larger amounts of heterogeneous data in recent years. We identify the need to combine Earth System Sciences with Data Sciences, and give our perspective on how this could be accomplished within the sub-field of Marine Sciences. Marine data hold abundant information and insights that Data Science techniques can reveal. There is high demand and potential to combine skills and knowledge from Marine and Data Sciences to best take advantage of the vast amount of marine data. This can be accomplished by establishing Marine Data Science as a new research discipline. Marine Data Science is an interface science that applies Data Science tools to extract information, knowledge, and insights from the exponentially increasing body of marine data. Marine Data Scientists need to be trained Data Scientists with a broad basic understanding of Marine Sciences and expertise in knowledge transfer. Marine Data Science doctoral researchers need targeted training for these specific skills, a crucial component of which is co-supervision from both parental sciences. They also might face challenges of scientific recognition and lack of an established academic career path. In this paper, we, Marine and Data Scientists at different stages of their academic career, present perspectives to define Marine Data Science as a distinct discipline. We draw on experiences of a Doctoral Research School, MarDATA, dedicated to training a cohort of early career Marine Data Scientists. We characterize the methods of Marine Data Science as a toolbox including skills from their two parental sciences. All of these aim to analyze and interpret marine data, which build the foundation of Marine Data Science.

Keywords: Marine Data Science, interface science, emerging science, Ph.D training, Data Science, Marine Sciences, Earth System Sciences

MOTIVATION

Earth System Sciences have seen enormous technological progress within the past decades, generating huge data sets from various sources. Increasingly, applications of big data are being used to generate policy advice, monitor regulations, and test potential mitigation measures. Using science to guide decision making requires transparent analyses and impartial communication. Uncertainties and limitations of scientific output must be clear to public, policy makers, and the media.

At the same time, Data Scientists apply, redesign, and develop new methods in statistics, data mining, and machine learning. These methods can be established to tackle specific challenges of research questions in Earth System Sciences. They have to prove their usefulness in applications to data stemming from this area of research, as it is often unknown whether the assumptions that regulate Data Science methods are met by real data from other disciplines. Therefore, combining unique non-conforming data sets not typically used together, with relevant research questions, provides a scientific benefit. Such research also serves to improve the development of data and computer science methods themselves.

We argue that it is time to integrate the fields of Earth System Sciences and Data Science. Data Science should be established as a fourth paradigm (Hey et al., 2009) in Earth System Sciences beyond observations, theory and modeling, requiring its own experts and specialists. We will present Marine Sciences as an example for Earth System Sciences since these are the areas of expertise of our consortium. We call this emerging interface field Marine Data Science (MDS) and the scientists within this field Marine Data Scientists (MDSc). To define this field and identify its needs, we conducted a workshop with eight principal investigators from both Marine and Data Sciences, and 14 doctoral candidates conducting research within MDS projects. This expert team was formed by members of a graduate school (MarDATA), which aims to educate early career MDSc and shall serve as an example of how such a pathway can be implemented.

Marine Data Science as an Emerging Field

Marine Sciences have been generating huge amounts of heterogeneous data stemming from, for example, experiments, observations, and model results including high frequency data streams (Williams et al., 2016; Mayer et al., 2018; Tanhua et al., 2019). Major advances in automated and remote observation capacity and the simultaneous collection of increasingly diverse data challenge conventional data handling methods. As more of the ocean is measured and mapped, and modeling increases in complexity, the need for innovative tools and methods will increase. Marine Scientists must extract scientific information from sparse data through smart analyses. However, Marine Scientists have little or no formal training in Data Science methods such as machine learning approaches. Data Scientists, similarly, lack formal knowledge of Marine Sciences. By merging disciplinary expertise in Marine Data Science, both groups of scientists can harness mutual advantages and provide maximal insights from complex data.

The integration of Data and Marine Sciences already takes place for numerous scientific applications but mostly in an *ad-hoc* manner (with the exception the established research field of bioinformatics). Successful approaches have been implemented originating in Marine Sciences (Malde et al., 2020; Sonnewald et al., 2020), as well as Data Science (Faghmous et al., 2015; Adibi et al., 2020). These approaches highlight the potential in combining the knowledge and power of these two scientific fields. They need to be differentiated from efforts like pangeo.io[1] or Pangaea[2]. Pangeo focuses on making computer science technology (not necessarily Data Science methods) available to natural scientists. Pangaea focuses on offering a platform for publishing research data.

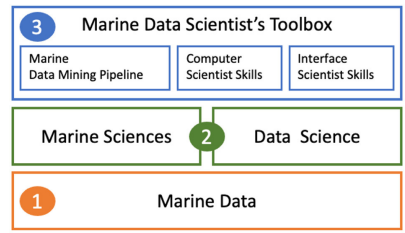

FIGURE 1 | Elements of Marine Data Science. The object of MDS research is marine data (1). Knowledge involved in MDS bases on the parental sciences, Marine and Data Sciences (2). Key methods of MDS are collected in the Marine Data Scientist's Toolbox (3). This includes the Marine Data Mining Pipeline as well as Computer and Interface Scientist Skills.

These examples show that establishing Marine Data Science can be approached from two perspectives: from a specialized field in Marine Sciences with an expansion toward Data Science methodologies, or from Data Science with a specialization toward the Marine Sciences.

In both cases new Data Science methods have the potential to generate added value to marine research, for example in ocean models by aiding the scientific interpretation of 4D model data. They help in improving the workflow and analysis of large volumes of model data. Also, they may help formulate and construct parameterizations of unresolved and underrepresented marine processes.

Even though we find the first way of establishing Marine Data Science important to expand the education of Marine Scientists into this new methodological field, the latter approach was the motivation for the establishment of the Helmholtz School for Marine Data Science (MarDATA)[3]. We view MarDATA as an example for strategic fusion of both methodologies and an application of Data Science methods in Marine Sciences.

In this paper, we present a template to establish a profile for Marine Data Scientists that offers structured training and career perspectives for researchers entering the field. In the following sections, we expand on the MDS components in **Figure 1**, which is followed by a concept on how to train MDSc. We provide an overview and discussion that can help both in designing and conducting MDS research, and guiding the research profile of early career researchers. This shall serve as an example of integrating Earth System Sciences and Data Science while focusing on the specific needs of Marine Sciences.

[1]https://pangeo.io/

[2]https://www.pangaea.de/

[3]MarDATA: https://www.mardata.de/

1. Perspectives on Marine Data Science

1. MARINE DATA

Marine data form the base of MDS research. Since Marine Sciences include the full range of natural sciences, MDSc deal with data that originate from small-scale experiments to globally operating autonomous instruments, satellites, and ocean model outputs. In consequence, their origin, format, scope, and characteristics are diverse. It is crucial for MDSc to understand the background of these data. This includes gathering knowledge about the method of data collection and learning about the suitability of the data for answering specific research questions. Interestingly, Marine and Data Scientists may apply different criteria to evaluate the usefulness of data. Marine Scientists are concerned with the ability of their data to resolve particular processes, while Data Scientists put more emphasis on completeness, consistency, and uncertainties in the data. Both the structural organization and content of data sets are vital for analyses to reach their full potential.

Accessing data from various sources, MDSc face the challenge of combining highly heterogeneous data. This heterogeneity can affect the following areas: *Sources, Data Formats and Data Structures*, *Origin*, *Processing Levels*, *Spatial and Temporal Resolution*.

Heterogeneity of *Sources, Data Formats and Data Structures* arises in the absence of a single, consistent, standardized, global, and generic infrastructure for marine data. Data acquisition, processing, and accessibility depends on national efforts, and repositories are often uncoordinated. Few ongoing efforts exist that coordinate and streamline data repositories, formats, and accessibility (e.g., Ocean Observatories[4], GOOS Moltmann et al., 2019, OOI Schofield et al., 2010, 2013 based on cyberstructure from Farcas et al. (2011)).

Heterogeneity of *Origin* distinguishes marine data by the disciplinary expertise who generated them. We identified three categories, that are not mutually exclusive:

1. *Observational Data* are collected and preprocessed by researchers. The researchers hold expertise in measurement devices and protocols, instrument calibration, data cleaning, and quality control. The data sources might be ship based measurements, moorings, gliders, autonomous underwater vehicles, drifters and floats, sea-floor optic cables, or laboratory measurements.
2. *Highly Processed Data Products* are extrapolated and interpolated in space and time, such as the objectively analyzed data of the World Ocean Atlas [WOA, (e.g,. Garcia et al., 2013; Locarnini et al., 2019)]. Remote sensing measurements can be combined with field observations. Algorithms, models and neural networks derive estimates of ocean properties, which can utilize and expand field observations.
3. *Synthetic data from Simulations and Models* are generated from models that are imperfect representations of the real world. They cover temporal and spatial scales beyond the observational data (e.g., Matthes et al., 2020) and include, for example, future climate projections (e.g., Eyring et al., 2016). Unlike data (1) and (2), simulation output data are usually available on a unique grid depending on the specific model simulation. Climate models, for example, typically provide a four-dimensional space-time grid. Thus, a comparison of model output and measurements always involves interpolation or data aggregation.

Heterogeneity of *processing levels* is concerned with the implicit uncertainty of the data in the specific level, depending on individual processing steps, as well as their underlying assumptions. The levels span raw measurements (level 0), quality-controlled data sets (level 1), derived data and data-model synthesis products (level 2 and higher) to synthetic data from simulations. Although the explicit assignment of processing levels has become common practice in Marine Science, the levels may be defined differently by scientists from different research perspectives.

Heterogeneity of *spatial and temporal resolution* is a common feature in ocean observations. Most field data describe properties of the upper ocean's pelagic layers (upper 500 m), where substantial variability can occur on much shorter time scales than changes in the deep ocean. Global oceanographic data from greater depth remain more scarce. Some ocean regions are still hardly covered at all, such as the southeastern Indian Ocean or the ice-covered polar oceans. Thus, observational data from great depths and from remote ocean regions are highly valuable and these data ought to be well-prepared and made accessible.

Marine data typically reflect multiple dynamical processes that are interconnected or simply overlap, while spanning a wide range of scales (Dickey, 2001). These scales can often not be regarded in isolation. **Figure 2** shows the continuum of features and processes changing in time and space in the marine environment. Failing to account for heterogeneity in the spatio-temporal resolution of data may lead to misinterpretation of results. It might even mask processes of interest, potentially mistaking a relevant signal for noise.

2. KNOWLEDGE FROM MARINE AND DATA SCIENCES

Having noted the challenges that arise from the heterogeneity of marine data, we now describe the core aspects of Marine and Data Sciences that join to form MDS.

2.1. Marine Sciences

An understanding of the ocean requires comprehension of its physical, chemical, biological, and geological processes as well as their interconnection with society. The methods and conceptual approaches of these disciplines differ substantially. They are based on mathematical, physical, biological, geological, or societal system understanding. They range from theoretical approaches, lab-based experiments and *in situ* measurements to global models. Consequently, specific toolboxes (methods and software) and the conceptual framework used depend strongly on the research question and the data type.

It is neither possible nor necessary for any individual MDSc to have a deep understanding of all Marine Sciences. As in

[4]https://oceanobservatories.org/

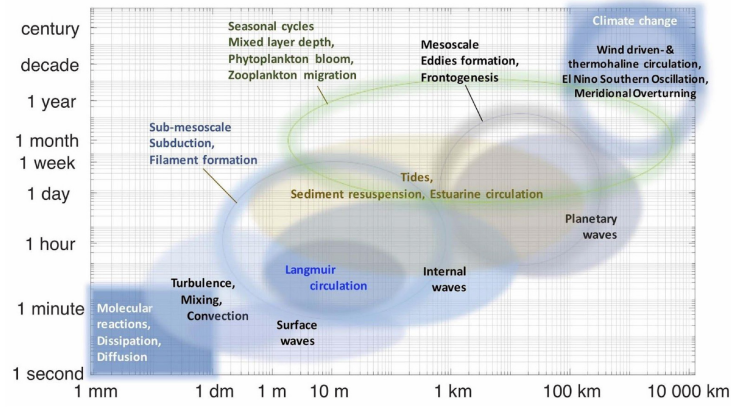

FIGURE 2 | Processes and features that generate variability in the ocean on various temporal and spatial scales. The sketch has been refined and redrawn, inspired by an illustration by Dickey (2001).

any field, it is possible to understand the big questions and the areas of possible breakthrough that big data can mediate. MDSc strive to attain a meta-level understanding of marine processes, and a focused deeper knowledge of the specific research theme they work on. They know the state-of-the-art analytical tools. The tool's strengths, weaknesses, and limitations influence the framing of MDS research questions.

At all scales, ocean models help to test hypotheses and simulate projections. These simulations solve systems of differential equations, using techniques from advanced numerics, parallelization, and high performance computing. Model calibration is usually a high-dimensional nonlinear optimization problem, which requires observational data as constraints. Model parameterizations, for example of ocean mixing or biogeochemical processes, are imperfect and data assimilation methods are one option to provide improved model solutions. Typically, model optimizations require a high number of simulations, increasing the computational power requirement. Model calibration and validation is difficult because of the sparseness and uncertainty of observational data.

Increasingly, Marine Sciences address global challenges such as climate change, biodiversity loss, and the sustainable use of natural resources. Scientific advances, often enabled by data-driven insights, provide the knowledge base for policy and societal action. Examples are the predictions of climate change given by earth system models (Masson-Delmotte et al., 2018), and the development of a digital twin for the ocean that can assess solution options to marine problems (Voosen, 2020). MDSc find themselves at the forefront of this research. They are challenged not just by the research complexity, but by the need to communicate its socioeconomic and ethical implications.

2.2. Data Science

As is the case for Marine Sciences, Data Science is also not a research discipline on its own. It encompasses fields such as statistics, probability theory, or machine learning from traditional disciplines such as mathematics and computer science.

Handling and processing data involves established computer science methods. These include standard algorithms (searching and sorting, usage and manipulation of graphs, etc.) and data structures. MDSc have to understand their functionalities in order to use these tools effectively and efficiently. This also applies to data management concepts. While every domain-specific database system has its own characteristics, basic concepts such as primary and secondary keys, queries, etc. must be known and understood.

Core definitions and theorems of pure mathematics are required in essentially all fields of Data Science. A strong background in these topics forms the foundation necessary to utilize and further develop Data Science tools. Calculations on data points are basically fundamental operations on elements of fields or vector spaces. Their foundations lie in pure mathematics, algebra, and analysis. Utilizing data for finding, for example, a best procedure, requires methods from optimization or optimal control theory. Implementing these concepts into computer programs is part of numerics. Knowledge about convergence, consistency and stability of such algorithms is important to judge their suitability to address a problem as well as to judge the reliability of the result. Application of statistical methods incorporates results from probability theory, hence understanding of basic stochastic calculus is essential to choose the correct statistical method and understand its outcome.

Ordinary and partial differential equations are essential to assess the existence and uniqueness of solutions of numerical models.

Finally, it is crucial to transform data into information, and ultimately into knowledge. Specifically in MDS, advanced machine learning and data mining methods that are designed for complex-structured data are required. Such data include high-dimensional data, sequence data, time series data, data streams, graph data as well as spatial and spatio-temporal data and unstructured data such as text. Awareness of the capabilities and limitations of these techniques is crucial.

3. THE MARINE DATA SCIENTIST'S TOOLBOX

We next discuss the skills that guide MDSc to successfully take on their research challenges (see **Figure 1**). We emphasize that substantial discussion with Marine Scientists about the expected scientific achievements is essential at the start of any MDS project. Only after this exchange can the choice of the specific tool be made.

3.1. Marine Data Mining Pipeline

To facilitate knowledge discovery, MDSc work along a data mining pipeline. This includes *selection*, *preprocessing*, and *transformation* of the data for feature selection for the *machine learning* and *pattern mining* algorithms. It is important to emphasize that data put into this pipeline's preprocessing step have already been processed, cleaned and maybe even imputed by Marine Scientists in their own data preprocessing routines. At the end of the marine data mining pipeline stands a meaningful *evaluation* of model performance utilizing expressive *visualization*. Along this pipeline, MDSc have to cope with data sparsity, problems of overfitting, treatment of outliers and noise, and sorting and weighting data according to quality and uncertainty. Due to this complexity, knowledge discovery is tackled by an iterative process with multiple loops over the pipeline steps. In the following we will focus on aspects of this data mining pipeline applicable to MDS.

During data *selection*, MDSc must keep in mind the scientific question, differences in processing levels and uncertainties between data types. Close collaboration with Marine Scientists is crucial in this step. This includes consideration of boundary conditions and an assessment of the plausibility of a solution, which distinguishes this approach from blind data mining.

In the *preprocessing* step the data is integrated, completed, and made consistent. MDSc consider the origin, temporal and spatial coverage, available metadata, and preprocessing performed on the data. When handed to MDSc, marine data is usually already preprocessed to a higher data level (see section 1).

The next step is *transformation* of the data into a format for machine learning and data mining. This includes feature selection, feature transformation, and dimensionality reduction. Gaussianity can facilitate some analyses and may even be a prerequisite. It can be met for e.g., by applying Gaussian anamorphosis for improved state estimations (Amezcua and Leeuwen, 2014) or logarithmic transformations. Data that exhibit non-Gaussian characteristics might be transformed by other parametric or non-parametric statistical measures (e.g., Tsybakov, 2009).

At the heart of Data Science are knowledge discovery methods such as *machine learning* and *pattern mining*. Their application presumes familiarity with the range of the spatio-temporal scales of the data and the processes involved (see sections 1 and 2). When working with complex, multi-source data, MDSc adapt methods of data mining to account for uncertainties in the data (Liu et al., 2016).

The *evaluation* of the results involves experts from Marine and Data Sciences. Techniques such as explainable artificial intelligence might be more useful than black-box solutions. They facilitate communication of the results and their origins to non-Data Scientists. Although blind data mining might expose unknown and unexpected interdependencies, it requires close collaboration (see section 3.3) to assess whether identified patterns are useful. Once the quality of the results is assessed, the pipeline can be backtracked to repeat steps, applying alternative approaches.

Visualization communicates the message extracted from the data. Descriptive statistics support visualization by removing noise, and summarizing core features. Graphic and dynamic visualization are utilized to communicate results to (Marine Science) colleagues. Innovative ways of presenting or animating data contribute significantly to dissemination of results.

3.2. Computer Science and Programming Skills

To facilitate the steps of the marine data mining pipeline, MDSc apply classical programming skills such as handling databases and UNIX platforms as well as different programming languages.

Database systems build the foundation to access and store marine data in a standardized and well-defined format. MDSc know how to work with relational as well as other database designs, such as NoSQL solutions. MDSc run and parallelize analyses on diverse systems, such as High-performance architectures with numerous CPUs or GPUs and associated storage systems.

Programming languages are essential for performing computer-supported calculations and analyses. Currently, huge marine models are often written in classical programming languages like C, C++, and Fortran[5]. Statistical analyses and data visualization in Marine Sciences are often performed in R[6], MATLAB[7], and programs such as Excel[8] out of convenience. For the data mining pipeline e.g., Python[9] is a helpful choice.

To ensure transferability, reproducibility and sustainability of the developed scripts and software they need to be created

[5]Fortran https://fortran-lang.org/ (accessed January 10, 2021).
[6]R Core Team (2020). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* Vienna. Available online at: https://www.R-project.org/ (accessed January 10, 2021).
[7]MATLAB (2010). *version 7.10.0 (R2010a)*. Natick, MA: The MathWorks Inc. Available online at: https://de.mathworks.com/ (accessed April 25, 2021).
[8]Microsoft Excel https://www.microsoft.com/de-de/microsoft-365/excel (accessed January 10, 2021).
[9]Python Software Foundation. Python Language Reference, version 3.9. https://www.python.org/ (accessed January 10, 2021).

using standard routines such as version control (e.g., git[10]) and modularity as well as good documentation. Hence MDSc need to apply best practices of software engineering.

3.3. Interface Scientists Skills

MDSc are scientists working across disciplines with different conceptual approaches, academic cultures, and disciplinary languages. Hence, they must develop personal and communication skills to allow them to contribute to a joint understanding of scientific questions and research design. In-depth bilateral scientific exchange between Marine and Data Scientists for co-definition of research questions and expected outcomes is the first step of any MDS project. A succinct guide to how such a collaboration could be established and nurtured is given by Ebert-Uphoff and Deng (2017). They suggest that rather than Data Scientists and Marine Scientists trying to gain proficiency in the field of the other, active and persistent collaboration is the way to join expertise. Defining the problem, approach and expectation of the scientific outcome will lead to the selection and application of appropriate data methods. Together with the interpretation of results these are a joint responsibility and must be iterated throughout the entire research process.

Interface skills for MDSc include grasping the conceptual approach and specific terminology of the marine problem while avoiding excessive detail. An innate, curiosity driven motivation, that enables research to be fun, stimulates lateral and innovative thinking and an openness for serendipity help greatly in this process. At a personal level, MDSc are continually operating beyond their comfort zone—they interact as non-experts in new fields. They must navigate among their collaborators and communicate their expertise at conferences and meetings with domain scientists. This requires confidence, questioning the input they receive and having an entrepreneurial mindset that shows resilience, determination, and an enthusiasm for multitasking.

In communication of results to stakeholders and decision makers, MDSc need to address and clarify uncertainties and limitations of their scientific output. This is fundamental to contributing MDS input to policy making and public understanding.

4. HOW TO TRAIN A MARINE DATA SCIENTIST

By defining MDS as a new research field with characteristic methods, workflows and required skills, we see the need for targeted education of early career scientists. This is motivated by Marine Data Science as an example of how Data Science could be fused with all fields of Earth System Sciences into a new interface discipline. We present here our experiences in developing and implementing targeted training for doctoral researchers in MDS within a dedicated graduate school of the Helmholtz Association. This can serve as an example of how data and earth sciences can be bridged to form a new interface discipline.

The Helmholtz School for Marine Data Science (MarDATA)[11], established in 2019, aims at training doctoral candidates (the German equivalent of a Ph.D) in MDS. It is a cooperation of GEOMAR - Helmholtz Centre for Ocean Research Kiel and the Alfred Wegener Institute (AWI) Helmholtz Centre for Polar and Marine Research with partner Universities in Kiel and Bremen, and was initiated by the Helmholtz Association to prepare the next generation of scientists for a data-heavy future. Conducting MDS requires highly specialized Data Science methods, thus we chose doctoral candidates with a Masters degree in Data Sciences. Their doctoral training is conceived to provide a Marine Sciences background as well as targeted in-depth training in information and Data Sciences. They also receive training to sharpen transferable skills that enhance their research output. Their research projects range from the improvement of autonomous underwater navigation over pattern recognition in large data sets to the development of new tools for data analysis. Most of the doctoral researchers aim at a degree in Data Science that could lead to a post-doctoral career inside Earth System Sciences (not necessarily restricted to Marine Sciences), and also prepares them for a career outside academia.

The core of MarDATA is the joint definition of research questions by professors and senior scientists from both Marine and Data Sciences. Regular meetings between the doctoral candidates and both their supervisors (one each from the Marine and Data Science disciplines) have proven to be the most effective. They are essential for a joint understanding of the research question and monitoring research progress. All participants share responsibility for exchange of disciplinary understanding and maintaining useful dialogue. It quickly became apparent, however, that doctoral candidates cannot be the only "glue" between their supervisors.

MarDATA supports scientific exchange by offering joint events, such as datathons[12], hacky hours[13], and other networking opportunities. Doctoral researchers in the MarDATA school gain lateral, interface skills by contributing lectures or workshops to early career Marine Scientists. This contributes visibility to the profile of MDSc in the marine field.

Training measures draw on project-specific expertise from the supervisors, as well as block courses and summer schools. The training is always open for a number of external participants providing an opportunity for knowledge transfer and exchange. Recurring lecture formats allow training in particular methodologies and provide an overview of state-of-the-art research in both domains. Workshop formats allow all involved researchers to strengthen their interface skills, for example in lateral thinking and design thinking.

[10]Software Freedom Conservancy. Git. https://git-scm.com/ (accessed January 10, 2021).

[11]MarDATA: https://www.mardata.de/.

[12]A *datathon* is an event where Data Scientists meet to solve Data Science challenges. These challenges can originate from applied fields or other data-heavy research disciplines.

[13]A *hacky hour* is a fixed informal weekly meeting, where researchers and programmers come together to discuss and solve code and programming related problems.

5. SUMMARY AND CHALLENGES

In this paper we provide our perspective of the current state and future of Marine Data Science (MDS) as a marine example for the fusion of Earth System Sciences with Data Science. To suggest a pathway for its development, we propose a model for the training for early career Marine Data Scientists (MDSc). The discussed ideas are inspired by the Helmholtz Graduate School for Marine Data Science (MarDATA), which is an example of training for a future generation of MDSc. MDSc should be trained both in classical Data Science skills as well as developing strong communication skills across disciplines. The Data Science skills include handling databases and programming languages, while maintaining software development standards, as well as dealing with diverse marine data types. The Marine Science skills include an overview of the marine environment and the characteristics of its data. MDS potentially extricates information from marine data, leading to new knowledge, and can identify new research questions.

Despite the obvious benefits of joining forces of Marine and Data Sciences, MDS comes with its own challenges as regards perspectives for a career after the doctorate. MDSc might struggle with appropriate scientific recognition, since publication strategies in Marine and Data Sciences differ greatly. Marine Scientists usually publish in journals whereas Data Scientists mostly publish in conference proceedings. The latter do not have the same assessment-metrics as journal publications, such as the impact factor. MDSc need to position themselves between these cultures. To attain publication recognition in Marine and Data Science, there is a risk that they will need to publish double the amount expected of pure Marine or Data Scientists.

The definition and education of MDSc is only the first step. Structural change is the necessary second step. Structural support could come through involving MDSc in new projects, assigning permanent positions to MDSs, and offering a pathway to an academic career including professorships in MDS. Besides their scientific expertise, MDSc draw from a range of interface and transferable skills. These qualify MDSc also for a career path outside of academia, in innovation sectors such as business, product development, public and private research and entrepreneurship. MDSc can thus easily transition between or merge the academic with the private and public sectors.

Although this is a Marine Sciences example, we believe that similar potentials and challenges exist for any variant of Earth System Data Science.

6. CONCLUSION

MDS is an example of a novel and highly demanded interdisciplinary research field between the Earth System Sciences and Data Science that needs to be properly defined and established. MDS comes with a high demand on knowledge and skills from both Marine and Data Sciences to be able to effectively work with marine data. It still has to develop a strategy for publishing with best impact, and recognition to facilitate entering a academic career. Also, it has to find a balance between incorporating experiences from the parental sciences while exploring new ways of combining and advancing Marine and Data Sciences simultaneously. We envision MDSc will be able to provide major benefits in advancing Marine as well as Data Sciences, understanding marine data and bridging two distinct scientific fields.

The approach and pitfalls of establishing Marine Data Science mapped out in this paper could be used as a blueprint for establishing other fields of research that fuse Earth System Sciences and Data Science.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CT and M-TV initiated and guided the work, hosted the workshop, developed the structure and framework of the manuscript, its introduction and discussion, the Data Science toolbox, the Marine Science toolbox, and edited the entire manuscript. AA developed the parts about the necessary depths of Marine Science knowledge, interface skills, soft skills, and edited the entire manuscript. AB is the head of the MarDATA research school; he developed the parts about the methods usually applied in Marine Sciences for data analyses. TD developed the parts about traditional education in mathematics, software engineering, and programming. MP developed the Data Science toolbox and the Marine Science toolbox. EP developed the parts about example Ph.D projects, interface skills, and soft skills. MR developed the parts about the general knowledge on machine learning and Data Science methods and solution patterns for common data problems. MS developed the parts about marine data. CS developed the parts about marine data repositories. TS developed the parts about simulation and simulation data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005).

ACKNOWLEDGMENTS

We want to thank all participants of the of the MarDATA workshop *Marine Data Science: Opportunities and Challenges* in November 2020 for their valuable input and enrichment of our discussion. We thank the referees and editors for their constructive feedback regarding the initial version of the manuscript.

REFERENCES

Adibi, P., Pranovi, F., Raffaetà, A., Russo, E., Silvestri, C., Simeoni, M., et al. (2020). "Predicting fishing effort and catch using semantic trajectories and machine learning," in *Lecture Notes in Computer Science* (Würzburg: Springer International Publishing), 83–99.

Amezcua, J., and Leeuwen, P. J. V. (2014). Gaussian anamorphosis in the analysis step of the enkf: a joint state-variable/observation approach. *Tellus A* 66:23493. doi: 10.3402/tellusa.v66.23493

Dickey, T. D. (2001). The role of new technology in advancing ocean biogeochemical research. *Oceanography* 14, 1078–120. doi: 10.5670/oceanog.2001.11

Ebert-Uphoff, I., and Deng, Y. (2017). Causal discovery in the geosciences using synthetic data to learn how to interpret results. *Comput. Geosci.* 99, 50–60. doi: 10.1016/j.cageo.2016.10.008

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., et al. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* 9, 1937–1958. doi: 10.5194/gmd-9-1937-2016

Faghmous, J. H., Frenger, I., Yao, Y., Warmka, R., Lindell, A., and Kumar, V. (2015). A daily global mesoscale ocean eddy dataset from satellite altimetry. *Sci. Data* 2:150028. doi: 10.1038/sdata.2015.28

Farcas, C., Meisinger, M., Stuebe, D., Mueller, C., Ampe, T., Arrott, M., et al. (2011). "Ocean observatories initiative scientific data model," in *OCEANS'11 MTS/IEEE KONA*, 1–10.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., et al. (2013). "World ocean atlas 2013," in *Dissolved Inorganic Nutrients* (Phosphate, Nitrate, Silicate). Vol. 4, eds S. Levitus and Mishonov (NOAA Atlas NESDIS 76), 25. doi: 10.7289/V5J67DWD

Hey, A. J., Tansley, S., Tolle, K. M. (eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. (Redmond, WA; Microsoft Research). Available online at: http://fourthparadigm.org

Liu, Y., Qiu, M., Liu, C., and Guo, Z. (2016). Big data challenges in ocean observation: a survey. *Pers. Ubiquitous Comput.* 21, 55–65. doi: 10.1007/s00779-016-0980-2

Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Boyer, T. P., Zweng, M. M., Garcia, H. E., et al. (2019). *World Ocean Atlas 2018, Volume 1: Temperature*. A. Mishonov (NOAA Atlas NESDIS 81), 52.

Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A.-B. (2020). Machine intelligence and the data-driven future of marine science. *ICES J. Mar. Sci.* 77, 1274–1285. doi: 10.1093/icesjms/fsz057

Masson-Delmotte, V., Zhai, P., Prtner, H.-O., Roberts, D., Skea, J., Shukla, P., et al. (2018). *Ipcc, 2018: Global Warming of 1.5c. an Ipcc Special Report on the Impacts of Global Warming of 1.5c Above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*.

Matthes, K., Biastoch, A., Wahl, S., Harlaß, J., Martin, T., Brücher, T., et al. (2020). The flexible ocean and climate infrastructure version 1 (FOCI1): mean state and variability. *Geosci. Model Dev.* 13, 2533–2568. doi: 10.5194/gmd-13-2533-2020

Mayer, L., Jakobsson, M., Allen, G., Dorschel, B., Falconer, R., Ferrini, V., et al. (2018). The nippon foundation—GEBCO seabed 2030 project: the quest to see the world's oceans completely mapped by 2030. *Geosciences* 8:63. doi: 10.3390/geosciences8020063

Moltmann, T., Turton, J., Zhang, H.-M., Nolan, G., Gouldman, C., Griesbauer, L., et al. (2019). A global ocean observing system (GOOS), delivered through enhanced collaboration across regions, communities, and new technologies. *Front. Mar. Sci.* 6:291. doi: 10.3389/fmars.2019.00291

Schofield, O., Glenn, S., Orcutt, J., Arrott, M., Meisinger, M., Gangopadhyay, A., et al. (2010). Automated sensor network to advance ocean science. *Eos Trans. Am. Geophys.l Union* 91, 345–346. doi: 10.1029/2010EO390001

Schofield, O., Glenn, S. M., Moline, M. A., Oliver, M., Irwin, A., Chao, Y., et al. (2013). *Ocean Observatories and Information: Building a Global Ocean Observing Network*. New York, NY: Springer, 319–336.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Sci. Adv.* 6:eaay4740. doi: 10.1126/sciadv.aay4740

Tanhua, T., Pouliquen, S., Hausman, J., O'Brien, K., Bricher, P., de Bruin, T., et al. (2019). Ocean FAIR data services. *Front. Mar. Sci.* 6:440. doi: 10.3389/fmars.2019.00440

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York, NY: Springer Science & Business Media.

Voosen, P. (2020). Europe builds 'digital twin' of earth to hone climate forecasts. *Science* 370, 16–17. doi: 10.1126/science.370.6512.16

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., et al. (2016). A global repository for planet-sized experiments and observations. *Bull. Am. Meteorol. Soc.* 97, 803–816. doi: 10.1175/BAMS-D-15-00132.1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RS declared a past collaboration with one of the authors AB to the handling editor.

Chapter 2

Description of a global marine particulate organic carbon-13 isotope data set

First author paper published in Earth System Science Data on the 26th October 2021.

Earth Syst. Sci. Data, 13, 4861–4880, 2021
https://doi.org/10.5194/essd-13-4861-2021

Open Access Earth System Science Data

Description of a global marine particulate organic carbon-13 isotope data set

Maria-Theresia Verwega[1,2], Christopher J. Somes[1], Markus Schartau[1], Robyn Elizabeth Tuerena[3], Anne Lorrain[4], Andreas Oschlies[1], and Thomas Slawig[2]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]Department of Computer Science, Kiel University, Kiel, Germany
[3]Scottish Association for Marine Science, Dunstaffnage, Oban, PA37 1QA, UK
[4]LEMAR, Univ Brest, CNRS, IRD, Ifremer, 29280 Plouzané, France

Correspondence: Christopher J. Somes (csomes@geomar.de)

Received: 14 May 2021 – Discussion started: 26 May 2021
Revised: 28 September 2021 – Accepted: 29 September 2021 – Published: 26 October 2021

Abstract. Marine particulate organic carbon stable isotope ratios ($\delta^{13}C_{POC}$) provide insights into understanding carbon cycling through the atmosphere, ocean and biosphere. They have for example been used to trace the input of anthropogenic carbon in the marine ecosystem due to the distinct isotopically light signature of anthropogenic emissions. However, $\delta^{13}C_{POC}$ is also significantly altered during photosynthesis by phytoplankton, which complicates its interpretation. For such purposes, robust spatio-temporal coverage of $\delta^{13}C_{POC}$ observations is essential. We collected all such available data sets and merged and homogenized them to provide the largest available marine $\delta^{13}C_{POC}$ data set (https://doi.org/10.1594/PANGAEA.929931; Verwega et al., 2021). The data set consists of 4732 data points covering all major ocean basins beginning in the 1960s. We describe the compiled raw data, compare different observational methods, and provide key insights in the temporal and spatial distribution that is consistent with previously observed large-scale patterns. The main different sample collection methods (bottle, intake, net, trap) are generally consistent with each other when comparing within regions. An analysis of 1990s median $\delta^{13}C_{POC}$ values in a meridional section across the best-covered Atlantic Ocean shows relatively high values (≥ -22 ‰) in the low latitudes ($< 30°$) trending towards lower values in the Arctic Ocean (~ -24 ‰) and Southern Ocean (≤ -28 ‰). The temporal trend since the 1960s shows a decrease in the median $\delta^{13}C_{POC}$ by more than 3 ‰ in all basins except for the Southern Ocean, which shows a weaker trend but contains relatively poor multi-decadal coverage.

1 Introduction

Carbon is an essential element for life, and it regulates climate via its atmospheric form CO_2, a long-living greenhouse gas. Understanding carbon cycling is fundamental to reliably projecting changes in the Earth's future climate. Carbon is subject to transformation and cycling throughout the ocean, land and atmosphere. It is a major part of organic matter of all living organisms which can both consume (e.g., photosynthesis) and produce (e.g., respiration) inorganic carbon. Besides the natural cycling processes, the total amount and distribution of carbon is strongly perturbed by human activity caused by industrialization, most notably due to fossil fuel emissions, deforestation, farming, cement production and other industrial processes. Anthropogenic CO_2 emissions are one of the main driving forces of modern climate change which is likely to continue in the future (IPCC, 2013). Only about 60 % of anthropogenic CO_2 emissions have been compensated for by natural sinks, including the dissolution of inorganic carbon in the ocean. This means the atmosphere has already been enriched with anthropogenic carbon by about 880 Gt CO_2 since 1750 (IPCC, 2014), which is driving the increase in global temperature levels. The ocean serves as an important buffer as it absorbs a significant amount of an-

thropogenic carbon, with the ocean interior being the largest readily exchangeable reservoir of carbon in the Earth system.

Marine phytoplankton convert dissolved inorganic carbon (e.g., aqueous CO_2) into organic carbon via photosynthesis in the euphotic surface layer. This organic carbon forms the base of the food web for higher tropic levels in marine ecosystems. Some particulate organic carbon (POC) sinks down to ocean depths, where it either is respired back to dissolved inorganic carbon by heterotrophic organisms or becomes buried in ocean sediments (Suess, 1980). This process is known as the soft-tissue biological carbon pump, an important mechanism for sequestering carbon to the deep ocean from the atmosphere (Volk and Hoffert, 1985; Banse, 1990; McConnaughey and McRoy, 1979). Since the deep ocean has a residence time of about a millennium, it is a key carbon reservoir influencing long-term climate change.

Carbon isotopes provide additional insights into the cycling of carbon in the Earth system (Zeebe and Wolf-Gladrow, 2001). The element carbon exists in two naturally occurring stable isotopes, ^{12}C and ^{13}C, with abundances of around 98.9 % and 1.1 %, respectively. Knowledge of their pathways through carbon reservoirs can support deeper understanding of carbon transfer and can help identify carbon sources with different isotopic ratios (Rounick and Winterbourn, 1986). Relative abundances of carbon isotopes are usually given in δ notation, which is based on the carbon isotope ratio $\frac{^{13}C}{^{12}C}$, standardized and given in parts per thousands as

$$\delta^{13}C = \left(\frac{\frac{^{13}C}{^{12}C}}{R_{std}} - 1 \right). \quad (1)$$

The constant $R_{std} = 0.0112372$ is a standard ratio, originally referring to the calcareous fossil Pee Dee Belemnite. The values ^{12}C and ^{13}C are the absolute concentrations of the individual isotopes (Hayes, 2004).

Distributed within the carbon cycle, the fractionation of $\delta^{13}C$ is influenced by biological and thermodynamic processes (Gruber et al., 1999). Air–sea gas exchange plays a dominant role at the ocean surface. Phytoplankton photosynthesis and POC remineralization increase their influence in the ocean interior (Gruber et al., 1999; Morée et al., 2018). The processes are dependent on circulation and temperature, and thus their individual influence varies with geographic location (Gruber et al., 1999; Schmittner et al., 2013).

Phytoplankton preferentially incorporate (i.e., fractionate) the lighter ^{12}C carbon isotope into its organic matter. This fractionation causes phytoplankton organic $\delta^{13}C$ to be 10 ‰ to 25 ‰ lower than that of inorganic $\delta^{13}C$, which depends on a variety of environmental, ecological and physiological conditions (e.g., Popp et al., 1989, 1998; Rau et al., 1989, 1996). The main factors that control phytoplankton fractionation are concentrations of CO_2(aq), species-specific effects enforced by the phytoplankton composition and the cellular growth rate, although uncertainties remain regarding the quantification of the specific processes and mechanisms that cause variations in phytoplankton fractionation (e.g., Fry, 1996; Laws et al., 1995; Popp et al., 1998; Bidigare et al., 1997; Cassar et al., 2006).

$\delta^{13}C_{POC}$ provides insights into physical and biological carbon cycle processes in the ocean (e.g., Fry and Sherr, 1989). It helps to diagnose carbon pathways from the atmosphere to the deep ocean including the biological carbon pump (e.g., Jasper and Hayes, 1990; Popp et al., 1989; Freeman and Hayes, 1992) and assists reconstruction of oceanic carbon cycling and even plankton cell size and community structure (e.g., Tuerena et al., 2019; Lorrain et al., 2020). For example, anthropogenic carbon emissions have a distinctly low $\delta^{13}C$ content, making $\delta^{13}C$ a useful property for tracing anthropogenic carbon throughout the Earth system (Eide et al., 2017; Levin et al., 1989; Ndeye et al., 2017). Atmospheric $\delta^{13}C_{CO_2}$ has decreased from -6.5 ‰ in preindustrial times to -8.4 ‰ presently (Rubino et al., 2013). The measurable decrease due to anthropogenic fossil carbon emissions is known as the Suess effect (Keeling, 1979), which enters the ocean via air–sea gas exchange. However, changes in marine $\delta^{13}C_{POC}$ are also significantly influenced by changes in phytoplankton fractionation due to other anthropogenic controls. For example increasing CO_2(aq) concentrations increase surface $\delta^{13}C$ fractionation (Young et al., 2013) and changes in phytoplankton composition and temperature influence phytoplankton growth rates and $\delta^{13}C$ fractionation over the air–sea interface (Zhang et al., 1995). But determination of the driving processes(es) of $\delta^{13}C_{POC}$ spatial and temporal trends remains a challenge. We also stress that all of these processes are sensitive to temperature changes which adds additional complexity to understanding how fractionation may change in space and time. A better understanding of the contributions from all of these effects requires a robust global data set of $\delta^{13}C_{POC}$.

Theoretical projection and understanding of changes associated with $\delta^{13}C_{POC}$ can be executed by models of different scales, which include $\delta^{13}C_{POC}$ circulation. Earth system models serve to simulate and test hypotheses in different scenarios as unbiased assessments (e.g., IPCC, 2014) and may support future decision-making. Besides resolving the mass flux of carbon, many models also simulate stable carbon isotopes (e.g., Schmittner and Somes, 2016; Buchanan et al., 2019; Hofmann et al., 2000; Jahn et al., 2015; Tagliabue and Bopp, 2008; Morée et al., 2018; Magozzi et al., 2017). For reliable calibrations and validations of such processed-based mechanistic models, a spatially and temporally comprehensive data set is essential. This additional constraint provided by marine $\delta^{13}C_{POC}$ assists the reconstruction of oceanic carbon cycling including how much anthropogenic carbon is entering marine ecosystems and being exported to the deep ocean. But until today, there has been a lack of suitable data sets as constraints. This results in large and mostly unknown uncertainties in model results.

2. A global marine particulate organic carbon-13 isotope data set

Data sets of marine $\delta^{13}C_{POC}$ improve our understanding of marine carbon cycling by providing another independent constraint. Recent model approaches support long-term past climate projections (Tjiputra et al., 2020) and assess estimations of the Suess effect (Liu et al., 2021). To date, numerous individual $\delta^{13}C_{POC}$ data sets exist, while the number of accessible, merged data sets is lacking. Existing merged data sets contain data from several sources but have often been focused on a specific region or process (e.g., Goericke, 1994; Tuerena et al., 2019). Individual data sets are usually collected during a specific cruise or time series station and are often neglected since they contain relatively few data. Such data sets can easily be accessed on data platforms such as PANGAEA and, when combined, they can represent an important and significant source of data.

In this study, we provide a novel merged seawater $\delta^{13}C_{POC}$ data product (Verwega et al., 2021) that – to our knowledge – contains the most expansive spatio-temporal coverage to date. It contains all available $\delta^{13}C_{POC}$ seawater data from PANGAEA and the merged data sets by Goericke (1994), Tuerena et al. (2019) and Young et al. (2013), as well as unpublished data from different cruises by Anne Lorrain. No data were excluded, even if sampled at extreme locations (e.g., trenches, hydrothermal vents). The metadata comprise information about the sampling location, time, depth and method as well as the original source, which makes original raw data values, methods, and further technical description easily accessible. Provided data files are Network Common Data Form (NetCDF) files interpolated onto two different global grids and a csv file that includes the data and their anomalies with respect to their overall mean together with all corresponding available meta-information.

The paper is structured as follows: we provide a brief overview of $\delta^{13}C_{POC}$ data acquisition in Sect. 2 and their compilation and metadata in Sect. 3. The characteristics of the collected $\delta^{13}C_{POC}$ data are shown in Sect. 4. We present their spatial distribution in Sect. 5 and temporal distribution in Sect. 6. Lastly, we provide a short summary and concluding remarks.

2 Data acquisition

The data set includes 4732 entries for $\delta^{13}C_{POC}$ from 185 different sources and ranges from the 1960s to the 2010s. In addition to many data sets from the data platform PANGAEA, we included unpublished data provided by Anne Lorrain and the data products from Tuerena et al. (2019), Goericke (1994) and Young et al. (2013). The adjustments that we conducted are described in the following.

2.1 Data sources

As a basis of our data set, we chose the 1990s data collection by Goericke (1994). This was established to investigate variations in $\delta^{13}C_{POC}$ with temperature and latitude. The $\delta^{13}C_{POC}$ sample data and measurements were conducted by investigating zooplankton, net plankton or particulate organic matter. We cross-checked and extended this data set by looking up all available primary sources. Goericke (1994) originally included 476 $\delta^{13}C_{POC}$ data points from 17 contributions. The largest contributions came from Fischer (1989) with 107 entries, Fontugne et al. (1991) with 97, and Fontugne and Duplessy (1981, 1978) with 78. Large extensions were possible, e.g., in the Fischer (1989) and Eadie and Jeffrey (1973) data sets, incorporating more than 70 additional data points from these primary sources. With this extension, we could increase the data set to 626 data points for $\delta^{13}C_{POC}$.

We collected most data from the PANGAEA data platform, an open-access online library archiving and providing geo-referenced Earth system data, hosted and monitored by the Alfred-Wegener-Institut (2020) – Helmholtz Centre for Polar and Marine Research (AWI) – and the Center for Marine Environmental Sciences, University of Bremen (MARUM). With the data made available therein, we could further extend the data set by an additional $\sim$ 3500 measurements of $\delta^{13}C_{POC}$. Most $\delta^{13}C_{POC}$ data from PANGAEA are associated with samples collected during the Joint Global Ocean Flux Study (JGOFS, 2020), with more than 2000 $\delta^{13}C_{POC}$ data points. Additionally, 529 samples are contributions by the Antarctic Environment and Southern Ocean Process Study (AESOPS, 2020), 342 are by the Archive of Ocean Data (EurOBIS Data Management Team, 2020) and 279 are by the SFB313 research project (Thiede et al., 1988).

Other collected data were provided by Robyn Tuerena and Anne Lorrain. Robyn Tuerena provided a data contribution coming from the data set mentioned in Tuerena et al. (2019), which we will refer to as the Tuerena data set. This contains 595 data points including 501 from Young et al. (2013) and covers samples within the euphotic zone and an observation time frame of 1964–2012. Moreover, we included 69 unpublished data points provided by Anne Lorrain, covering the years 2012–2015 and sampled during the cruises CASSIOPEE, PANDORA, OUTPACE, NECTALIS-3, NECTALIS-4 and KH-13. We refer to this data set as the Lorrain data set.

A recent collection of 303 measurements of $\delta^{13}C_{POC}$ has been provided by Close and Henderson (2020), largely based on data gathered from individual publications referenced therein. Since our analyses originally relied on data sources that differed from those of Close and Henderson (2020), we find our collection to be as yet incomplete. Especially measurements from national databases might provide a huge future benefit.

2.2 Adjustments made

All data were taken with as many details as possible from the sources and have been reshaped to fit the structure of the data set. No rounding or cutoff of detailed data was

made. Spatial coordinates originally given as depth intervals were replaced by their respective midpoints. Time intervals were not changed in this way. If they contained just 1 month or year, this was included; otherwise the time information was omitted. Sample depth given as "surface" was denoted as 1 m. Longitude values were converted to the format $[-180°, 180°]$ by the transformation

$$\mathrm{Long}_{\mathrm{new}} = \begin{cases} \mathrm{Long}_{\mathrm{old}} - 360° & \text{for all } \mathrm{Long}_{\mathrm{old}} \in (180°, 360°] \\ \mathrm{Long}_{\mathrm{old}} & \text{otherwise.} \end{cases} \quad (2)$$

Wherever possible the data were taken from their original publication. Changes made to the data by Goericke (1994) are described in Table 1 and changes to all other data in Table 2. The complete structure is presented in Table 3.

Most data listed in the Goericke (1994) data set could be gathered from the original publications directly. Some data are not accessible from an original source, including those data labeled as "Harrison", "Hobson" and "Schell", which were included as unpublished data by personal communication in Goericke (1994). Also, we could not identify the original data sources of "Voss (1991)" and "Sackett et al. (1966)". Data from these sources are used as provided by Goericke (1994). All other data could be directly compared with and linked to their origin. According to Table 3 we complemented the data with the month, year, depth, sample method, cruise, trap duration and references wherever available. Special notes given in Goericke (1994) were conserved in our "project/cruise"-named meta-information. Rounded values were adjusted to their source values as well as data with interchanged longitudinal information, which is shown in detail in Table 1.

In two cases we identified multiple $\delta^{13}C_{POC}$ data sets from a single event (time, place, investigator) where the data had been subject to different stages of processing or different types of measurement: in Westerhausen and Sarnthein (2003), we chose the "mass spectrometer" data set because this was the originally measured one. In Trull and Armand (2013a, b), we used the "blank corrections" data set of $\delta^{13}C$, since this set of $\delta^{13}C_{org}$ values is recommended to be considered (Trull and Armand, 2001).

The primary source of the Tuerena and Lorrain data was mentioned in our data set in the project/cruise column. In the data set from Tuerena et al. (2019), this was originally labeled as "source" and in the Lorrain data set as "campaign". In both data sets the longitude was converted to $[-180°, 180°]$ from a $[0°, 360°]$ format by Eq. (2). In the data of MacKenzie et al. (2019) we deleted a typo where the depth value was set equal to the negative longitude value. We disregarded the trap duration given in Voss and von Bodungen (2003), which was given as the negative value -1.

3 Content and structure of the data set

The data collection is made available in files of raw and interpolated values (Verwega et al., 2021). The raw data are in a csv file that includes the $\delta^{13}C_{POC}$ measurements, their anomalies with respect to their mean and all available meta-information. The interpolated data are provided as NetCDF files on two different global grids: a $1.8° \times 3.6°$ resolution and 19 depth layers from a model that simulates $\delta^{13}C_{POC}$ (e.g., Schmittner and Somes, 2016), in the following referred to as the UVic grid, and the $1° \times 1°$ resolution and 102-depth-layer grid of the World Ocean Atlas (Garcia et al., 2018), in the following referred to as the WOA grid. Interpolation required the availability of the full spatial information (latitude, longitude and depth) of included $\delta^{13}C_{POC}$ data to locate them on the grid.

On the WOA grid we provide 13 NetCDF files containing only data with full spatio-temporal metadata: one averages all observations from each year together, each year accounting for a time increment on the time axis, and the other 12 files average only observations from an individual month with again each year accounting for a time increment on the time axis. These files provide a variety of analysis opportunities but also limited the content of $\delta^{13}C_{POC}$ data.

On the UVic grid we provide seven individual NetCDF files: six of them each represent one of the decades from the 1960s to the 2010s containing all data which were able to be assigned to their respective decade. One file contains all available $\delta^{13}C_{POC}$ data completely independent of their measurement time. This individual provision of data on a decadal and overall timescale increases the fraction of usable $\delta^{13}C_{POC}$ data for the following analyses.

3.1 Raw data file

The csv-format data file includes $\delta^{13}C_{POC}$ measurements, anomalies and meta-information in its columns. A full description of the content, value range and coverage of the individual columns is given in Table 3. Anomalies of $\delta^{13}C_{POC}$ were calculated, based on the arithmetic mean of the full data collection. The mean was calculated, rounded to two digits after the floating point and used as

$$\mathrm{mean}_{\delta^{13}C_{POC}} = -23.96‰. \quad (3)$$

Anomalies contain all the relevant information with respect to the variability in the $\delta^{13}C_{POC}$ data in space and time. This way it becomes easier to analyze bias information separately, e.g., during the first steps of model calibration.

The reference includes the citations in as much detail as possible. Wherever available, this is taken from the original source. Otherwise, we tried to include the author, title, publication year and platform, and DOI. For unpublished data like Harrison's (unpublished data, quoted from Goericke, 1994) from the Goericke (1994) data set or those included by the co-authors, we specified from where we took the data.

Coordinates are given in decimal degrees over $[-90°, 90°] \times [-180°, 180°]$. The sample depth is given in meters measured positively from the ocean surface downwards. Data published as measured at 0 m were included

Table 1. Changes that were introduced into data taken from Goericke (1994): the first column names the publication or author of the primary data set. The second column lists in which part of the data we applied changes. The third and fourth columns show what these changes were, and the last column gives the reason for this.

Data set	Changed	From	To	Reason
Degens et al. (1968)	long	Goericke (1994)	source value	E and W interchanged
Eadie and Jeffrey (1973)	long	Goericke (1994)	source value	E and W interchanged
Fischer (1989)	long	Goericke (1994)	source value	E and W interchanged
Fontugne and Duplessy (1978)	long	Goericke (1994)	source value	E and W interchanged
Fontugne and Duplessy (1981), MD13 Osiris III	long	Goericke (1994)	source value	E and W interchanged
Francois et al. (1993)	long	Goericke (1994)	source value	E and W interchanged
Harrison*	long	Goericke (1994)	source value	E and W interchanged
Sacket et al. (1965)	long	Goericke (1994)	source value	E and W interchanged
Saupe et al. (1989)	long	Goericke (1994)	source value	E and W interchanged
Wada et al. (1987)	long	Goericke (1994)	source value	E and W interchanged
Eadie and Jeffrey (1973)	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Fischer (1989) all but INDOMED leg 12	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Fontugne and Duplessy (1978)	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Fontugne and Duplessy (1981)	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Francois et al. (1993)	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Sacket et al. (1965)	lat, long	Goericke (1994)	source value	rounded in Goericke (1994)
Eadie and Jeffrey (1973)	$\delta^{13}C_{POC}$	not included	added	not included in Goericke (1994)
Fischer (1989)	$\delta^{13}C_{POC}$	not included	added	not included in Goericke (1994)
Sacket et al. (1965)	$\delta^{13}C_{POC}$	not included	added	not included in Goericke (1994)
Wada et al. (1987)	$\delta^{13}C_{POC}$	not included	added	not included in Goericke (1994)
Fischer (1989)	$\delta^{13}C_{POC}$	Goericke (1994)	source value	rounded in Goericke (1994)
Fontugne and Duplessy (1978)	$\delta^{13}C_{POC}$	Goericke (1994)	source value	rounded in Goericke (1994)
Fontugne and Duplessy (1981)	$\delta^{13}C_{POC}$	Goericke (1994)	source value	rounded in Goericke (1994)
Fischer (1989)	temperature	Goericke (1994)	source value	rounded in Goericke (1994)
Fontugne and Duplessy (1981)	temperature	Goericke (1994)	source value	rounded in Goericke (1994)
Francois et al. (1993)	temperature	Goericke (1994)	source value	rounded in Goericke (1994)
Sacket et al. (1965)	temperature	Goericke (1994)	source value	rounded in Goericke (1994)
Fischer (1989)	$\delta^{13}C_{POC}$	Goericke (1994)	deleted	not found in source
Fontugne and Duplessy (1978)	temperature	Goericke (1994)	deleted	not found in source

* The original source was not available, but we highly suspected an error in the coordinates that interchanged east and west.

Table 2. Changes made in other data: this table's structure is equivalent that of to Table 1. It refers to all changes made in general and any data other than the Goericke (1994) data.

Data set	Changed	From	To	Reason
Any	depth	"surface"	1	comparability
Any	depth	depth range	average[1]	comparability
Trull and Armand (2013a)	$\delta^{13}C_{POC}$	three available	"blank correction"	mentioned in Trull and Armand (2001)
Trull and Armand (2013b)	$\delta^{13}C_{POC}$	three available	"blank correction"	mentioned in Trull and Armand (2001)
Any using sediment traps	month, year	range	explicit value[2]	comparability
Chang et al. (2013)	month, year	range	explicit number	just one date for trap sampling given
Lorrain	project/cruise		"campaign"	provided by Anne Lorrain
Tuerena	project/cruise		"source"	provided by Robyn Tuerena
Tuerena	long	$[0°, 360°]$	$[-180°, 180°]$[3]	comparability
Lorrain	long	$[0°, 360°]$	$[-180°, 180°]$[3]	comparability
MacKenzie et al. (2019)	depth	original	deleted	suspected typo
Voss and von Bodungen (2003)	trap duration	original	deleted	suspected typo
De Jonge et al. (2015a)	method	multiple investigations (MULT)	in situ pump	found in De Jonge et al. (2015b)

[1] By arithmetic mean. [2] Only for sample durations entirely within an explicit month and year, otherwise information on time frames has been discarded. [3] We applied Eq. (2).

Table 3. Available data and meta-information: the columns of the raw data set correspond to the provided data and meta-information. Their names are given in the first column of this table. The second holds a short description of their content and the third their ranges of values. In the final column we give how well this data kind is covered relative to the size of the full data set.

Column	Content	Range of values	Coverage[1]
Reference	citation[2]	description	full[3]
No.	running index	$\{1, \ldots, 4732\}$	full
Lat	latitude in decimal degrees[4]	$[-90°, 90°]$	4604/4732
Long	longitude in decimal degrees[4]	$[-180°, 180°]$	4604/4732
d13C	$\delta^{13}C^{4}_{POC}$	$[-55.15, -4.5]$	full
d13Canomaly	$\delta^{13}C_{POC} - \text{mean}^{5}_{\delta^{13}C_{POC}}$	$[-31.19, 19.46]$	full
Temp	temperature in degrees Celsius[4]	$[-1.8, 31.12]$	1622/4732
Month	month as number	$\{1, \ldots, 12\}$	4114/4732
Year	year CE	$\{1964, \ldots, 2015\}$	4483/4732
Depth	depth in meters	$[0, 4850]$	3917/4732
Method	measurement method of $\delta^{13}C_{POC}$	description	3164/4732
Origin	associated project or cruise	description	3921/4732
Note	special circumstances	description	140/4732
Trap duration	duration of trap activity in days	$[1, 133]$	533/587[6]

[1] Ratio of available entries relative to the full number of data points. [2] Wherever possible, this includes: author(s), year, title, journal name, full, number, issue, pages and DOI. [3] Primary source was not available in every case as a reference. A note, where the data were taken is included in this case. [4] With as many decimal places as available. [5] Rounded to two decimal places. [6] Here, abundance is given relative to the full number of sediment trap samples.

as this, while no surface microlayer measurements were included. The month and year were used to describe the sample date; specific days are neglected.

Anomalies of $\delta^{13}C_{POC}$ are given in the δ ratio described in Eq. (1). A sample method was added, wherever available. Any special sampling circumstances were given in the "Note" column. Activity duration of sediment traps was denoted in the last column.

The "Origin" columns listed the associated project or cruise or author note. Some samples were given with multiple project connections; all of them were given in this column.

3.2 Interpolated data sets

The interpolated $\delta^{13}C_{POC}$ data are available as NetCDF files on two global grids with different resolutions. NetCDF files are machine-independent and support the creation, accessing and sharing of array-oriented scientific data. On the UVic grid, we provide seven different files, each of them independent of time and averaged over the available spatial information. Six of them contain an individual decade each (from the 1960s through the 2010s). The seventh file comprises a combined set of all interpolated $\delta^{13}C_{POC}$ data. On the WOA grid, we provide 13 files including all $\delta^{13}C_{POC}$ measurements with complete spatial–temporal information, averaged across time and space.

One major aim of this work is to support reliable validation and calibration of $\delta^{13}C_{POC}$-simulating models. Hence, we chose the grid of the UVic model version 2.9, as used, e.g., in Schmittner and Somes (2016). Horizontally, it consists of 100×100 cells with a resolution of $1.8° \times 3.6°$, arranged from 0 to 360° in longitude (LONG) and -90 to 90° in latitude (LAT). Vertically, it is split up into 19 vertical layers (DEPTH), decreasing in resolution with depth. The two uppermost layers reach down to depths of 50 and 130 m, respectively, and they are supposed to comprise the upper ocean's euphotic zone.

The WOA grid is based on the $1° \times 1°$ grid of the World Ocean Atlas (Garcia et al., 2018). It has a horizontal resolution of 360 arranged from -180 to 180° in the longitude (LONG) and 180 arranged from -90 to 90° in the latitude (LAT) direction. Vertically, it is split up into 102 layers (DEPTH). The time axis (TIME) increases in increments for each year from 1964 to 2015 by 1 and has a size of 52. This interpolation includes only $\delta^{13}C_{POC}$ data with full spatio-temporal metadata coverage; i.e., additionally to latitude, longitude and depth, we also required and included year and month information.

Ferret scripts were used for the interpolations. These averaged the irregularly measured data points within the ocean grid to one single data point representing each covered grid cell. The interpolation function SCAT2GRIDGAUSS by NOAA's Pacific Marine Environmental Laboratory (2020) performed the spatial averaging under PyFerret v7.5. Calculations in this function are based on a work by Kessler and McCreary (1992) and can be summarized as follows: let $(x_1, y_1), \ldots, (x_n, y_n) \subseteq \mathbb{R}^2$ be an equidistant grid and $(\tilde{x_1}, \tilde{y_1}), \ldots, (\tilde{x_m}, \tilde{y_m}) \subseteq \mathbb{R}^2$ be irregular measurement locations of a real tracer D_j, $j \in \{1, \ldots, m\}$. Then the value $D_i \in$

$\mathbb{R}$ at grid point $(x_i, y_i), i \in \{1, \ldots, n\}$ becomes interpolated as

$$D_i := \frac{\sum_{j=1}^{m} D_j W_{i,j}}{\sum_{j=1}^{m} W_{i,j}}, \tag{4}$$

where

$$W_{i,j} := \begin{cases} 0; & \tau_{i,j} < e^{-CX} \\ 0; & \tau_{i,j} < e^{-CY} \\ \tau_{i,j} & \text{otherwise,} \end{cases} \tag{5}$$

where $\tau_{i,j} := \exp\left(-\left(\frac{(x_j - x_i)^2}{X^2} + \frac{(y_j - y_i)^2}{Y^2}\right)\right)$ is the Gaussian weight function, $X, Y \in \mathbb{R}$ comprises scaling arguments and $C \in \mathbb{R}$ the cutoff parameter. We set $X = 1.8$, $Y = 0.9$ and $C = 1$ in our script.

Since the interpolation into the WOA grid excluded all data without full spatio-temporal metadata coverage, we focus the following descriptions of interpolated data on the UVic grid interpolations. These also include data without month information in the six decadal files and even completely without temporal information in the seventh time-independent file.

4 Main data set characteristics

The final data set includes 4732 individual $\delta^{13}C_{POC}$ measurements of seawater samples. We show the distribution of $\delta^{13}C_{POC}$ values by Gaussian kernel density estimation (KDE) in Fig. 1. KDEs are non-parametric density estimations (Silverman, 1986) for the approximation of probability density functions, which are theoretically similar to histograms but with continuous curves not dependent on rigid intervals. We applied a Python implementation from the SciPy stats package (Virtanen et al., 2020) to create the results presented here. Likewise, we derived conditional probability densities of $\delta^{13}C_{POC}$ values, given the different measurement method applied (Fig. 3).

4.1 Range and outlier values

The data distribution is presented by its KDE in Fig. 1. The interval of $\delta^{13}C_{POC}$ values ranges over $[-55.15, -4.5]$ with a mostly smooth distribution. Most of our data exhibit values around $\delta^{13}C_{POC} \approx -24\,‰$, which becomes clearly identifiable as a single maximum in the KDE. Two smaller modes are visible at around $\delta^{13}C_{POC} \approx -27.5\,‰$ and $\delta^{13}C_{POC} \approx -22\,‰$ (see also Table A1 in the Appendix). A steep decline to zero is visible outside the two outer modes. The steep decline in the KDE stops at around $\delta^{13}C_{POC} = -37\,‰$ and $\delta^{13}C_{POC} \approx -14\,‰$. Between $\delta^{13}C_{POC} \approx -37\,‰$ and $\delta^{13}C_{POC} \approx -55.15\,‰$ as well as between $\delta^{13}C_{POC} \approx -14\,‰$ and $\delta^{13}C_{POC} \approx -4.5\,‰$ the KDE closely aligns to the x axis, which indicates very few data points lie in this range.

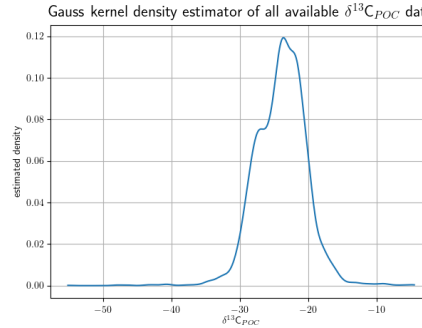

Figure 1. The density function of all individual $\delta^{13}C_{POC}$ measurements approximated by Gaussian kernel density estimation: values of the estimated density are drawn on the y axis; the $\delta^{13}C_{POC}$ values run on the x axis. The higher the value of the estimated density is, the more $\delta^{13}C_{POC}$ points have been measured around this value.

Below $\delta^{13}C_{POC} = -37\,‰$ we find 17 data points ranging down to $\delta^{13}C_{POC} = -55.15\,‰$. Down to $\delta^{13}C_{POC} = -48‰$ these were all taken from Lein and Ivanov (2009) and Lein et al. (2006), measured in September or October 2003, around the location 10° N, 104° W and below 2500 m depth in the vicinity a hydrothermal field close to the Pacific coast of middle America. The lowest outlier at $\delta^{13}C_{POC} = -55.15\,‰$ was taken from Altabet and Francois (2003a) from November 1996 and at 62.52° S, 169.99° E at the ocean surface south of New Zealand.

Above $\delta^{13}C_{POC} = -10\,‰$ we find 15 data points ranging up to $\delta^{13}C_{POC} = -4.5‰$. Three of them were taken from Lein et al. (2007) and measured at 800 m depth at a hydrothermal vent located 30.125° N, 42.117° W in the middle north Atlantic. Ten were taken from Calvert and Soon (2013b, c, a). All of these were measured between 636 and 901 m depth around 49° N, 130° W close to the American coast of the Pacific, and all of them were measured in February or May, except one in August. The final two were part of the Lorrain data set. Both were measured at the ocean surface in the South Pacific, in July at 5.3° S, 164.9° E and December at 20.9° S, 159.6° E.

Since more than 98 % of the data (4668 of the 4732 data points) have values that lie between $\delta^{13}C_{POC} = -35\,‰$ and $\delta^{13}C_{POC} = -15\,‰$, we will focus on this range in our following analyses.

We tested the robustness of our KDE approach in a subsampling experiment. We considered 500 random subsets of 20 % of the original data over the range with the highest data density $[-35, -15]$ and visualize their KDEs in Fig. 2. They show peaks at $\delta^{13}C_{POC} \approx -23\,‰$, fitting the maximum and the second smaller mode to the right of it, and

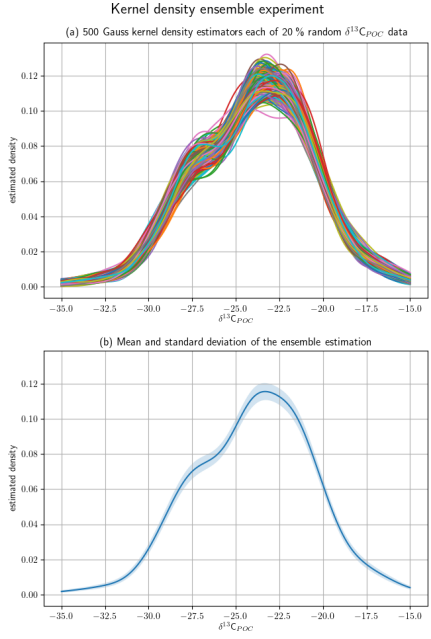

Figure 2. A random sample of 20 % of the $\delta^{13}C_{POC}$ data was taken from the full data set for 500 times to generate an ensemble of subsets. Their densities were approximated with a Gaussian kernel density estimator. Panel **(a)** shows all 500 estimated densities by individual lines. Panel **(b)** shows the mean and the variance of the full ensemble of densities by a graph and the shaded area around it, respectively.

at $\delta^{13}C_{POC} \approx -27.5$ ‰. Outside $[-27, -22]$ the KDEs are closely aligned. The mean of and standard variation in the KDE ensemble also show the highest variability around the two modes at $\delta^{13}C_{POC} = -23$ ‰ and $\delta^{13}C_{POC} = -27.5$ ‰.

4.2 Sampling methods

Various sampling methods were involved in obtaining the $\delta^{13}C_{POC}$ data. Around 67 % of the data had associated sampling-method information, which included 18 different sampling methods. In principle, all 18 methods could be grouped into five main observational types: bottles, intake, nets, traps and diverse. "Bottle" data include samples taken from Niskin bottles and samples collected via Sea-Bird submersible pumps. By "intake" we refer to all versions of pumps and underway cruise track measurements, as well as multiple-unit large-volume filtration systems (MULVFSs). "Net" data represent all occurring versions of plankton nets, and "traps" refers to all represented sediment traps and moorings. Finally, the deep-sea manned submersible (MIR2) is not classified into any of these groups and was assigned to a cluster that we refer to as "diverse".

All sample devices provided data over all sample depths. Deeper samples were mainly taken from traps and pump systems and the upper samples from bottle and net data. Most data sampled deeper than 2600 m were collected by sediment traps. At 3800 m there were several trap contributions by Calvert (e.g., Calvert, 2002), mostly from the late 1980s. Data sampled by a deep-sea manned submersible were given at locations down to 2520 m (Lein and Ivanov, 2009).

For resolving differences between sampling methods we chose data from the Atlantic Ocean which comprise all four major methods (with data embracing a region between 45° S and 80° N and 70° W and 20° E). In addition, data were distinguished by tropical, temperate and polar subregions. By crudely sorting the data according to their sampling locations, we gain some insight into methodological variability within a subregion and may relate this to variations between the three subregions (Fig. 3). Overall, we do not find any severe bias with respect to any particular method. Bottle data seem to cover most of the lower $\delta^{13}C_{POC}$ values that typically range between −28 ‰ and −21 ‰, which could be due to samples collected at greater depths. Intake and net measurements are rather restricted to the upper ocean layers, and these methods often yield $\delta^{13}C_{POC}$ values larger than −25 ‰, with some polar net measurements being a notable exception (Fig. 3d). For the tropical Atlantic (30° S–30° N) the net and intake measurements vary around −21 ‰, with 95 % confidence limits between −24 ‰ and −18 ‰ (see Table A2 in the Appendix). According to our comparison, we could not identify any method that yields much greater variance of $\delta^{13}C_{POC}$ values than others. The spatio-temporal variations of the $\delta^{13}C_{POC}$ compare well among different methods, but we advise caution when comparing bottle measurements with data of other methods because of potential differences in the depth range covered.

In the full Atlantic Ocean, densities of intake and net data are most representative of the maximum full $\delta^{13}C_{POC}$ sample. From the intake data shown here, ∼ 80 % were sampled within 30° S and 30° N. When restricting data to this area, net data better resemble the full data. Of all net sample data, ∼ 80 % were collected between 30 and 60° N, where they fit the overall $\delta^{13}C_{POC}$ density best, followed by trap data. Trap and bottle data deliver the lowest $\delta^{13}C_{POC}$ measurements in the Atlantic Ocean. Of both data kinds, ∼ 74 % to 85 % were sampled north of 60° N. A restriction to this area shows trap and bottle samples being closely aligned to the full data in this region.

The variance of the intake and trap data is ∼ 3 ‰ and lower than the variance of all $\delta^{13}C_{POC}$ together, which is

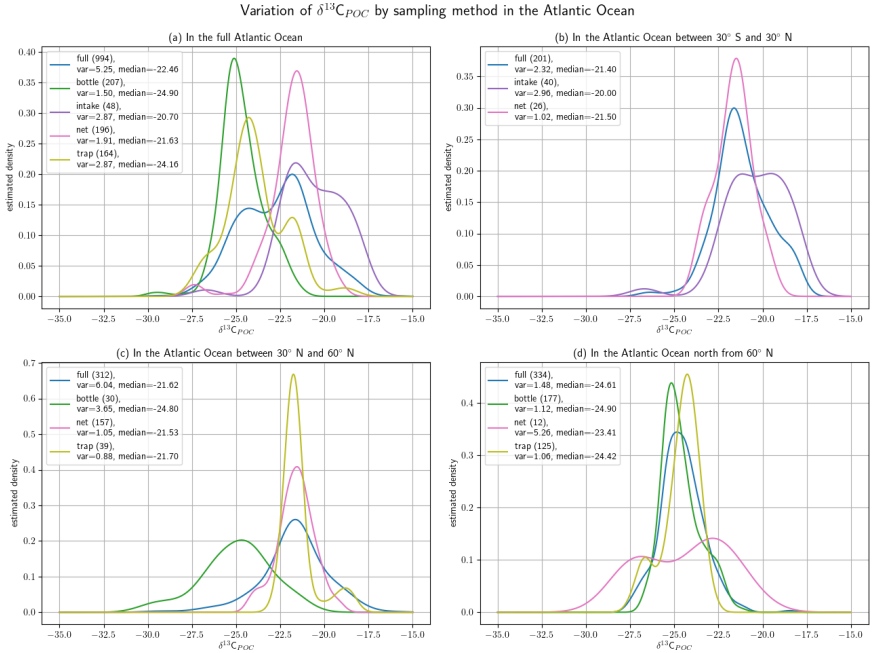

Figure 3. Separation of $\delta^{13}C_{POC}$ in the Atlantic Ocean data by four main sample methods: bottle, intake net and trap data. Panel **(a)** shows the full Atlantic Ocean, panel **(b)** the equatorial core of the Atlantic Ocean, panel **(c)** the Atlantic between 30° S and 30° N, and panel **(d)** its most northern area. In each plot, the density of the $\delta^{13}C_{POC}$ sample groups with enough data was approximated by Gaussian KDEs and drawn with an individual color. An additional graph shows the comparison to the full-$\delta^{13}C_{POC}$-data density in the respective area. The numbers of used data points are indicated in each KDE label.

$\sim 5\,‰$, the highest value observed here. Both bottle and net data show a variance of less than $2\,‰$. Furthermore, trap, net and full $\delta^{13}C_{POC}$ show a pronounced second mode in their densities, while bottle and net data show a clear individual maximum. Median values of net and intake data are $\sim 1\,‰$ to $\sim 2\,‰$ higher, respectively, than the one of the full data. This has a median of $\delta^{13}C_{POC} = 22.46\,‰$. Both bottle and trap data show a $\sim 2\,‰$ lower median. Analytical errors and uncertainties are typically $0.2\,‰$ or lower (Young et al., 2013) and thus are not likely to significantly contribute to the much larger variance in the observations

5 Spatial distribution

We show the spatial distribution of $\delta^{13}C_{POC}$ measurements across the global ocean surface and depths. Most $\delta^{13}C_{POC}$ data have been measured in the uppermost few ocean meters, and the best surface coverage is available for the Atlantic Ocean. Changes in $\delta^{13}C_{POC}$ on the ocean surface were evaluated based on the UVic grid.

5.1 Vertical distribution of the data set

Depth values are available for more than 80 % of the sample data with most of them located in the upper ocean. The distribution of depth measurements is shown in Fig. 4. An approximation of the depth measurements by Gaussian KDE is visualized in Fig. 5 along with the $\delta^{13}C_{POC}$ value distribution over them in the main ocean basins. The KDE resolves the best data coverage for the uppermost ~ 500 m of the oceans and a second far smaller maximum at ~ 3800 m. The depth ranges presented in Fig. 4 correspond to the depth intervals of

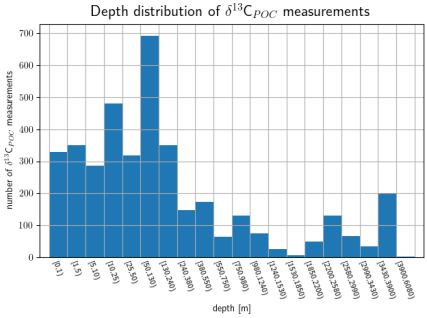

Figure 4. Vertical data coverage in depth layers based on the UVic grid: the uppermost 50 m is divided into subranges; below they are according to the UVic grid. The number of $\delta^{13}C_{POC}$ data points available is plotted against its respective depth range.

the UVic grid; only the two uppermost layers are presented in more detail, and the last four are combined. Within the first 130 m we observe the highest data density and find nearly 2500 measurements of $\delta^{13}C_{POC}$, where nearly 1000 of them were measured within [0 m, 10 m). A total of 200 $\delta^{13}C_{POC}$ values were available in the depth interval [3430 m, 3900 m). The two deepest values were taken from Fischer (1989) and Altabet and Francois (2003b) and sampled at 4500 and 4850 m depth, respectively.

Values of $\delta^{13}C_{POC}$ are, apart from in the North Pacific, closely aligned within the individual ocean basins. The Atlantic, South Pacific and Indian Ocean show values mostly of $-28‰$ to $-19‰$. The $\delta^{13}C_{POC}$ values in the Arctic reach down to approximately $-30‰$ and those in the Southern Ocean even to approximately $-35‰$. The North Pacific shows a wide spread of $\delta^{13}C_{POC}$ values, especially between 50 and 100 m depth and at 2500 m depth. There they reach either down to less than $-40‰$ or up to more than approximately $-10‰$ at a depth of 2500 m.

Measurements in the North Atlantic, North Pacific and Indian Ocean reach down to more than 3500 m. Measurements down to nearly 5000 m were sampled in the Southern Ocean. The South Pacific was sampled down to a depth of 2500 m and the Arctic Ocean and South Atlantic only in the uppermost few hundred meters.

5.2 Horizontal distribution of the data set

All global oceans are covered with $\delta^{13}C_{POC}$ data. In Fig. 6 the horizontal distribution of available data is depicted for both grids. For the UVic grid we show data from the file including all data independent of time; the WOA grid is averaged over all times. In both cases, we averaged data over all depths and also added data without depth information to best visualize the horizontal coverage. A similar plot, although with a different purpose, is given later in this work in Fig. 10 showing only surface data locations.

Many cruises are visible as lines formed by connected grid cells in Fig. 6, especially in the Atlantic and Indian Ocean and less so in the Southern Ocean. Also, smaller sample spots occur, mainly located in the Pacific, Arctic and Southern Ocean. The Atlantic Ocean provides the best data coverage. Then the Southern and Indian oceans contain the next best coverage with the North Pacific having the sparsest.

The highest $\delta^{13}C_{POC}$ values are evident in low-latitude regions. In the Atlantic Ocean the highest values were measured between 0–30° N and 30–60° W as well as close to the western coast of France, reaching up to at least $-17‰$. The Indian Ocean shows generally high values of approximately $-20‰$. In the Pacific Ocean the highest values are close to the Peruvian coast and Papua New Guinea. We also find high values in the Bering Strait and on the northern edge of the Southern Ocean at around 65° E.

The lowest $\delta^{13}C_{POC}$ values are mostly found in the Southern Ocean. Nearly all measured grid cells here belong to $\delta^{13}C_{POC}$ values lower than around $-28‰$. The Arctic Ocean shows low values as well, for instance in the Kara Sea. The lowest values in the Pacific Ocean occur in the Southern Ocean at high latitudes.

5.3 Meridional trend of $\delta^{13}C_{POC}$ values

We show the north–south trend of $\delta^{13}C_{POC}$ over the Atlantic Ocean based on the time-independent UVic grid and restricted to the uppermost 130 m, which resembles the euphotic zone in the UVic model. We chose this section due to it having the best data coverage. A biome mask according to Fay and McKinley (2014) was applied to the gridded data, thereby defining latitudinal zones in the entire Atlantic Ocean. Distributions of $\delta^{13}C_{POC}$ within the biomes are shown in Fig. 7 (see also Table A3 in the Appendix).

The biomes derived by Fay and McKinley (2014) are areas with consistent biological and ecological properties. The chosen biomes cover the Atlantic Ocean and extend to the Arctic Sea and parts of the Southern Ocean. The biomes are numbered 9 to 17, excluding 14. The biomes 15 to 17 represent parts of the Southern Ocean and were restricted to 70° W and 20° E. Their locations are shown in Fig. 7.

Observations by the biomes are consistent with the ones from Fig. 6. The two biomes showing the lowest $\delta^{13}C_{POC}$ values from $-28‰$ to $-29‰$ are those located farthest south. The biome located farthest north contains the next-lowest values of about $-24‰$. The biomes with more positive $\delta^{13}C_{POC}$ values are in the lower latitudes and show similarly higher values from $-23‰$ to $-21‰$.

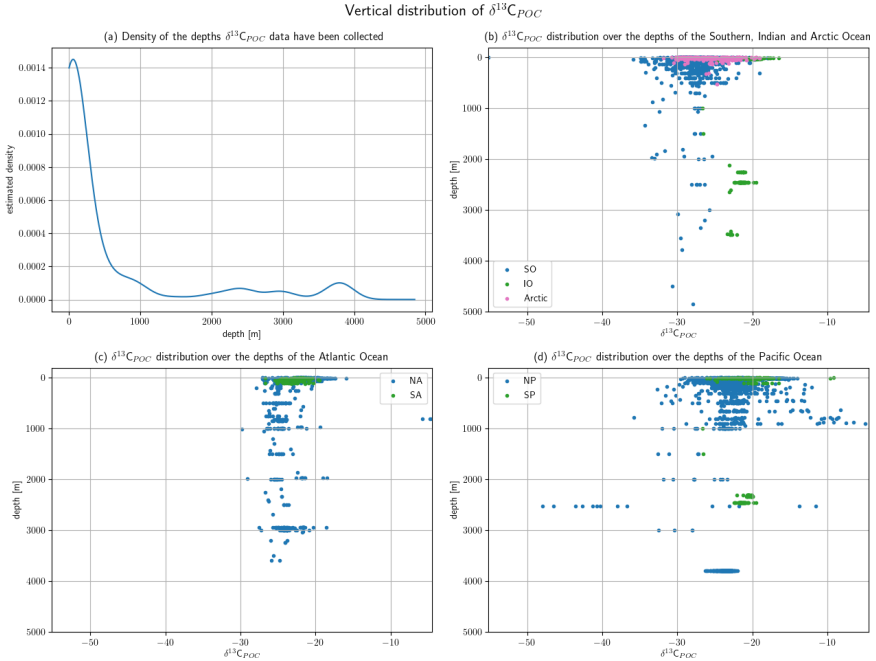

Figure 5. The vertical distribution of available $\delta^{13}C_{POC}$ samples is shown **(a)** as the approximated density of the measurement depths and **(b–d)** as measured $\delta^{13}C_{POC}$ values relative to their respective measurement depth. Panel **(a)** provides the estimated density of the depth values on the y axis and the depth in meters on the x axis. The estimation was realized by a Gaussian KDE. Panel **(b)** resolves the measurements of the Southern, Indian and Arctic Ocean, **(c)** the North Atlantic and South Atlantic, and **(d)** the North Pacific and South Pacific. The last three panels show the depth in meters on the y axis and the measured $\delta^{13}C_{POC}$ value on the x axis. Different colors are used to mark different ocean basins.

6 Temporal distribution of the data set

The full $\delta^{13}C_{POC}$ data cover a time period of around 50 years over 1964–2015 and all 12 calendar months. The number of samples measured during individual decades varies considerably with most measurements in the 1990s. Coverage within the months is quite comparable; only winter months in both hemispheres exhibit fewer data.

The distribution of $\delta^{13}C_{POC}$ samples over the years is resolved in Table 4 and is visually approximated by Gaussian KDE in Fig. 8. The 1990s shows the best data coverage. More than half of the data points are associated with a year in this decade, which is visible by a pronounced maximum in the estimated density. The sparsest data are found in the 1960s, when only 74 data points were sampled. All other decades come with between around 300 and 600 $\delta^{13}C_{POC}$ data points. The latest data are mostly from Anne Lorrain, MacKenzie et al. (2019) and Kaiser et al. (2019). The oldest data were taken from the data sets by Robyn Tuerena, Degens et al. (1968), and Eadie and Jeffrey (1973).

6.1 Monthly variations

Monthly clustered data of the Northern Hemisphere and Southern Hemisphere show monthly variations, but more observations are required to demonstrate robust seasonality within different regions. Since more than 50 % of the available $\delta^{13}C_{POC}$ data originate in the 1990s, we selected data from this decade to exclude changes that might be introduced by longer-term trends. Furthermore, we restricted our data

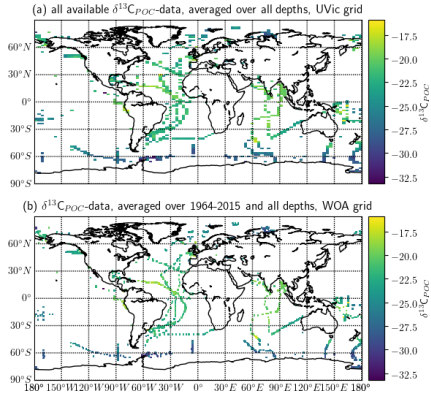

Figure 6. Global distribution of the $\delta^{13}C_{POC}$ data is visualized based on the **(a)** UVic grid and **(b)** WOA grid. The data used for **(a)** are independent of time and include all available measurements with latitude and longitude information. The data shown in **(b)** include only data with complete temporal metadata and are averaged over the years 1964–2015. Both kinds of data are averaged over all measurements including data with missing depth information. Each colored square refers to a grid cell with available $\delta^{13}C_{POC}$ measurements. The colors indicate the $\delta^{13}C_{POC}$ value in the respective grid cell.

Table 4. Data coverage within the available decades: the first column lists the available decades and the second column the number of sampled $\delta^{13}C_{POC}$ data points within this time frame.

Decade	$\delta^{13}C_{POC}$ values available
1960s	74
1970s	321
1980s	463
1990s	2403
2000s	614
2010s	589

to the uppermost 130 m, which resembles the euphotic zone in the UVic model. In Fig. 9 we displayed all months with enough data points by a KDE and indicate the same months by the same colors. We excluded July, November and December in the Northern Hemisphere from this KDE representation because these months provided three or fewer data points each, which resulted in a KDE that overgrew the others by magnitudes and made their visual comparison difficult. The KDEs are supported by comparison of the median values of the individual months in Table 5.

The monthly resolved variations in $\delta^{13}C_{POC}$ do not reveal any significant seasonal pattern (Fig. 9; see also Table A4 in the Appendix). In general we find the highest $\delta^{13}C_{POC}$ values in the Northern Hemisphere, with a median $\delta^{13}C_{POC}$ of $-20.4\,‰$ in April and a median $\delta^{13}C_{POC}$ of $-21.5\,‰$ in October, which are typical months with enhanced primary production (Northern Hemisphere spring and autumn blooms). Similarly high median $\delta^{13}C_{POC}$ values cannot be ascertained for any month with data of the Southern Hemisphere, where values of $\delta^{13}C_{POC}$ above $-20\,‰$ have rarely been observed at any time of the year. In fact, there is an overall tendency towards low $\delta^{13}C_{POC}$ values for the Southern Hemisphere, which becomes well expressed during the months April and September, with medians of $\delta^{13}C_{POC} = -28.1\,‰$ and $\delta^{13}C_{POC} = -28.5\,‰$, respectively. However, interpretations of this north–south trend should be treated with caution because the apparent tendency is likely conditioned by some imbalance in the number of high-latitude data points. Compared to the number of data points from the Southern Ocean, samples from the Arctic Ocean are considerably underrepresented (see also Fig. 10). Furthermore, the discrimination between data of the Northern Hemisphere and Southern Hemisphere is crude, and we encourage the use of our data collection for more advanced analyses of seasonal, monthly based changes in the $\delta^{13}C_{POC}$ signal.

6.2 Decadal variations

The decadal UVic grid NetCDF files are the basis for showing long-term changes in the $\delta^{13}C_{POC}$ data. An overview of where the data within the individual decades were sampled is given in Fig. 10. This shows that the sparsest coverage was obtained in the 1960s, located close to the central American continent. Most data in the Indian Ocean were sampled in the 1970s. A cruise across the southern part of the Atlantic Ocean up to 30° N and some samples close to Iceland were also measured in this decade. The 1980s is similarly sparse in spatial coverage to the 1960s. Measurements of the 1980s were taken at locations in the Southern Ocean, in the Arctic and in the Atlantic close to the Equator. The 1990s has the best coverage including most ocean basins. Most Southern Ocean data were sampled within the 1990s. The 2000s provides good coverage of the Arctic Ocean. Finally, the 2010s data were mostly sampled in the Southern Hemisphere in the open Pacific and Atlantic. A smaller number of Eurasian continental sea data were also part of the 2010s samples.

We show the changes in $\delta^{13}C_{POC}$ values over the available decades by density estimates in Fig. 11 (see also Table A5 in the Appendix) and by their median in Fig. 12. Figure 11 visualizes the sparse coverage of the Southern Ocean outside of the 1990s, which is why the area is not part of any further discussion here. The Southern Ocean is defined as the ocean area south of 45° S. All presented analyses were restricted to the euphotic zone, i.e., the uppermost 130 m resembling the two first layers of the UVic grid.

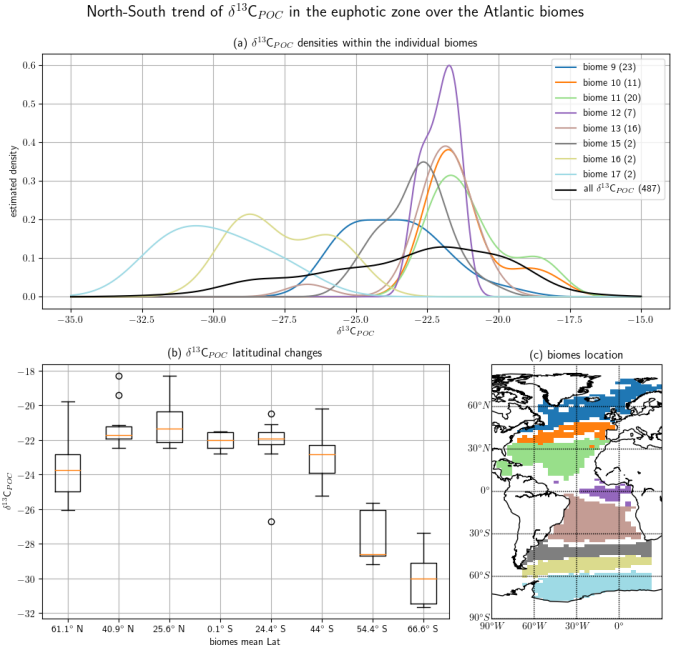

Figure 7. The north–south trend of sampled $\delta^{13}C_{POC}$ values is visualized by a cross section over the Atlantic Ocean. Biomes (Fay and McKinley, 2014) define the latitudinal bands of the interpolated data set. Panel **(a)** presents a Gaussian KDE for each biome approximating the density of the contained $\delta^{13}C_{POC}$ data. Different colors mark the individual biomes, and a black line shows the general global $\delta^{13}C_{POC}$ distribution. The number in parentheses in each KDE label counts the number of $\delta^{13}C_{POC}$ measurements used for the respective graph. Panel **(b)** shows in a box plot the steep decline in $\delta^{13}C_{POC}$ values from the tropical biomes towards the higher latitudes. The x axis provides the mean latitudes of the biomes introduced in **(a)**. The y axis measures the $\delta^{13}C_{POC}$ value. Panel **(c)** shows the biome locations. Each biome is drawn in the color of its corresponding density estimate in **(a)** above. The biome numbers increase from the north to the south.

Table 5. Monthly median change in $\delta^{13}C_{POC}$. Due to their having the best data coverage, the analyses were carried out within the 1990s and in the uppermost 130 m.

Hemisphere	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
North		−24.815	−24.12	−20.	−24.06	−24.7	−21.746	−23.67	−22.83	−21.4	−23.5455	−23.368
South	−26.45	−26.41	−23.34	−28.2					−28.65	−27.95	−27.9	−26.08

A clear decrease in $\delta^{13}C_{POC}$ densities in Fig. 11 can be identified for the global ocean outside of the Southern Ocean. All decades but the 1980s show one clear maximum in their approximated densities. The 1980s shows a second expressed density maximum at lower values. The main maximum shifts from the 1960s at $\delta^{13}C_{POC} \approx -19.9\,‰$ to the 2010s at $\delta^{13}C_{POC} \approx -23\,‰$. This decrease is also clearly visible in the comparison of the decadal medians (Fig. 12). The Southern Ocean provides far worse data coverage. Only the 1980s and 1990s include enough data to construct a comparable KDE. Due to this very low data availability, all of these results must be taken with the highest caution.

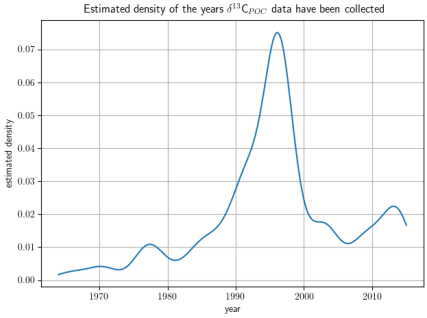

Figure 8. The distribution of $\delta^{13}C_{POC}$ data samples over the years approximated by Gaussian KDE. The density is drawn on the y axis; the sample year is on the x axis. A higher altitude of the graph indicates years with more available data.

7 Data availability

The described $\delta^{13}C_{POC}$ data are available at https://doi.org/10.1594/PANGAEA.929931 (Verwega et al., 2021).

8 Conclusions

The aim of this work was to construct the largest publicly accessible $\delta^{13}C_{POC}$ data set. The starting point of our collection and analyses was the readily available data collection of Goericke (1994), which comprised 467 data points. Our primary objective was to elaborate this set of data by adding useful meta-information from the original publications and by introducing additional $\delta^{13}C_{POC}$ measurements, as recorded in the world ocean database PANGAEA and made available by Robyn Tuerena and Anne Lorrain. This way we could expand the data collection substantially, from the original 467 to 4732 data points. This new $\delta^{13}C_{POC}$ data set provides the best coverage to date and will be a useful tool to help constrain many marine carbon cycling processes and pathways from ocean–atmosphere exchange to marine ecosystems, as well as to better understand observations and validate models. To ensure dynamic growth of our data collection, the corresponding author will provide annual updates of the data set. Furthermore, he may be contacted by any interested researcher who would like to add their data to this collection.

The data are provided in a csv structure and interpolated onto two different global grids in NetCDF format. The csv file contains the $\delta^{13}C_{POC}$ values, their anomalies to their mean and all available meta-information. The interpolations are provided on a coarse $1.8° \times 3.6°$ grid of a $\delta^{13}C_{POC}$-simulating model and a finer $1° \times 1°$ grid by the World Ocean Atlas. We have provided a detailed description of our data

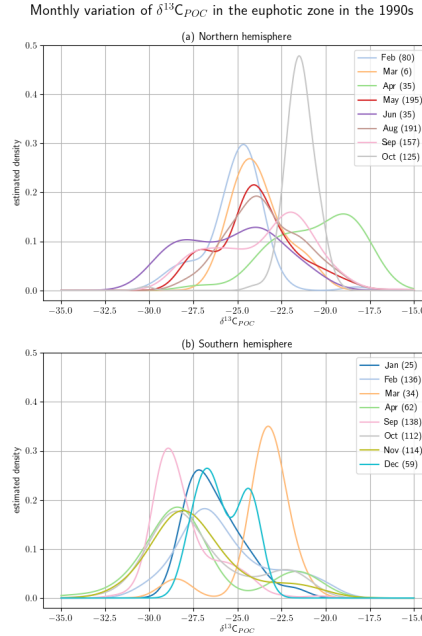

Figure 9. Monthly variations are split up by hemisphere with the Northern Hemisphere in **(a)** and Southern Hemisphere in **(b)**. Due to their having the best data coverage, the analyses are carried out within the 1990s and in the uppermost 130 m. The $\delta^{13}C_{POC}$ values are split up by sample month, and for every month with enough available data points (here more than three) a Gaussian KDE approximates their density. The number of used data points is given in each KDE label. For each hemisphere the densities are drawn all together; each month is indicated by an individual color.

collection procedure, all added meta-information and data coverage as well of the interpolation procedure carried out. We took the utmost care to make all data coherent, comparable and back-trackable and all adjustments transparent. Assumptions, changes and deletions of the used data sets have been described in detail.

We have described the general spatial and temporal trends of the sampled $\delta^{13}C_{POC}$ data of the raw data file. Distributions were always approximated by Gaussian kernel density estimators. The data range from 1964–2015 with by far the best coverage in the 1990s. Sample locations reach down to a depth of nearly 5000 m and best cover the uppermost 10 m, especially in the Atlantic and Indian Ocean. We were able to

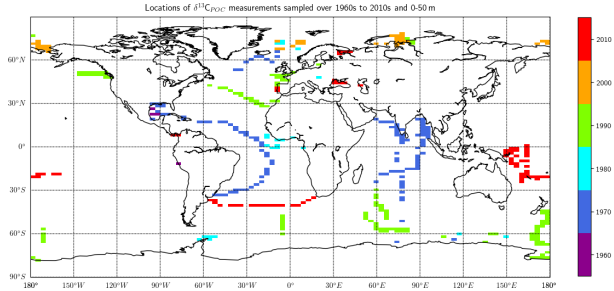

Figure 10. Grid locations of the $\delta^{13}C_{POC}$ data, colored by sampling decades. Only data of the uppermost layer are considered in this plot. The different colors indicate the different sample decades and were plotted increasing in time above each other.

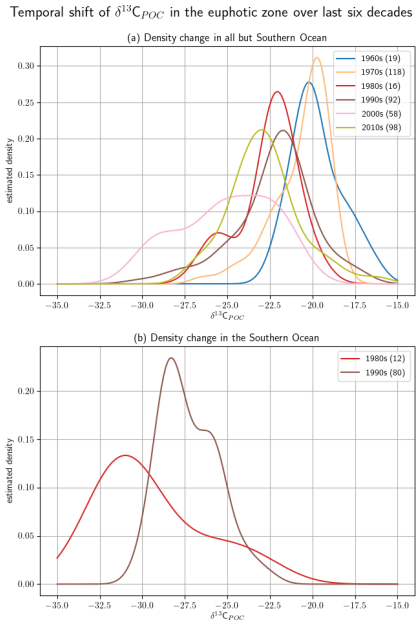

Figure 11. The decadal shift in $\delta^{13}C_{POC}$ values for all but the Southern Ocean **(a)** and only the Southern Ocean **(b)** shown by estimated densities of $\delta^{13}C_{POC}$ values. The differently colored graphs refer to the individual decades. Southern Ocean data are sparsely covered, and the region does not provide enough data for a reasonable comparison.

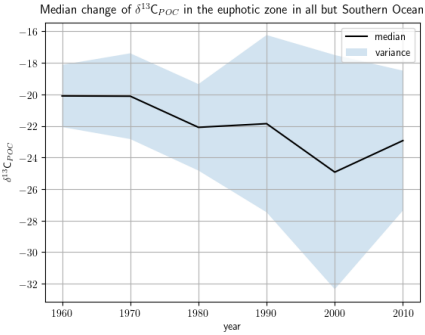

Figure 12. The decadal shift in $\delta^{13}C_{POC}$ values in the uppermost 130 m for all but the Southern Ocean: $\delta^{13}C_{POC}$ decadal median against the decades. The shaded area around the graph marks the variance of the respective decade in each direction.

show our $\delta^{13}C_{POC}$ data values are mostly located between $\delta^{13}C_{POC} = -15\,‰$ and $\delta^{13}C_{POC} = -35\,‰$ with two maxima at around $\delta^{13}C_{POC} = -27\,‰$ and $\delta^{13}C_{POC} = -23\,‰$, the latter one being more pronounced. A comparison of the main sample methods showed consistent results when compared with regions. $\delta^{13}C_{POC}$ data separated by months indicate counteracting seasonal trends in both hemispheres, but more data are required to demonstrate robust seasonality.

The interpolated data provide insights into geographical behavior of the sampled $\delta^{13}C_{POC}$ data. We showed a good general coverage of all global oceans by $\delta^{13}C_{POC}$ but observed a lack of data in PANGAEA that cover North Pacific regions. Since the Atlantic Ocean provides the best coverage, corresponding data were used for a north–south trend anal-

ysis, where we observed that the lowest values ($\lesssim -28\,‰$) can be found in the Southern Ocean, whereas the highest ($\gtrsim -22\,‰$) are restricted to low-latitude regions. This might also have influenced the observed lower $\delta^{13}C_{POC}$ values in the Southern Hemisphere compared to the Northern Hemisphere, due to the relatively good coverage of the Southern Ocean. Finally, we showed the sample locations and value development of $\delta^{13}C_{POC}$ over the observed decades. Since the Southern Ocean data were mainly sampled in the 1990s, a significant multi-decadal trend could not be detected there. In all other oceans our $\delta^{13}C_{POC}$ data show a decrease by about 3 ‰ over the observed time frame, which is about double the rate of the known Suess effect (Keeling, 1979) on aqueous $\delta^{13}CO_2$ (Young et al., 2013). This corroborates an increase in phytoplankton carbon fractionation that may be associated with a change in phytoplankton communities as previously suggested (Lorrain et al., 2020; Young et al., 2013). The data set shows promise for better understanding, constraining and prediction of carbon cycling as it provides a validation tool for mechanistic models and supports separation of non-spatial components in $\delta^{13}C_{POC}$ variations.

Appendix A: Statistical properties of $\delta^{13}C_{POC}$ kernel density estimates

In Tables A1, A2, A3, A4 and A5 we present the modes, medians and confidence limits of the KDEs derived in Figs. 1, 3, 7, 9 and 11, respectively.

Table A1. Statistical properties of the KDE derived for Fig. 1 evaluated on an equidistant grid over $[-55.15, -4.5]$ with 1001 grid points: the first column indicates the respective KDE, the following two its modes, the fourth the median and the fifth the 95 % confidence interval of the respective KDE. All values are given in per mill.

$\delta^{13}C_{POC}$ KDE	Dominant mode	Second mode	Median	95 % confidence interval
Figure 1	−23.6	−26.9	−23.8	[−30.9, −17.0]

Table A2. Statistical properties of the KDEs derived for Fig. 3 evaluated on an equidistant grid over $[-35, -15]$ with 1001 grid points: the first column indicates the respective KDE, the following two its modes, the fourth the median and the fifth the 95 % confidence interval of the respective KDE. All values are given in per mill.

$\delta^{13}C_{POC}$ KDE	Dominant mode	Second mode	Median	95 % confidence interval
Figure 3a, full	−21.8	−24.3	−24.3	[−26.8, −18.3]
Figure 3a, bottle	−25.1	–	−24.8	[−26.9, −22.0]
Figure 3a, intake	−21.6	–	−20.7	[−24.0, −17.4]
Figure 3a, net	−21.6	−27.4	−21.7	[−26.4, −19.5]
Figure 3a, trap	−24.3	−21.6	−24.1	[−27.2, −20.0]
Figure 3b, full	−21.6	–	−21.3	[−24.2, −18.0]
Figure 3b, intake	−19.5	−21.1	−20.3	[−24.8, −17.2]
Figure 3b, net	−21.5	–	−21.6	[−23.9, −19.4]
Figure 3c, full	−21.6	–	−21.6	[−26.4, −17.6]
Figure 3c, bottle	−24.7	–	−24.9	[−29.8, −21.1]
Figure 3c, net	−21.6	–	−21.6	[−24.0, −19.5]
Figure 3c, trap	−21.8	−18.8	−21.7	[−22.8, −18.5]
Figure 3d, full	−24.9	–	−24.6	[−27.1, −21.9]
Figure 3d, bottle	−25.2	–	−24.8	[−26.5, −22.1]
Figure 3d, net	−22.8	−26.9	−24.2	[−29.4, −19.7]
Figure 3d, trap	−24.3	−26.6	−24.5	[−27.2, −22.9]

Table A3. Statistical properties of the KDEs derived for Fig. 7 evaluated on an equidistant grid over $[-35, -15]$ with 1001 grid points: the first column indicates the respective KDE, the following two its modes, the fourth the median and the fifth the 95 % confidence interval of the respective KDE. All values are given in per mill.

$\delta^{13}C_{POC}$ KDE	Dominant mode	Second mode	Median	95 % confidence interval
Figure 7a, all	−21.8	–	−22.8	[−29.9, −18.1]
Figure 7a, biome 9	−24.0	–	−23.8	[−27.5, −18.5]
Figure 7a, biome 10	−21.7	–	−21.5	[−25.0, −17.9]
Figure 7a, biome 11	−21.6	−21.1	−21.3	[−24.4, −17.7]
Figure 7a, biome 12	−21.7	–	−21.9	[−23.2, −20.8]
Figure 7a, biome 13	−21.9	−24.9	−22.0	[−24.4, −20.4]
Figure 7a, biome 15	−22.7	–	−22.8	[−26.5, −19.2]
Figure 7a, biome 16	−28.7	−26.0	−27.7	[−30.7, −24.1]
Figure 7a, biome 17	−27.8	–	−28.5	[−32.7, −24.9]

Table A4. Statistical properties of the KDEs derived for Fig. 9 evaluated on an equidistant grid over [−35, −15] with 1001 grid points: the first column indicates the respective KDE, the following two its modes, the fourth the median and the fifth the 95 % confidence interval of the respective KDE. All values are given in per mill.

$\delta^{13}C_{POC}$ KDE	Dominant mode	Second mode	Median	95 % confidence interval
Figure 9a, Feb	−24.7	–	−25.0	[−29.2, −22.3]
Figure 9a, Mar	−24.3	–	−24.0	[−26.6, −20.3]
Figure 9a, Apr	−19.0	–	−20.4	[−26.2, −16.5]
Figure 9a, May	−24.1	−27.0	−24.0	[−28.3, −19.0]
Figure 9a, Jun	−24.0	−27.9	−25.2	[−30.5, −20.1]
Figure 9a, Aug	−23.9	–	−23.6	[−27.7, −18.9]
Figure 9a, Sep	−22.0	−26.0	−23.1	[−28.9, −18.7]
Figure 9a, Oct	−21.5	–	−21.5	[−23.4, −19.7]
Figure 9b, Jan	−27.2	–	−26.5	[−28.9, −22.1]
Figure 9b, Feb	−29.9	–	−26.2	[−30.3, −19.8]
Figure 9b, Mar	−23.3	−28.5	−23.3	[−29.0, −21.0]
Figure 9b, Apr	−28.4	−21.6	−28.1	[−32.6, −19.9]
Figure 9b, Sep	−28.9	−20.4	−28.5	[−30.8, −23.6]
Figure 9b, Oct	−28.5	−22.3	−27.7	[−31.7, −20.8]
Figure 9b, Nov	−28.1	–	−27.7	[−31.8, −20.1]
Figure 9b, Dec	−26.7	−24.4	−26.0	[−28.3, −23.3]

Table A5. Statistical properties of the KDEs derived for Fig. 11 evaluated on an equidistant grid over [−35, −15] with 1001 grid points: the first column indicates the respective KDE, the following two its modes, the fourth the median and the fifth the 95 % confidence interval of the respective KDE. All values are given in per mill.

$\delta^{13}C_{POC}$ KDE	Dominant mode	Second mode	Median	95 % confidence interval
Figure 11a, 1960s	−20.0	–	−19.9	[−26.8, −16.5]
Figure 11a, 1970s	−19.8	–	−20.4	[−25.0, −18.0]
Figure 11a, 1980s	−21.7	−25.3	−22.1	[−26.9, −18.5]
Figure 11a, 1990s	−21.8	−27.3	−22.1	[−27.6, −18.2]
Figure 11a, 2000s	−22.4	–	−23.2	[−30.4, −19.2]
Figure 11a, 2010s	−23.1	–	−23.3	[−27.4, −17.6]
Figure 11b, 1960s	−27.5	−30.3	−27.7	[−31.4, −25.2]
Figure 11b, 1980s	−31.0	–	−29.8	[−34.3, −15.0]

Author contributions. MTV collected and merged the data, performed the analyses, and drafted the manuscript. CJS initiated and supported the data collection, conducted the grid interpolations, guided analyses of the data, and structured and proofread the manuscript. MS supported the data collection, guided data analyses and proofread the manuscript. RET provided additional data and ideas for data analyses and proofread the manuscript. AL provided additional data and proofread the manuscript. AO guided the analysis of the data and proofread the manuscript. TS guided the elaboration of the manuscript, and structured and proofread it.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We would like to thank Tronje Kemena for providing the basic global biomes masks, used for analyzing the interpolated data sets on the coarse grid.

We thank the referees and editor for their constructive feedback regarding the initial version of the manuscript.

Financial support. This research has been supported by the Helmholtz School for Marine Data Science (MarDATA) (grant no. HIDSS-0005) and by the Deutsche Forschungsgemeinschaft (DFG) (project no. 445549720).

Review statement. This paper was edited by Attila Demény and reviewed by Anne Morée and one anonymous referee.

References

AESOPS: U.S. JGOFS Antarctic Environment and Southern Ocean Process Study, available at: http://usjgofs.whoi.edu/southern.html, last access: 3 December 2020.

Alfred-Wegener-Institut: PANGAEA Data Publisher for Earth & Environmental Science, available at: https://www.pangaea.de, last access: 3 December 2020.

Altabet, M. A. and Francois, R.: Natural nitrogen and carbon stable isotopic composition in surface water at cruise NBP96-05, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.128266, 2003a.

Altabet, M. A. and Francois, R.: Natural nitrogen and carbon stable isotopic composition of station NBP96-05-06-4, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.128229, 2003b.

Banse, K.: New views on the degradation and disposition of organic particles as collected by sediment traps in the open sea, Deep-Sea Res., 37, 1177–1195, https://doi.org/10.1016/0198-0149(90)90058-4, 1990.

Bidigare, R. R., Fluegge, A., Freeman, K. H., Hanson, K. L., Hayes, J. M., Hollander, D., Jasper, J. P., King, L. L., Laws, E. A., Milder, J., Millero, F. J., Pancost, R., Popp, B. N., Steinberg, P. A., and Wakeham, S. G.: Consistent fractionation of ^{13}C in nature and in the laboratory: Growth-rate effects in some haptophyte algae, Global Biogeochem. Cy., 11, 279–292, https://doi.org/10.1029/96gb03939, 1997.

Buchanan, P. J., Matear, R. J., Chase, Z., Phipps, S. J., and Bindoff, N. L.: Ocean carbon and nitrogen isotopes in CSIRO Mk3L-COAL version 1.0: a tool for palaeoceanographic research, Geosci. Model Dev., 12, 1491–1523, https://doi.org/10.5194/gmd-12-1491-2019, 2019.

Calvert, S. E.: Stable isotope data of sediment trap P84-4, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.68555, 2002.

Calvert, S. E. and Soon, M.: Carbon and nitrogen data measured on water samples from the multiple unit large volume filtration system (MULVFS) during John P. Tully cruise IOS_96-09, PAN-

GAEA [data set], https://doi.org/10.1594/PANGAEA.808319, 2013a.

Calvert, S. E. and Soon, M.: Carbon and nitrogen data measured on water samples from the multiple unit large volume filtration system (MULVFS) during John P. Tully cruise IOS_96-18, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.808320, 2013b.

Calvert, S. E. and Soon, M.: Carbon and nitrogen data measured on water samples from the multiple unit large volume filtration system (MULVFS) during John P. Tully cruise IOS_97-02, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.808321, 2013c.

Cassar, N., Laws, E. A., and Popp, B. N.: Carbon isotopic fractionation by the marine diatom Phaeodactylum tricornutum under nutrient- and light-limited growth conditions, Geochim. Cosmochim. Ac., 70, 5323–5335, https://doi.org/10.1016/j.gca.2006.08.024, 2006.

Chang, A. S., Bertram, M. A., Ivanochko, T. S., Calvert, S. E., Dallimore, A., and Thomson, R. E.: (Supplement 2) Total mass flux, geochemistry and abundance of selected diatom taxa of Effingham Inlet OSU Trap samples, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.806329, 2013.

Close, H. G. and Henderson, L. C.: Open-Ocean Minima in δ^{13}C Values of Particulate Organic Carbon in the Lower Euphotic Zone, Frontiers in Marine Science, 7, 540165, https://doi.org/10.3389/fmars.2020.540165, 2020.

Degens, E. T., Behrendt, M., Gotthardt, B., and Reppmann, E.: Metabolic fractionation of carbon isotopes in marine plankton – II. Data on samples collected off the coasts of Peru and Ecuador, Deep Sea Research and Oceanographic Abstracts, 15, 11–20, https://doi.org/10.1016/0011-7471(68)90025-9, 1968.

De Jonge, C., Stadnitskaia, A., Hopmans, E. C., Cherkashov, G. A., Fedotov, A., Streletskaya, I., Vasiliev, A. A., and Sinninghe Damsté, J. S.: (Table 2) Particulate organic carbon contentand the stable carbon isotope signal of suspended particulate matter samples, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.877962, 2015a.

De Jonge, C., Stadnitskaia, A., Hopmans, E. C., Cherkashov, G. A., Fedotov, A., Streletskaya, I., Vasiliev, A. A., and Sinninghe Damsté, J. S.: Drastic changes in the distribution of branched tetraether lipids in suspended matter and sediments from the Yenisei River and Kara Sea (Siberia): Implications for the use of brGDGT-based proxies in coastal marine sediments., Geochim. Cosmochim. Ac., 165, 200–225, https://doi.org/10.1016/j.gca.2015.05.044, 2015b.

Eadie, B. J. and Jeffrey, L. M.: δ^{13}C analyses of oceanic particulate matter, Mar. Chem., 1, 199–209, https://doi.org/10.1016/0304-4203(73)90004-2, 1973.

Eide, M., Olsen, A., Ninnemann, U. S., and Johannessen, T.: A global ocean climatology of preindustrial and modern ocean δ^{13}C, Global Biogeochem. Cy., 31, 515–534, https://doi.org/10.1002/2016gb005473, 2017.

EurOBIS Data Management Team: PANGAEA – data from Archive of Ocean Data, available at: http://ipt.vliz.be/eurobis/resource?r=pangaea_2724, last access: 3 December 2020.

Fay, A. R. and McKinley, G. A.: Global open-ocean biomes: mean and temporal variability, Earth Syst. Sci. Data, 6, 273–284, https://doi.org/10.5194/essd-6-273-2014, 2014.

Fischer, G.: Stabile Kohlenstoff-Isotopen in partikulärer organischer Substanz aus dem Südpolarmeer (Atlantischer Sektor), PhD thesis, Bremen University, Bremen, Germany, 1989.

Fontugne, M. and Duplessy, J. C.: Carbon isotope ration of marine plankton related to surface water masses, Earth Plamet. Sc. Lett., 41, 365–371, https://doi.org/10.1016/0012-821X(78)90191-7, 1978.

Fontugne, M. and Duplessy, J. C.: Oceanic carbon isotopic fractionation by marine plankton in the temperature range of −1 to 31 °C, Oceanol. Acta, 4, 85–90, 1981.

Fontugne, M., Descolas-Gros, C., and de Billy, G.: The dynamics of CO_2 fixation in the Southern Ocean as indicated by carboxylase activities and organic carbon isotopic ratios, Mar. Chem., 35, 371–380, https://doi.org/10.1016/S0304-4203(09)90029-9, 1991.

Francois, R., Atlabet, M. A., Goericke, R., McCorkle, D. C., Brunet, C., and Posson, A.: Changes in the δ^{13}C of surface water particulate organic matter across the subtropical convergence in the SW Indian Ocean, Global Biogeochem. Cy., 7, 627–644, https://doi.org/10.1029/93GB01277, 1993.

Freeman, K. H. and Hayes, J. M.: Fractionation of carbon isotopes by phytoplankton and estimates of ancient CO_2 levels, Global Biogeochem. Cy., 6, 185–198, https://doi.org/10.1029/92GB00190, 1992.

Fry, B.: ^{13}C/^{12}C fractionation by marine diatoms, Mar. Ecol. Prog. Ser., 134, 283–294, https://doi.org/10.3354/meps134283, 1996.

Fry, B. and Sherr, E. B.: δ^{13}C Measurements as Indicators of Carbon Flow in Marine and Freshwater Ecosystems, in: Stable Isotopes in Ecological Research. Ecological Studies (Analysis and Synthesis), edited by: Rundel, P. W., Ehleringer, J. R., and Nagy, K. A., Springer, New York, NY, USA, vol. 68, 196–229, https://doi.org/10.1007/978-1-4612-3498-2_12, 1989.

Garcia, H. E., Weathers, K., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Baranova, O. K., Seidov, D., and Reagan, J. R.: Dissolved Inorganic Nutrients (phosphate, nitrate and nitrate+nitrite, silicate), World Ocean Atlas 2018, 4, 35 pp., nOAA ATLAS NESDIS 84, NOAA National Centers for Environmental Information (NCEI), Silver Spring, Maryland, USA, 2018.

Goericke, R.: Variations of marine plankton δ^{13}C with latitude, temperature, and dissolved CO_2 in the world ocean, Global Biogeochem. Cy., 8, 85–90, https://doi.org/10.1029/93GB03272, 1994.

Gruber, N., Keeling, C. D., Bacastow, R. B., Guenther, P. R., Lueker, T. J., Wahlen, M., Meijer, H. A. J., Mook, W. G., and Stocker, T. F.: Spatiotemporal patterns of carbon-13 in the global surface oceans and the oceanic suess effect, Global Biogeochem. Cy., 13, 307–335, https://doi.org/10.1029/1999GB900019, 1999.

Hayes, J. M.: An Introduction to Isotopic Calculations, Woods Hole Oceanographic Institution, available at: http://www.whoi.edu/cms/files/jhayes/2005/9/IsoCalcs30Sept04_5183.pdf (last access: 12 May 2020), 2004.

Hofmann, M., Wolf-Gladrow, D. A., Takahashi, T., Sutherland, S. C., Six, K. D., and Maier-Reimer, E.: Stable carbon isotope distribution of particulate organic matter in the ocean: a model study, Mar. Chem., 72, 131–150, https://doi.org/10.1016/s0304-4203(00)00078-5, 2000.

IPCC: Summary for policymakers, Cambridge University Press, Cambridge, UK, 3–29, https://doi.org/10.1017/CBO9781107415324.004, 2013.

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Core Writing Team, Pachauri, R. K., and Meyer, L. A., Geneva, Switzerland, 2014.

Jahn, A., Lindsay, K., Giraud, X., Gruber, N., Otto-Bliesner, B. L., Liu, Z., and Brady, E. C.: Carbon isotopes in the ocean model of the Community Earth System Model (CESM1), Geosci. Model Dev., 8, 2419–2434, https://doi.org/10.5194/gmd-8-2419-2015, 2015.

Jasper, J. P. and Hayes, J. M.: A carbonisotopic record of CO_2 levels during the late Quaternary, Nature, 347, 462–464, https://doi.org/10.1038/347462a0, 1990.

JGOFS: Joint Global Ocean Flux Study, available at: http://ijgofs.whoi.edu, last access: 3 December 2020.

Kaiser, D., Konovalov, S. K., Arz, H. W., Voss, M., Krüger, S., Pollehne, F., Jeschek, J., and Waniek, J. J.: Black Sea water column dissolved nutrients and dissolved and particulate organic matter from winter 2013, Maria S. Merian cruise MSM33, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.898717, 2019.

Keeling, C. D.: The Suess effect: 13Carbon-14Carbon interrelations, Environ. Int., 2, 229–300, https://doi.org/10.1016/0160-4120(79)90005-9, 1979.

Kessler, W. S. and McCreary, J. P.: The annual wind-driven Rossby wave in the subthermocline equatorial Pacific, J. Phys. Oceanogr., 23, 1192–1207, 1992.

Laws, E. A., Popp, B. N., Bidigare, R. R., Kennicutt, M. C., and Macko, S. A.: Dependence of phytoplankton carbon isotopic composition on growth rate and [CO2]aq: Theoretical considerations and experimental results, Geochim. Cosmochim. Ac., 59, 1131–1138, https://doi.org/10.1016/0016-7037(95)00030-4, 1995.

Lein, A. Y. and Ivanov, M. V.: (Table 9.4.3) Concentrations of suspended matter in water samples from the 9°50′ N EPR hydrothermal field and contents and isotopic compositions of organic carbon in suspended matter, PANGAEA [data set], PANGAEA, https://doi.org/10.1594/PANGAEA.771566, 2009.

Lein, A. Y., Bogdanov, Y. A., Grichuk, D. V., Rusanov, I. I., and Sagalevich, A. M.: (Table 5) Concentration of particulate organic carbon and its isotopic composition in water samples from hydrothermal fields at the axis of the East Pacific Rise near 9°50′ N, PANGAEA [data set], PANGAEA, https://doi.org/10.1594/PANGAEA.745910, 2006.

Lein, A. Y., Bogdanova, O. Y., Bogdanov, Y. A., and Magazina, L. O.: (Table 6) Isotopic composition of organic carbon from microbial communities within the Lost City hydrothermal field, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.765164, 2007.

Levin, I., Schuchard, J., Kromer, B., and Münnich, K. O.: The Continental European Suess Effect, Radiocarbon, 31, 431–440, https://doi.org/10.1017/s0033822200012017, 1989.

Liu, B., Six, K. D., and Ilyina, T.: Incorporating the stable carbon isotope ^{13}C in the ocean biogeochemical component of the Max Planck Institute Earth System Model, Biogeosciences, 18, 4389–4429, https://doi.org/10.5194/bg-18-4389-2021, 2021.

Lorrain, A., Pethybridge, H., Cassar, N., Receveur, A., Allain, V., Bodin, N., Bopp, L., Choy, C. A., Duffy, L., Fry, B., Goni, N., Graham, B. S., Hobday, A. J., Logan, J. M., Ménard, F., Menkes, C. E., Olson, R. J., Pagendam, D. E., Point, D., Revill, A. T., Somes, C. J., and Young, J. W.: Trends in tuna carbon isotopes suggest global changes in pelagic phytoplankton communities, Glob. Change Biol., 26, 458–470, https://doi.org/10.1111/gcb.14858, 2020.

MacKenzie, K. M., Robertson, D. R., Adams, J. N., Altieri, A. H., and Turner, B. L.: Carbon and nitrogen stable isotope data from organisms in the Bay of Panama ecosystem, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.903842, 2019.

Magozzi, S., Yool, A., Zanden, H. B. V., Wunder, M. B., and Trueman, C. N.: Using ocean models to predict spatial and temporal variation in marine carbon isotopes, Ecosphere, 8, e01763, https://doi.org/10.1002/ecs2.1763, 2017.

McConnaughey, T. and McRoy, C. P.: Food-Web structure and the fractionation of Carbon isotopes in the bering sea, Mar. Biol., 53, 257–262, https://doi.org/10.1007/bf00952434, 1979.

Morée, A. L., Schwinger, J., and Heinze, C.: Southern Ocean controls of the vertical marine $\delta^{13}C$ gradient – a modelling study, Biogeosciences, 15, 7205–7223, https://doi.org/10.5194/bg-15-7205-2018, 2018.

Ndeye, M., Sène, M., Diop, D., and Saliège, J.-F.: Anthropogenic CO_2 in the Dakar (Senegal) Urban Area Deduced from ^{14}C Concentration in Tree Leaves, Radiocarbon, 59, 1009–1019, https://doi.org/10.1017/rdc.2017.48, 2017.

NOAA's Pacific Marine Environmental Laboratory: Ferret Support, available at: http://ferret.pmel.noaa.gov/Ferret, last access: 26 November 2020.

Popp, B. N., Takigiku, R., Hayes, J. M., Louda, J. W., and Baker, E. W.: The post-palaeozoic chronology and mechanism of ^{13}C depletion in primary marine organic matter, Am. J. Sci., 289, 436–454, 1989.

Popp, B. N., Laws, E. A., Bidigare, R. R., Dore, J. E., Hanson, K. L., and Wakeham, S. G.: Effect of Phytoplankton Cell Geometry on Carbon Isotopic Fractionation, Geochim. Cosmochim. Ac., 62, 69–77, https://doi.org/10.1016/s0016-7037(97)00333-5, 1998.

Rau, G. H., Takahashi, T., and Des Marais, D. J.: Latitudinal variations in plankton $\delta^{13}C$: implications for CO_2 and productivity in past oceans, Nature, 341, 516–518, https://doi.org/10.1038/341516a0, 1989.

Rau, G. H., Riebesell, U., and Wolf-Gladrow, D.: A model of photosynthetic ^{13}C fractionation by marine phytoplankton based on diffusive molecular CO_2 uptake, Mar. Ecol. Prog. Ser., 133, 275–285, 1996.

Rounick, J. S. and Winterbourn, M. J.: Stable carbon isotopes and carbon flow in ecosystems – Measuring ^{13}C to ^{12}C ratios can help to trace carbon pathways, BioScience, 36, 171–177, https://doi.org/10.2307/1310304, 1986.

Rubino, M., Etheridge, D. M., Trudinger, C. M., Allison, C. E., Battle, M. O., Langenfelds, R. L., Steele, L. P., Curran, M., Bender, M., White, J. W. C., Jenk, T. M., Blunier, T., and Francey, R. J.: A revised 1000 year atmospheric $\delta^{13}C$-CO_2 record from Law Dome and South Pole, Antarctica, J. Geophys. Res.-Atmos., 118, 8482–8499, https://doi.org/10.1002/jgrd.50668, 2013.

Sacket, W. M., Eckelmann, W. R., Bender, M. L., and Bé, A. W. H.: Temperature Dependence of Carbon Isotope Composi-

tion in Marine Plankton and Sediments, Science, 148, 235–237, https://doi.org/10.1126/science.148.3667.235, 1965.

Saupe, S. M., Schell, D. M., and Griffiths, W. B.: Carbon-isotope ratio gradients in western arctic zooplankton, Mar. Biol., 103, 427–432, https://doi.org/10.1007/BF00399574, 1989.

Schmittner, A. and Somes, C. J.: Complementary constraints from carbon (^{13}C) and nitrogen (^{15}N) isotopes on the glacial ocean's soft-tissue biological pump, Paleoceanography, 31, 669–693, https://doi.org/10.1002/2015PA002905, 2016.

Schmittner, A., Gruber, N., Mix, A. C., Key, R. M., Tagliabue, A., and Westberry, T. K.: Biology and air–sea gas exchange controls on the distribution of carbon isotope ratios (δ^{13}C) in the ocean, Biogeosciences, 10, 5793–5816, https://doi.org/10.5194/bg-10-5793-2013, 2013.

Silverman, B. W.: Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability, Chapman and Hall, London, UK, 1986.

Suess, E.: Particulate organic carbon flux in the oceans–surface productivity and oxygen utilization, Nature, 288, 260–263, 1980.

Tagliabue, A. and Bopp, L.: Towards understanding global variability in ocean carbon-13, Global Biogeochem. Cy., 22, GB1025, https://doi.org/10.1029/2007gb003037, 2008.

Thiede, J., Gerlach, S. A., Altenbach, A., and Henrich, R.: Sedimentation im europaeischen Nordmeer – Organisation und Forschungsprogramm des Sonderforschungsbereiches 313 fuer den Zeitraum 1988–1990, Tech. rep., Kiel University, Kiel, Germany, 1988.

Tjiputra, J. F., Schwinger, J., Bentsen, M., Morée, A. L., Gao, S., Bethke, I., Heinze, C., Goris, N., Gupta, A., He, Y.-C., Olivié, D., Seland, Ø., and Schulz, M.: Ocean biogeochemistry in the Norwegian Earth System Model version 2 (NorESM2), Geosci. Model Dev., 13, 2393–2431, https://doi.org/10.5194/gmd-13-2393-2020, 2020.

Trull, T. W. and Armand, L. K.: Insights into Southern Ocean carbon export from the δ^{13}C of particles and dissolved inorganic carbon during the SOIREE iron release experiment, Deep-Sea Res. Pt. II, 48, 2655–2680, https://doi.org/10.1016/S0967-0645(01)00013-3, 2001.

Trull, T. W. and Armand, L. K.: δ^{13}C content of particulate organic carbon measured on samples from traps during TANGAROA cruise SOIREE, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.807904, 2013a.

Trull, T. W. and Armand, L. K.: δ^{13}C content of fractionated particulate organic carbon measured on samples from traps during TANGAROA cruise SOIREE, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.807906, 2013b.

Tuerena, R. E., Ganeshram, R. S., Humphreys, M. P., Browning, T. J., Bouman, H., and Piotrowski, A. P.: Isotopic fractionation of carbon during uptake by phytoplankton across the South Atlantic subtropical convergence, Biogeosciences, 16, 3621–3635, https://doi.org/10.5194/bg-16-3621-2019, 2019.

Verwega, M.-T., Somes, C. J., Tuerena, R. E., and Lorrain, A.: A global marine particulate organic carbon-13 isotope data product, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.929931, 2021.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat. Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Volk, T. and Hoffert, M. I.: Ocean carbon pumps: analysis of relative strengths and efficiencies in ocean-driven atmospheric CO_2 changes, American Geophysical Union; Geophysical Monograph, 32, 99–110, 1985.

Voss, M. and von Bodungen, B.: Carbon and nitrogen from mooring NB2, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.106805, 2003.

Wada, E., Terazaki, M., Kabaya, Y., and Nemoto, T.: ^{15}N and ^{13}C abundances in the Antarctic Ocean with emphasis on the biogeochemical structure of the food web, Deep-Sea Res., 34, 829–841, https://doi.org/10.1016/0198-0149(87)90039-2, 1987.

Westerhausen, L. and Sarnthein, M.: δ^{13}C of plankton from surface water (Table A2), PANGAEA [data set], https://doi.org/10.1594/PANGAEA.89388, 2003.

Young, J. N., Bruggeman, J., Rickaby, R. E. M., Erez, J., and Conte, M.: Evidence for changes in carbon isotopic fractionation by phytoplankton between 1960 and 2010, Global Biogeochem. Cy., 27, 505–515, https://doi.org/10.1002/gbc.20045, 2013.

Zeebe, R. E. and Wolf-Gladrow, D.: CO_2 in Seawater: Equilibrium, Kinetics, Isotopes, Elsevier Science B.V., Elsevier Oceanography Series, Amsterdam, the Netherlands, 65, 2001.

Zhang, J., Quay, P., and Wilbur, D.: Carbon isotope fractionation during gas-water exchange and dissolution of CO2, Geochim. Cosmochim. Ac., 59, 107–114, https://doi.org/10.1016/0016-7037(95)91550-d, 1995.

Chapter 3

The second version of the global marine particulate carbon-13 isotope data set

The originally published $\delta^{13}C_{POC}$ data set [VST+21] has been extended and re-published in a second version [PST+22]. The number of incorporated data points have increased from 4732 data points to 6952 data points in the second version. The new data are mainly Southern Ocean samples [GEH+21; CH20]. Again, the second version is available on the global WOA grid and as a spreadsheet file including all data points and their meta information on the data platform PANGAEA. In the following, I give an impression into the second database version, how the data has changed from the first to the second version and what new insights generate from the updated data.

Fig. 3.1 presents the global distribution of the second data set version. The figure is an adapted version of Fig. 6 in [VSS+21] showing the values of the second data base version in their respective grid cells. The used grid is the WOA grid [GWP+18], which is the only grid the second data version is available on. It is visible that mainly Southern Ocean data has been added, but also some data points in the prior sparsely sampled northern Pacific Ocean.

The added Southern Ocean values are in agreement with the previous observations and similarly widespread. The new data show mainly values around $\delta^{13}C_{POC} = -30$ ‰. Most of them are south of New Zealand, Cape Horn and Tasmania, all reaching down to the Antarctic coast. Furthermore, cruises have been added reaching from below South Africa to the Antarctic coast at around 0° and 60° E, the latter one extending along the coastline

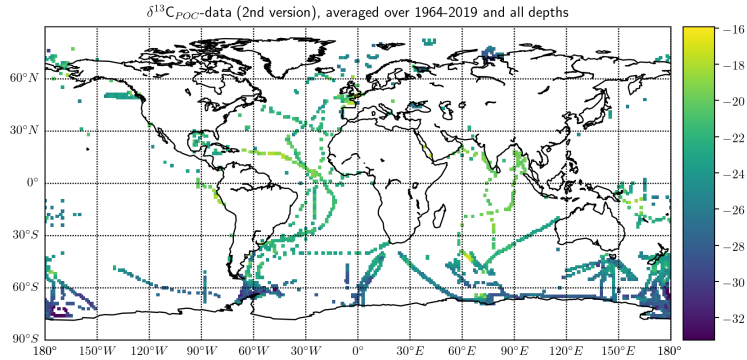

Figure 3.1. All available $\delta^{13}C_{POC}$ data in the second data base version: The colored regions mark grid cells with available $\delta^{13}C_{POC}$ data. The colorscale indicates the values of the $\delta^{13}C_{POC}$ measurements at the respective locations.

up to around 150° E. The lowest values are now measured between 150° E and 150° W and around 60° W reaching down to less than $\delta^{13}C_{POC} = -32$ ‰. The highest values were measured close to the Antarctic coast at around 75° E reaching up to around $\delta^{13}C_{POC} = -22$ ‰.

The northern Pacific Ocean is still sparsely covered. Nevertheless, individual data points have been added between 120° E and 150° W, mostly north from 30° N. West from 180° E there were no $\delta^{13}C_{POC}$ data available in the first version. All added data show values around $\delta^{13}C_{POC} = -23$ ‰.

In Fig. 3.2 I provide KDEs of all $\delta^{13}C_{POC}$ data from the first and the second data set version in comparison. The KDEs are diffKDEs [PS23] from the first and second data base versions and the originally published Gaussian KDE of the first data base version (see Fig. 2) in [VSS+21]). The

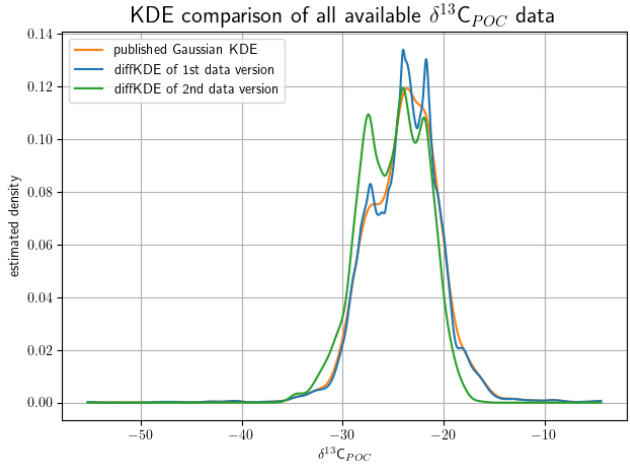

Figure 3.2. All available $\delta^{13}C_{POC}$ data of the first and second data base version: Both data set versions are visualized with the diffusion KDE, the first data set version also with the originally published Gaussian KDE.

comparison of the Gaussian KDE to the new diffKDE on the first data base version reveals how the Gaussian KDE oversmoothed at least two important data features with (local) minima at around $\delta^{13}C_{POC} = -22$ ‰and $\delta^{13}C_{POC} = -27$ ‰. These are well resolved by the diffKDE. The comparison of the diffKDEs of the first and second data set version present the influence of the amount of added Southern Ocean data. These are generally far lower $\delta^{13}C_{POC}$ values, reaching down to $\delta^{13}C_{POC} = -32$ ‰, which is well visible by a far stronger pronunciation of the smallest mode at $\delta^{13}C_{POC} = -28$ ‰.

In Fig. 3.3, I show the long term decadal trend of $\delta^{13}C_{POC}$ values in the second data base version. This is an update of Fig. 12 in [VSS+21]. In [VST+21] the data were taken from decadal averages, which are not available from the second data base version. For the second version, the data means are calculated from the spreadsheet file. Analogously to the first version, selected data is restricted to the euphotic zone and excluding the Southern Ocean. A comparison with the plot from [VSS+21] shows a

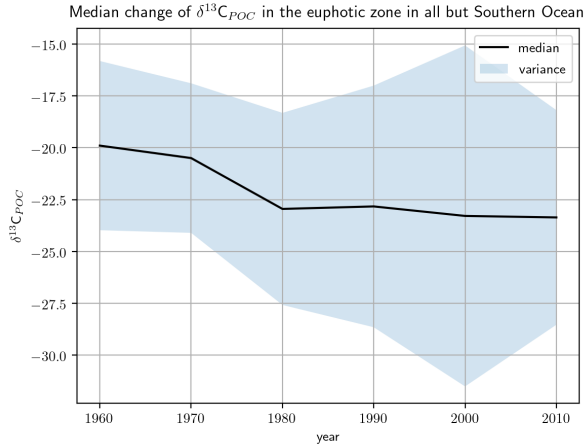

Figure 3.3. Decadal changes of $\delta^{13}C_{POC}$ drawn from the second data base version: Decadal averages of $\delta^{13}C_{POC}$ data are drawn in the black line against their respective decades. The grey shaded area in the back visualizes the variance. All $\delta^{13}C_{POC}$ data is restricted to the euphotic zone and excluding Southern Ocean data.

comparable long term trend from around $\delta^{13}C_{POC} = -20$ ‰in the 1960 to around $\delta^{13}C_{POC} = -23$ ‰in the 2010s. Only this time the trend is more a monotically decreasing function exhibiting the prior dip at the 2000s and already starting to decrease in the 1970s.

Chapter 4

A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research

First author paper submitted to Geoscientific Model Development and published in the discussion forum on the 13th February 2023.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research

Maria-Theresia Pelz[1,2], Markus Schartau[2], Christopher J. Somes[2], Vanessa Lampe[2], and Thomas Slawig[1]
[1]Department of Computer Science, Kiel University, 24118 Kiel, Germany
[2]GEOMAR Helmholtz Centre for Ocean Research Kiel, 24105 Kiel, Germany
Correspondence: mpelz@geomar.de

Abstract. Probability density functions (PDFs) comprise basic information about the variability of observed or simulated variables within a system of interest. In geoscience data distributions are often expressed by a parametric estimation of their PDF, such as e.g. a Gaussian distribution. At present there is a growing attention towards the analysis of non-parametric estimation of PDFs, where no prior assumptions about the type of PDF are required. A common tool for such non-parametric estimation is a kernel density estimator (KDE). Existing KDEs are valuable but incomplete, because of the difficulty of specifying optimal bandwidths for the individual kernels. A diffusion-based KDE provides a useful approach to mitigate the difficulty in identifying bandwidths that resolve desired details of multi-modal data while being insensitive to noise. Therefore we designed and developed a new implementation of a diffusion-based KDE as an open source Python tool. We tested our implementation on artificial and real marine biogeochemical data individually and against other popular KDEs. Our estimator is able to detect relevant multiple modes and resolve boundary close data while suppressing details induced by noise and individual outliers. The convergence rate is comparable to the Gaussian estimator, but with a generally smaller error, most notably for small data sets with up to around 5000 data points. We exemplify and discuss the general applicability of such KDEs for data-model comparison in geoscience, in particular for sparse data. We also provide an example for how our approach can be efficiently utilized for the derivation of plankton size spectra in ecological research.

1 Introduction

In geoscience the application of Earth system models (ESMs) has become an integral part of climate research (IPCC, 2022). Given the complexity of ESMs and the associated manifold of model solutions, there is strong demand for assessing the agreement of simulation results with observational data. In fact, such necessity is not only restricted to simulations with ESM but is transferable to other model applications as well, like in social science, in financial- or ecological research. A viable evaluation procedure is to compare probability density functions (PDFs) of the data with their simulated counterparts, which may also be quantified by some distance measure or divergence between respective PDFs (Thorarinsdottir et al., 2013). Along with the examination of the suitability of specific divergence functions for data-model assessment as done by Thorarinsdottir et al. (2013), a necessary prerequisite is the approximation of PDFs based on available data and model results.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions

Mathematically formulated, PDFs are integrable non-negative functions $f : \mathcal{A} \to [0, \infty]$ from a probability space $(\Omega, \mathcal{A}, P)$ into the non-negative real numbers. By definition, they allow to directly read the probability of the occurrence of a data value $X \in \mathbb{R}$ within a specific range $[a, b] \subseteq \mathbb{R}$ via the relationship

$$P(a < X < b) = \int_a^b f(x)\,dx \text{ for all } a < b \in \mathbb{R}. \tag{1}$$

The application of kernel density estimators (KDEs) has become a common approach for approximating PDFs in a *non-parametric* way (Parzen, 1962), which means that no probability parameters (like expectation or variance) of the data and no type of the underlying probability distribution (as, e.g., normal or log-normal) are prescribed or assumed. The general concept of KDEs takes into account information of every single data point and treats all of them equally. Consequently, every point's information weighs the same in the resulting estimate, without introducing additional assumptions.

A KDE is based on a kernel function and a smoothing parameter. The kernel function is ideally chosen to be a PDF itself. It is usually unimodal and centered around zero (Sheather, 2004). In the estimation process, the kernel function is sequentially centered around each data point. The sum of these individual kernels is standardized by the number of data points. This ensures that the final estimate is again a PDF by inheriting all properties of its kernels. The smoothing parameter, referred to as bandwidth, determines the smoothness of the estimate. If it is chosen to be small, more details of the underlying data structure become visible. If it is larger, more structure becomes smoothed out (Jones et al., 1996), and information from single data points might get lost. Hence, it is crucial to determine some kind of an optimal size of the bandwidth parameter to represent a suitable signal-to-noise ratio that allows a separation of significant distinctive features from ambiguous details. The question of optimal bandwidth selection is widely discussed in the literature (e.g., Sheather and Jones, 1991; Jones et al., 1996; Heidenreich et al., 2013). It also takes into account that there might not be one single "optimal" choice for such bandwidth (Abramson, 1982; Terrell and Scott, 1992; Chaudhuri and Marron, 2000; Scott, 2012).

The reformulation of the most common Gaussian KDEs (Sheather, 2004) into a diffusion equation provides a different view on KDE (Chaudhuri and Marron, 2000). This perspective change is possible, because the Gaussian kernel function solves the partial differential equation describing the diffusion heat process as the Green function. The time parameter of this differential equation corresponds to the smoothness of the estimate, and thus becomes tantamount to the estimate's bandwidth parameter (Chaudhuri and Marron, 2000). The initial value is typically set to include the δ-distribution of the input data. This differentiates the initial value problem from classical problems, since the δ-distribution is not a proper function itself. In specific applications this diffusion approach delivered convincing results (e.g., Botev et al., 2010; Deniz et al., 2011; Qin and Xiao, 2018). However, it tends to resolve too many details or overfit the data in others (e.g., Ma et al., 2019; Chaudhuri and Marron, 2000; Farmer and Jacobs, 2022). One main benefit of the diffusion KDE is that it provides a series of PDF estimates for a sequence of bandwidths by default (Chaudhuri and Marron, 2000). As a consequence, it offers the chance to choose between different grades of smoothness by design.

In this study, we present a new, modified diffusion-based KDE, for which we provide a Python implementation. Our aim is to retain the original idea of diffusion-based KDEs by Chaudhuri and Marron (2000) and Botev et al. (2010), but to avoid the

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

complex fixed-point iteration by Botev et al. (2010). The main objective of our refined approach is to achieve high performance for analyses of high variance and multimodal data sets. Our diffusion-based KDE is based on an iterative approximation that differs from others, using a default optimal bandwidth and two preliminary, so-called *pilot* estimates. This way the KDE can provide a family of estimates at different bandwidths to choose from in addition to a default solution optimally designed for data from geoscience and ecological research. Thus, an interactive investigation option of these different estimates becomes possible in an easy way.

This paper is structured as follows: At first, we will briefly recall the general concept of KDEs. Afterwards, our specific KDE approach will be introduced and described, as developed and implemented in a software package. We explain the two pilot estimation steps and the selection of the smoothing parameters. Then the performance of our refined estimator will be compared with other state-of-the-art KDEs, while considering known distributions and real marine biogeochemical data. The real test data include carbon isotope ratios of particulate organic matter ($\delta^{13}C_{POC}$) and plankton size data. Our analyses presented here involve investigations of KDE error, runtime, the sensitivity to data noise and the characteristics of convergence w.r.t. increasing sample size.

2 Theory and methods

2.1 Kernel density estimation

A kernel density estimator (KDE) is a non-parametric statistical tool for the estimation of probability density functions (PDFs). In practice, diverse specifications of KDEs exist that may improve the performance with respect to individual needs. Before we explain our specifications of the diffusion-based KDE, we will provide basic background information about KDEs.

For all following let $\Omega \subseteq \mathbb{R}$ be a domain, $X_j \in \Omega$, $j \in \{1, ..., N\}$, be $N \in \mathbb{N}$ independent real random variables.

2.1.1 The general kernel density estimator

The most general form of a KDE approximates the true density f of the input data $(X_j)_{j=1}^N$ by

$$\hat{f} : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}, \quad (x;h) \mapsto \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right). \tag{2}$$

In this formula the *kernel function* $K : \mathbb{R} \to \mathbb{R}_{>0}$ has to satisfy the following conditions (Parzen, 1962):

$$\sup_{y \in \mathbb{R}} |K(y)| < \infty, \quad \int_{\mathbb{R}} |K(y)|\, dy < \infty, \quad \lim_{y \to \infty} |yK(y)| = 0, \quad \int_{\mathbb{R}} K(y) dy = 1. \tag{3}$$

The parameter h determines the smoothness of the estimate calculated by Eq. 2 and is called the *bandwidth parameter* (Silverman, 1986). In the following we will exclusively deal with the squared bandwidth (h^2) and therefore adapt a notation where some t is defined as $h^2 =: t \in \mathbb{R}$. An optimal choice for the bandwidth parameter is regarded as the minimizer of the asymptotic mean squared error between the true density of $(X_j)_{j=1}^N$ and their KDE (Sheather and Jones, 1991). The mean

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

integrated squared error is defined as

$$\mathbf{MISE}\left(\hat{f}\right) : \mathbb{R}_{>0} \to \mathbb{R}, t \mapsto \mathbf{E}\left(\int_{\mathbb{R}} \left(\hat{f}\left(x;t\right) - f\left(x\right)\right)^2 dx\right) \tag{4}$$

for all PDFs f and respective KDEs $\hat{f}$ (Scott, 1992). If now $\hat{f}$ is a KDE and there exists a $t^* \in \mathbb{R}_{>0}$ with

$$\mathbf{AMISE}\left(\hat{f}\right)\left(t^*\right) = \min_{t \in \mathbb{R}_{>0}} \mathbf{AMISE}\left(\hat{f}\right)\left(t\right), \tag{5}$$

we call t^* the *optimal bandwidth* of $\hat{f}$ by $\left(X_j\right)_{j=1}^N$ (Scott, 1992). For the general KDE from Equation 2, this can be calculated according to Parzen (1962) as

$$t^* = \left(\frac{f\left(x\right)\int K^2\left(y\right)dy}{N4\left\|f^2\right\|^2}\right)^{\frac{2}{5}}. \tag{6}$$

As we see from Eq. 6, the true density f is involved in the calculation of the optimal bandwidth t^*, which is in turn needed for the approximation of f by a KDE. Thus, a direct derivation of an optimal bandwidth is precluded. One possibility of how this implicit relation can be solved is the calculation of pilot estimation steps. Our specific approach to this is shown in Sec. 2.1.3 and Sec. 2.1.4.

There exists a variety of available choices for the type of kernel function K, which all have their individual benefits and shortcomings. Amongst them are for example the uniform, triangle, or the Epanechnikov kernel (Scott, 1992):

$$K_E : \mathbb{R} \to \mathbb{R}_{\geq 0}, w \mapsto \frac{3}{4}\left(1 - w^2\right). \tag{7}$$

A common choice for K is the Gaussian kernel (Sheather, 2004):

$$\Phi : \mathbb{R} \to \mathbb{R}_{\geq 0}, w \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2}. \tag{8}$$

The standard KDE from Eq. 2 – despite being widely applied and investigated – comes with several disadvantages in practical applications (Khorramdel et al., 2018). For example, severe boundary bias can occur when applied on a compact interval (Marron and Ruppert, 1994). It means that a kernel function with a specified bandwidth, attributed to a single point nearby the boundary, may actually exceed the boundary. Furthermore, it can lack a proper response to variations in the magnitude of the true density f (Breiman et al., 1977). The introduction of a parameter that depends on the respective data region can address the latter (Breiman et al., 1977). Unfortunately, no true independent local bandwidth strategy exists (Terrell and Scott, 1992), meaning that in all local approaches there is still an influence of neighboring data points on each locally chosen bandwidth.

2.1.2 The diffusion-based kernel density estimator

The diffusion-based KDE provides a different approach to Eq. 2. It solves the partial differential equation describing the diffusion heat process, starting from an initial value based on the input data $\left(X_j\right)_{j=1}^N$, that progresses up to an estimate at a final time $T \in \mathbb{R}_{>0}$. An advantageous connection to Eq. 2 is that the widely applied Gaussian kernel is a fundamental solution of this

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

differential equation. Precisely, the Gaussian kernel from Eq. 8 as applied in the construction of a Gaussian KDE depends on the location $x \in \mathbb{R}$ and the smoothing parameter $h \in \mathbb{R}_{>0}$ and has the form $(x;t) \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X_j}{\sqrt{t}}\right)^2}$ for any $j \in \{1,...,N\}$. This function solves

$$\frac{\partial}{\partial t} u(x;t) = \frac{1}{2} \frac{d^2}{dx^2} u(x;t), x \in \Omega, t \in \mathbb{R}_{>0} \tag{9}$$

as the Green's function, where the time parameter $t \in \mathbb{R}_{>0}$ equals the squared bandwidth parameter h^2 (Chaudhuri and Marron, 2000). This idea to use the diffusion heat equation to calculate a KDE was first proposed by Chaudhuri and Marron (2000). Its benefits were widely explored in Botev et al. (2010).

Our implementation of the diffusion KDE is based on Chaudhuri and Marron (2000) and is extended by some advancements proposed by Botev et al. (2010): We included a parameter function $p \in C^2(\Omega, \mathbb{R}_{>0})$ with $\|p''\|_\infty < \infty$ into Eq. 9, acting inversely to a diffusion quotient. Boundary conditions are set to be Neumann and the initial value being a normalized sum of the δ-distributions centered around the input data points. In the following, we call a function $u \in C^{2,1}(\Omega \times \mathbb{R}_{>0}, \mathbb{R}_{\geq 0})$ the *diffusion kernel density estimator* (diffKDE), if it solves the diffusion partial differential equation

$$\frac{\partial}{\partial t} u(x;t) = \frac{1}{2} \frac{d^2}{dx^2} \left(\frac{u(x;t)}{p(x)} \right), \qquad x \in \Omega, t \in \mathbb{R}_{>0}, \tag{10}$$

$$\frac{\partial}{\partial x} \left(\frac{u(x;t)}{p(x)} \right) = 0, \qquad x \in \partial\Omega, t \in \mathbb{R}_{>0}, \tag{11}$$

$$u(x;0) = \frac{1}{N} \sum_{j=1}^{N} \delta(x - X_j), \qquad x \in \Omega. \tag{12}$$

In Eq. 12, the data are incorporated as initial values via the Dirac δ-distribution, i.e., a generalized function which takes the value infinity at its argument and zero anywhere else. Regarded as PDF, it puts all probability in the corresponding data point. The δ-distribution can be defined exactly as a limit of functions, the so-called Dirac sequence. In actual implementations, it has to be approximated, see Sec. 2.2.3.

The final iteration time $T \in \mathbb{R}_{>0}$ of the solution process of Eq. 10 is called the squared *bandwidth* of the diffKDE.

This specific type of KDE has several advantages. First of all, it naturally provides a sequence of estimates for different smoothing parameters (Chaudhuri and Marron, 2000). This obliterates identifying one single optimal bandwidth whose existence is questioned (e.g., Abramson, 1982; Terrell and Scott, 1992; Chaudhuri and Marron, 2000; Scott, 2012). Even more, such a sequence allows a specification of the estimate's smoothness that is most appropriate for the analysis. The parameter function p introduces adaptive smoothing properties (Botev et al., 2010). Thus, setting p properly solves the prior problem of having to locally adjust the bandwidth to the respective region to prevent oversmoothing of local data structure (Breiman et al., 1977; Terrell and Scott, 1992; Pedretti and Fernàndez-Garcia, 2013). In contrast to local bandwidth adjustments, local variations of the smoothing intensity can be applied to resolve multimodal data as well as values close to the boundary.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

2.1.3 Bandwidth selection

According to the relationship $T = h^2$ between final iteration time T of the diffKDE and bandwidth parameter h (Chaudhuri and Marron, 2000), we from now on focus on the selection of the optimal squared bandwidth $T \in \mathbb{R}_{>0}$ and refer to this as the *bandwidth selection* for simplicity.

In Eq. 6 we stressed that the optimal choice of the bandwidth parameter depends on the true density f. In our setup of the diffKDE, the analytical solution for T of Eq. 5 depends not only on the true density f, but also on the parameter function p. It can be calculated as

$$T^* = \left(\frac{\mathbf{E}\left(\sqrt{p(X)}\right)}{2N\sqrt{\pi}\left\| \left(\frac{f}{p}\right)'' \right\|_{L^2}^2} \right)^{\frac{2}{5}} \tag{13}$$

(Botev et al., 2010). The role of the parameter function p is in detail described in Sec. 2.1.4.

In the simplified setup with $p = 1$ as in Eq. 9, the analytical optimal solution of Eq. 6 becomes

$$T^*_{(p=1)} = \left(\frac{1}{2\,N\sqrt{\pi}\,\|f''\|_{L^2}^2} \right)^{\frac{2}{5}} . \tag{14}$$

Still, the smoothing parameter depends on the unknown density function f and its derivatives. So we will need to find a suitable approximation of f, which might again be dependent of f and so on. Botev et al. (2010) use an iterative scheme to solve this implicit dependency. This additional effort is avoided in our approach.

The prior claimed possibility of no existence of one single optimal bandwidth for complicated densities (e.g. Scott, 2012) is, by default, no problem for the diffKDE. A solution to this problem is to create a family of estimates from different bandwidth parameters (Breiman et al., 1977), ranging from oversmoothed estimates to those with beginning oscillations (Sheather, 2004). For the diffKDE the progression of the time t up to a final iteration time T is equivalent to the creation of such a family of estimates. For the diffKDE we thus only need to find a suitable optimal final iteration time T^*. Then, the temporal solution of Eq. 10 provides solutions for the diffKDE for the whole sequence of the temporal discretization time steps smaller than T^*, which we can then use as the requested family of estimates.

2.1.4 Pilot estimation

A crude first estimate of the true density f can serve as a pilot estimation step for several purposes (Abramson, 1982; Sheather, 2004). The most obvious in our case is to obtain an estimate of f for the calculations of the optimal bandwidth in Eq. 13. The second purpose is its usage for the definition of the parameter function p in Eq. 10. Setting this as an estimate of the true density itself introduces locally adaptive smoothing properties (Botev et al., 2010). Since p appears in the denominator in the diffusion equation, it operates conversely to a classical diffusion coefficient. Choosing p to be a function allows for a spatially dependent influence on the smoothing intensity: at points where the function p is small, the smoothing becomes more pronounced, whereas if p is larger, the smoothing is less intense. This resolves the expected structure in data dense areas,

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

but expands in sparsely sampled areas. Eventually, we calculate two pilot estimates – one for p and one for f – to support the calculation of the diffKDE. We set both pilot estimates to be the solution of Eq. 9 with an optimal smoothing parameter approximating Eq. 14. This approach combines Gaussian and diffKDE interchangeably to make best use of both of their benefits (Chung et al., 2018).

2.2 Discretization of diffusion kernel density estimator

Equation 10 is solved numerically, using a spatial and temporal discretization. The discretization is based on finite differences and sparse matrices in Python. A similar approach can be found in a diffusion-based kernel density estimator for linear networks implemented in R by McSwiggan et al. (2016).

2.2.1 Spatial discretization

We start with the description of the discretization of the spatial domain $\Omega \subseteq \mathbb{R}$. This will reduce the partial differential equation in Eq. 10 into a system of linear ordinary differential equations.

Let $n \in \mathbb{N}$ and $(x_i)_{i=0}^n \subseteq \bar{\Omega}$, an equidistant discretization of Ω with $x_{i-1} < x_i$ and $\mathbb{R}_{>0} \ni h := x_i - x_{i-1}$ for all $i \in \{1,\dots,n\}$. For the following calculations, we set $x_{-1} := x_0 - h \in \mathbb{R}$ and $x_{n+1} := x_n + h \in \mathbb{R}$. Let u be the solution of the diffKDE and p its parameter function, both as defined in Sec. 2.1.2. We assume that u and p are both defined on x_{-1} and x_{n+1} and we set $u_i = u(x_i)$ and $p_i = p(x_i)$ for all $i \in \{-1,\dots,n+1\}$.

Let $t \in \mathbb{R}_{>0}$. We approximate Eq. 11 at $x = x_0$ by applying a first order central difference quotient as

$$0 = \frac{\partial}{\partial x}\left(\frac{u(x_0;t)}{p(x_0)}\right) = \frac{1}{2h}\left(\frac{u_1(t)}{p_1} - \frac{u_{-1}(t)}{p_{-1}}\right).$$

This implies

$$\frac{u_{-1}(t)}{p_{-1}} = \frac{u_1(t)}{p_1}.$$

We approximate Eq. 10 at $x = x_0$ by applying a second order central difference quotient

$$u_0'(t) = \frac{1}{2}\frac{1}{h^2}\left(\frac{u_1(t)}{p_1} - 2\frac{u_0(t)}{p_0} + \frac{u_{-1}(t)}{p_{-1}}\right) = \frac{1}{2}\frac{1}{h^2}\left(2\frac{u_1(t)}{p_1} - 2\frac{u_0(t)}{p_0}\right). \tag{15}$$

Analogously, we approximate Eq. 11 and Eq. 10 at $x = x_n$ again by first and second order central difference quotients, respectively. This gives

$$u_n'(t) = \frac{1}{2}\frac{1}{h^2}\left(\frac{u_{n+1}(t)}{p_{n+1}} - 2\frac{u_n(t)}{p_n} + \frac{u_{n-1}(t)}{p_{n-1}}\right) = \frac{1}{2}\frac{1}{h^2}\left(2\frac{u_{n-1}(t)}{p_{n-1}} - 2\frac{u_n(t)}{p_n}\right). \tag{16}$$

Finally, we derive from Eq. 10 by applying a second order central difference quotient for all $i \in \{1,\dots,n-1\}$:

$$u_i'(t) = \frac{1}{2}\frac{1}{h^2}\left(\frac{u_{i+1}(t)}{p_{i+1}} - 2\frac{u_i(t)}{p_i} + \frac{u_{i-1}(t)}{p_{i-1}}\right). \tag{17}$$

Now, we identify $\boldsymbol{p} := (p_0,\dots,p_n) \in \mathbb{R}^{n+1}$, $\boldsymbol{u}'(t) := (u_0'(t),\dots,u_n'(t)) \in \mathbb{R}^{n+1}$ and $\boldsymbol{u}(t) := (u_0(t),\dots,u_n(t)) \in \mathbb{R}^{n+1}$ with their spatial discretizations. Furthermore, we define $\boldsymbol{v}_{\text{upper}} := (2,1,\dots,1) \in \mathbb{R}^n$, $\boldsymbol{v}_{\text{main}} := (-2,\dots,-2) \in \mathbb{R}^{n+1}$ and $\boldsymbol{v}_{\text{lower}} :=$

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023


Geoscientific Model Development Discussions

$(1, ..., 1, 2) \in \mathbb{R}^n$ to be the upper, main and lower diagonal of the tridiagonal matrix $\mathbf{V} \in \mathbb{R}^{(n+1)\times(n+1)}$. Now, we set

$$\frac{1}{2}\frac{1}{h^2}\mathbf{V}\frac{1}{\boldsymbol{p}} =: \mathbf{A} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{18}$$

where the division by $\boldsymbol{p}$ is meant to be column-wise. Then, Eq. 15, Eq. 16 and Eq. 17 can be summarized as a linear system of ordinary differential equations:

$$\boldsymbol{u}'(t) = \frac{1}{2}\frac{1}{h^2}\mathbf{V}\frac{\boldsymbol{u}(t)}{\boldsymbol{p}} = \mathbf{A}u(t). \tag{19}$$

By these calculations the partial differential equation from Eq. 10 becomes a system of ordinary differential equations:

$$\boldsymbol{u}'(t) = \mathbf{A}\boldsymbol{u}(t), \quad t \in \mathbb{R}_{>0}. \tag{20}$$

2.2.2 Temporal discretization

The time-stepping applied to solve the ordinary differential equation from Eq. 20 and Eq. 12 is again built on equidistant steps forward in time. Let $\Delta \in \mathbb{R}_{>0}$ small and set $t_0 := 0$ and $t_k := t_{k-1} + \Delta$ for all $k \in \mathbb{N}$. Set $u_{k,i} := u(x_i, t_k) \in \mathbb{R}$ for all $i \in \{0, \ldots, n+1\}$ and $k \in \mathbb{N}_0$ and identify $\boldsymbol{u_k} := (u_{k,i})_{i=0}^{n} \in \mathbb{R}^{n+1}$ for all $k \in \mathbb{N}_0$ with their discretizations.

We use an implicit Euler method to approximate Eq. 20 for all $k \in \mathbb{N}_0$

$$\boldsymbol{u_{k+1}} = \Delta\mathbf{A}\boldsymbol{u_{k+1}} + \boldsymbol{u_k} \tag{21}$$

from which we obtain

$$\boldsymbol{u_k} = (\mathbf{I}_{n+1} - \Delta\mathbf{A})\,\boldsymbol{u_{k+1}} \text{ for all } k \in \mathbb{N}_0. \tag{22}$$

Together with the initial value Eq. 12 this describes an implementation-ready time stepping procedure. The linear equation for $\boldsymbol{u_{k+1}}$ will be solved in every time step $k \in \mathbb{N}_0$.

2.2.3 Initial value

The initial value in Eq. 12 depends on the δ-distribution (Dirac, 1927). The δ-distribution is not a proper function, but can be calculated as a limit of a suitable function sequence. A common approximation for the δ-distribution is to use a Dirac sequence (Hirsch and Lacombe, 1999). Such is a sequence $(\Phi_n)_{n\in\mathbb{N}}$ of integrable functions that are non-negative and satisfy

$$\int \Phi_n(x)\,dx = 1 \text{ for all } n \in \mathbb{N} \tag{23}$$

and

$$\lim_{n\to\infty} \int_{\mathbb{R}\setminus B_\rho(0)} \Phi_n dx = 0 \text{ for all } \rho \in \mathbb{R}_{>0}. \tag{24}$$

For our implementation we define a Dirac sequence $(\Phi_h)_{h\in\mathbb{R}_{>0}}$ depending on the spatial discretization fineness $h \in \mathbb{R}_{>0}$ as an approximation of δ in Eq. 12. The relationship $\frac{|\Omega|}{n} = h$ provides the dependency of Φ_h on $n \in \mathbb{N}$ and the equivalence of the

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

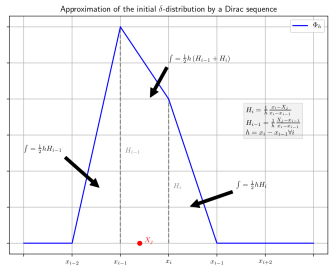

Figure 1. Dirac sequence $(\Phi_h)_{h\in\mathbb{R}_{>0}}$ for the approximation of the δ-distribution in the initial value in Eq. 12. The function Φ_h is depending on the spatial discretization fineness h and converges to δ for $h \to 0$. The function Φ_h is piecewise linear with a peak at each data point $X_j, j \in \{1, ..., N\}$ integrating to 1.

limits $n \to \infty$ and $h \to 0$ in this framework. In the following we give the specific definition of our function sequence of choice and proof that this indeed defines a proper Dirac sequence.

We assume $0 \in \Omega$. Then there exists an $i \in \{0, \dots, n\}$ with $0 \in [x_{i-1}, x_i)$. If not readily defined, we set $x_{i-2} := x_{i-1} - h \in \mathbb{R}$ and $x_{i+1} := x_i + h \in \mathbb{R}$. We define (see also Fig. 1)

$$\Phi_h : \Omega \to \mathbb{R}, x \mapsto \begin{cases} \frac{x_i}{h^3}x + \frac{x_i|x_{i-2}|}{h^3}, & x \in [x_{i-2}, x_{i-1}) \\ \frac{x_i+x_{i-1}}{h^3}x + x_i\frac{x_i+x_{i-1}}{h^3} - \frac{x_{i-1}}{h^2}, & x \in [x_{i-1}, x_i) \\ \frac{x_{i-1}}{h^3}x + \frac{x_{i+1}|x_{i-1}|}{h^3}, & x \in [x_i, x_{i+1}] \\ 0 & \text{else.} \end{cases} \tag{25}$$

Then $\Phi_h \in L^1(\mathbb{R})$ is non-negative for all $h \in \mathbb{R}_{>0}$ and $\int \Phi_h(x)\,dx = 1$ (see Appendix).

Now, let $\rho \in \mathbb{R}_{>0}$ and set $h = \frac{\rho}{2} \in \mathbb{R}_{>0}$. Then we have by Eq. 25

$$\int\limits_{\mathbb{R}\setminus\mathcal{B}_\rho(0)} \Phi_h dx = \int\limits_{-\infty}^{\rho} \Phi_h dx + \int\limits_{\rho}^{\infty} \Phi_h dx = \int\limits_{-\infty}^{\rho} 0 dx + \int\limits_{\rho}^{\infty} 0 dx = 0,$$

and it follows

$$\lim_{h\to 0} \int\limits_{\mathbb{R}\setminus\mathcal{B}_\rho(0)} \Phi_h dx = 0 \text{ for all } \rho \in \mathbb{R}_{>0}. \tag{26}$$

Hence Eq. 25 defines a Dirac sequence. We use Φ_h for the approximation of the δ-distribution in our implementation of Equation 12.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

The concept of the Dirac sequence also provides the justification to generally rely on the δ-distribution in the construction of the initial value of the diffKDE. The Gaussian kernel defined in Eq. 8 that solves the diffusion equation as a fundamental solution is again a Dirac-sequence (Boccara, 1990). This link connects the diffKDE directly back to the δ-distribution.

2.3 Implementation of the diffusion kernel density estimator

The selected implementation is a straight forward approach using equidistant finite differences in space and time and a direct solution of the diffusion equation by an implicit Euler. It is build on the three times sequent solution of the diffusion equation, providing two pilot estimates for the calculation of the final diffKDE. The three chosen bandwidths increase in complexity and accuracy over this iteration. The implementation is realized in Python 3.

2.3.1 Selection of pilot function and optimal bandwidth

For the optimal bandwidth T^* from Eq. 13 we need the parameter function p as well as the true density f. We approximate them both by a simple KDE, each as pilot estimation steps. We use for both cases the simplified diffKDE defined in Eq. 9, without additional parameter functions. We denote the bandwidths for p and f as $T_p, T_f \in \mathbb{R}_{>0}$, respectively. We use a simple bandwidth as variants of the *rule of thumb* by Silverman (1986) for both of them.

We begin to estimate T_p, which is the bandwidth for the KDE that serves as p. It shall be the smoothest of the three estimates, since p limits the resolution fineness of the diffKDE as a lower boundary. This is, because the diffKDE converges to this parameter function and hence never resolves less details than p itself (Botev et al., 2010).

As seen in Eq. 14, the optimal bandwidth for the approximation of p is depending on the second derivative of f. We therefore need to make some initial assumption about f. For a first simplification, we assume that f belongs to the normal distribution family. Then the variance can be estimated by the standard deviation of the data. This leads us to the parametric approximation of the bandwidth T_P (Silverman, 1986)

$$T_p = \left(\frac{1}{2N\sqrt{\pi}\,\|f''\|_{L^2}^2}\right)^{\frac{2}{5}} = \left(\frac{1}{2N\sqrt{\pi}\sigma^{-5}\,\|\Phi''\|_{L^2}^2}\right)^{\frac{2}{5}} = \left(\frac{1}{2N\sqrt{\pi}\sigma^{-5}\frac{3}{8}\frac{1}{\sqrt{\pi}}}\right)^{\frac{2}{5}} = \sigma^2\left(\frac{4}{3}N\right)^{-\frac{2}{5}}, \tag{27}$$

whose estimate is known to be overly smooth on multimodal distributions.

To calculate the bandwidth T_f for the approximation of f in Eq. 13 we choose a refined approximation of Eq. 14, which has been proposed by Silverman (1986) as

$$T_f = \left(0.9min\left(\sigma, \frac{iqr\,(data)}{1.34}\right)\right)^2 N^{-\frac{2}{5}}. \tag{28}$$

We approximate optimal bandwidth T^* from Eq. 13 by calculating p and f by Eq. 9, based on Eq. 27, and Eq. 28 respectively. The nominator is approximated by the unbiased estimator (Botev et al., 2010)

$$\mathbf{E}\,(p(X)) = \frac{1}{n+1}\sum_{i=0}^{n+1}\sqrt{p(x_i)} =: E_\sigma \tag{29}$$

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

and the second derivative in the denominator by finite differences (McSwiggan et al., 2016)

$$\left(\frac{f}{p}\right)''(x_i) = \frac{1}{h^2}\left(\frac{f}{p}(x_{i+1}) - 2\frac{f}{p}(x_i) + \frac{f}{p}(x_{i-1})\right) =: q_i \tag{30}$$

for all $i \in \{1,...,n\}$. For the boundary values we set

$$\left(\frac{f}{p}\right)''(x_0) = \frac{1}{h^2}\left(2\frac{f}{p}(x_1) - 2\frac{f}{p}(x_0)\right) =: q_0 \tag{31}$$

and

$$\left(\frac{f}{p}\right)''(x_{n+1}) = \frac{1}{h^2}\left(2\frac{f}{p}(x_{n-1}) - 2\frac{f}{p}(x_n)\right) =: q_{n+1}. \tag{32}$$

We set the finite differences approximation from Eq. 30, Eq. 31 and Eq. 32 as a discrete function with image $\boldsymbol{q} := (q_0, ..., q_{n+1})$. In this way we derived an already discrete formula for approximation of the optimal squared bandwidth $T^* \in \mathbb{R}_{>0}$ of the diffKDE on the discretization Ω as

$$T^* = \left(\frac{E_\sigma}{2N\sqrt{\pi}\|\boldsymbol{q}\|_{L^2}^2}\right)^{\frac{2}{5}}. \tag{33}$$

The L^2-norm is calculated on the discretized versions of f and p by array operations. The integration is performed by the *trapz* function of the SciPy integrate package (Gommers et al., 2022), the square root is part of the math package (Van Rossum, 2020).

2.3.2 The diffKDE algorithm with optimized bandwidth

The implementation is realized in Python and its concept shown in Alg. 1. We use the Python libraries Numpy (Harris et al., 2020) and SciPy (Virtanen et al., 2020; Gommers et al., 2022) and the Python Math module (Van Rossum, 2020) for data preprocessing, calculation of the bandwidths, setup of the differential equations and their solution. The algorithm iteratively calculates three KDEs: first the two for the approximations of p and f as the pilot estimation steps described in Sec. 2.3.1 and the last one being u the solution of the diffKDE built on the two prior. All three KDEs are calculated by solving the diffusion equation up to the respective final iteration time. The solution is realized in *while*-loops solving Eq. 22. The two pilot estimation steps can be calculated simultaneoulsy, since they are independent of each other and only differ in their final iteration times T_p and T_f. All input variables are displayed in Tab. 1 the return values listed in Tab. 2.

The spatial grid Ω is setup according to the description in Sec. 2.2.1 in lines 1 and 2 of Alg. 1. It consists of $n \in \mathbb{N}$ intervals, where n can be set by the user. The boundary values are $x_{min} := \min X \in \mathbb{R}$ and $x_{max} := \max X \in \mathbb{R}$ by default, but can also be chosen individually. Setting the boundary values to an individually chosen interval in the function call results in a clipping of the used data to this smaller one before KDE calculation. Outside the interval boundaries, the diffKDE adds two additional discretization points to make it applicable for the case of a data point X_i, $i \in \{0, ..., n+1\}$ being directly located at one of the boundaries. This way it is possible to construct the initial value defined in Eq. 25, which takes into account the two neighbouring discretization points in each direction. This leads to a full set of $n+1$ equidistant discretization points saved

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

Algorithm 1 Finite differences based algorithm for the implementation of the diffusion KDE.

Note: the routine solve$(\mathbf{M}, \mathbf{b})$ means that the system $\mathbf{M}\mathbf{x} = \mathbf{b}$ is solved.

Require: $\boldsymbol{X} \in \mathbb{R}^N, n \in \mathbb{N}, timesteps \in \mathbb{N}, x_{min} \in \mathbb{R}, x_{max} \in \mathbb{R}$

1: $h \leftarrow (x_{max} - x_{min}) / (n-4)$
2: $\boldsymbol{\Omega} \leftarrow (x_{min} - 2h, x_{min} - h, ..., x_{max} + h, x_{max} + 2h) \in \mathbb{R}^{n+1}$
3: $\boldsymbol{p}, \boldsymbol{f}, \boldsymbol{u} \leftarrow \Phi_h$
4: $T_p \leftarrow \sigma^2 \left(\frac{4}{3}N\right)^{-\frac{2}{5}}$
5: $T_f \leftarrow \left(0.9min\left(\sigma, \frac{iqr(data)}{1.34}\right)\right)^2 N^{-\frac{2}{5}}$
6: $t \leftarrow 0, \Delta_p \leftarrow T_p/timesteps, \Delta_f \leftarrow T_f/timesteps$
7: **while** $t < T_p$ **do**
8: $\quad \boldsymbol{p} \leftarrow$ solve $(\mathbf{I}_{n+1} - \Delta_p \mathbf{A}_{pilot}, \boldsymbol{p})$
9: $\quad \boldsymbol{f} \leftarrow$ solve $(\mathbf{I}_{n+1} - \Delta_f \mathbf{A}_{pilot}, \boldsymbol{f})$
10: $\quad t \leftarrow t + \Delta_p$
11: **end while**
12: $\boldsymbol{q} \leftarrow \sqrt{\int_\Omega \left(\left(\frac{\boldsymbol{f}}{\boldsymbol{p}}\right)''\right)^2 dh}$
13: $E_\sigma \leftarrow \frac{1}{n+1} \sum_{i=0}^{n+1} \sqrt{p(x_i)}$
14: $T \leftarrow \left(\frac{E_\sigma}{2N\sqrt{\pi}q^2}\right)^{\frac{2}{5}}$
15: $t \leftarrow 0, \Delta \leftarrow T/timesteps$
16: **while** $t < T$ **do**
17: $\quad \boldsymbol{u} \leftarrow$ solve $(\mathbf{I}_{n+1} - \Delta \mathbf{A}, \boldsymbol{u})$
18: $\quad t \leftarrow t + \Delta$
19: **end while**
20: **return** $\boldsymbol{u}, \boldsymbol{\Omega}, \boldsymbol{\Phi_h}, \boldsymbol{p}, \boldsymbol{stages}, \boldsymbol{times}$

in the variable Ω. The spatial discretization Ω includes an inner discretization between the handed in (or default set) interval endpoints x_{min} and x_{max} of $n-4$ equally sized inner discretization intervals.

The Dirac sequence Φ_h for the implementation of the initial value is defined in Eq. 25 and we use the same for initialization of all three approximations of the the PDF (p, f, u) in line 3 of Alg. 1. In its calculation, the algorithm searches for each $j \in \{1, ..., N\}$ for the $i \in \{1, ..., n+1\}$ with x_i being the closest right neighbour of X_j. Then the initial value is constructed by assigning the values $\frac{1}{h}\frac{x_i - X_j}{x_i - x_{i-1}}$ and $\frac{1}{h}\frac{X_j - x_{i-1}}{x_i - x_{i-1}}$ at grid point x_i and x_{i-1}, respectively, and zero elsewhere. These values are corresponding to the weighed heights H_i and H_{i-1} displayed in Fig. 1. The final initial value is the normalized sum of all these individual approximations of the δ-distribution. All three used KDEs (p, f, u) are initialized with this initial value.

In the pilot estimation steps, we calculate the KDEs for p and for f required for the set up of the bandwidth T^* for the diffKDE. The bandwidths T_p and T_f for p and f, respectively, are calculated based on the input data X in lines 4 and 5 of Alg. 1 as described in Sec. 2.3.1. Then, the KDEs are calculated by solving a linear ordinary differential equation by an implicit Euler in the first *while*-loop in lines 7 to 9 of Alg. 1. For the pilot estimation steps calculating p and f the matrix $\mathbf{A}$ defined in

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

Table 1. Input variables: The only required input variable for the calculation of the diffKDE is a one dimensional data set as an array like type. All other variables are optional, with some prescribed defaults. On demand the user can set individual lower and upper bounds for their data evaluation under the diffKDE as well as the number of used spatial and temporal discretization intervals. The individual selection of the final iteration time provides the opportunity to choose the specific smoothing grade on demand.

index	name	type	default	description
0	$data$	array like	required	input data $X \in \mathbb{R}$
1	$xmin$	float	$\min X$	lower data boundary for KDE calculation
2	$xmax$	float	$\max X$	upper boundary for KDE calculation
3	n	integer	1004	number of spatial discretization intervals in Ω
4	$timesteps$	integer	20	number of temporal discretization intervals
5	T	float	T^*	final iteration time for diffKDE

Eq. 18 does not incorporate a parameter function and reduces to

$$\frac{1}{2}\frac{1}{h^2}\mathbf{V} =: \mathbf{A}_{pilot} \in \mathbb{R}^{(n+1)\times(n+1)}. \tag{34}$$

Apart from this, the solutions for the pilot KDEs are the same as for the final diffKDE. The two pilot KDEs can be solved simultaneously, since they share their matrix $\mathbf{A}_{pilot}$ and have independent pre-computed bandwidths. The difference in their bandwidths is implemented in different time step sizes Δ_p and Δ_f for p and f, respectively, which are initialized in line 6 of Alg. 1 directly before this first *while*-loop. The time forward is calculated $timesteps \in \mathbb{N}$ times in equidistant time steps until each individual final iteration time derived by the respective bandwidths. Since we solve implicitly, there is no restriction to the time step size. But a larger $timesteps$ parameter reduces the numerical error proportional to the step size parameters Δ_p and Δ_f. In this temporal solution we rely on the fact that the involved matrices are sparsely covered. The applied solver is part of the SciPy Python library and designed for efficient solution of linear systems including sparse matrices (Virtanen et al., 2020; Gommers et al., 2022).

The final bandwidth T for the diffKDE solution u is calculated after the calculations of p and f, using them both as described in Sec. 2.3.1 in lines 12 to 14 of Alg. 1. For the diffKDE u the differential equation is given in Eq. 20 and the solution approach by the implicit Euler in Eq. 22. This is implemented in a second *while*-loop described in lines 16 to 18 in Alg. 1 and apart from the final iteration time T^* and the matrix $\mathbf{A}$ identical to the calculations in the pilot step.

The return value is a vector providing the user the diffKDE, along with some main parameters and the opportunity to also evaluate different approximation stages. It provides in the first and second entry the diffKDE and the spatial discretization Ω. The third entry is the initial value Φ_δ and the fourth pilot estimate p that influences the adaptive smoothing. The last return values two vectors are handed back: $stages$ and $times$. These include the approximation stages of the diffKDE and the respective times exceeding the default optimal solution stored in the diffKDE and providing also some oversmoothed solution

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

Table 2. Return values of the diffKDE: The return variable of the diffKDE is a vector. Its first entry is the diffKDE evaluated on the spatial grid. Its second entry is the spatial grid Ω.

index	name	type	size	description
0	u	Numpy array	$n+1$	diffKDE values on Ω
1	Ω	Numpy array	$n+1$	spatial discretization

for individual evaluations. The times are the 20 timesteps used for the calculation of u and 10 additional with doubled stepsize reaching up to the doubled approximated optimal final iteration time $2T^*$.

Possible problems are caught in *assert* and *if* clauses. First of all, the data is reshaped to a Numpy array for the case of a list handed in and it is made sure that this is non-empty. For the case of numerical issues leading to a pilot estimate including zero values, the whole pilot is set back equal to 1 to ensure numerical convergence. Similar is done for the case of NaN value being delivered for the optimal bandwidth for the diffKDE, in which case this is also set to the bandwidth chosen for f in Eq. 28.

2.4 Pre-implemented visual outputs

Besides the standard use to calculate a diffKDE at an approximated optimal final iteration time for direct usage, we also included three possibilities to generate a direct visual output, one of them being interactive. Matplotlib (Hunter, 2007) provides the software measures for creating the plots. Most methods are part of the submodule Pyplot, the interactive plot is based on the submodule Slider.

The function call *evol_plot* opens a plot showing the time evolution of the diffKDE. The plot includes drawings of the data points on the x-axis. In the background the initial values are drawn, but cut off at 20 % above the global maximum of the diffKDE to preserve focus of the graphic on the diffKDE and evolution. The evolution is presented by drawings of the individual time evolution stages using the sequential color map Viridis. In the front the diffKDE is drawn. This visualization of the evolution provides the user insight into the data distribution and their respective influence on the final form of the diffKDE.

The function call *pilot_plot* opens that shows the diffKDE together with its pilot estimate p, showing the intensity of local smoothing. With this the user has the possibility to gain insight to the influence of this pilot estimator on the performance of the diffKDE. This plot also includes the data points on the x-axis.

The function call *custom_plot* opens an interactive plot, allowing the user to slide through different approximation stages of the diffKDE. This feature is based on the Slider module from the Matplotlib library (Hunter, 2007) and opens a plot showing the diffKDE. On the bottom of this plot is a scale that shows the time, initially being set to the optimal iteration time derived from Eq. 13 in the middle of the scale. By clicking to the scale, the user can display the evolution stages at the respective (closest) iteration time. This reaches down to the initial value and up to the doubled optimal iteration time. This interactive tool provides the user a simple tool to follow the estimate at different bandwidths, the intensity of smoothing at different localizations. With

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions

the help of such plot it is possible to decide on whether the diffKDE is desired to be applied with a final iteration time that is different from the default.

3 Results and Discussion

In the following we document the performance of the diffKDE on artificial and real marine biogeochemical data. Different data sources are chosen to best show possibilities and performance of the diffKDE. Additionally, snapshots of the pre-implemented plot routines are given as examples. Whenever not stated otherwise, we used the default values of the input variables stated in Tab. 2 in the calculation of the diffKDE.

For testing our implementation against a known true PDF we first constructed a three-modal distribution. The objective is to assess the diffKDE's resolution and to exemplify the pre-implemented plot routines. The distribution was constructed from three Gaussian kernels centered around $\mu_1 = 3$, $\mu_2 = 6.5$ and $\mu_3 = 9$ and with variances $\sigma_1^2 = 1$, $\sigma_2^2 = 0.7^2$ and $\sigma_3^2 = 0.5^2$, each of them with a relative contribution of 30 %, 60 % and 10 %, respectively:

$$f(x) = 0.3\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-3)^2} + 0.6\frac{1}{0.7\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-6.5}{0.7}\right)^2} + 0.1\frac{1}{0.5\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-9}{0.5}\right)^2}. \tag{35}$$

The performance of the diffKDE is then illustrated with real data of a) measurements of carbon isotopes (Verwega et al., 2021; Verwega et al., 2021) and b) of plankton size (equivalent spherical diameter) (Lampe et al., 2021). We chose these data because we propose to apply the diffKDE for the analysis of field data for assessment and optimization of marine biogeochemical- as well as size-based ecosystem models. The carbon isotope data have been collected to constrain model parameter values of a marine biogeochemical model that incorporates this tracer as a prognostic variable (Schmittner and Somes, 2016).

3.1 Pre-implemented outputs

As described in Sec. 2.4, we included three plot functions in the diffKDE implementation. All of them open pre-implemented plots, to give an impression of the special features that come with the diffKDE. An overview of the three possible direct visual outputs of the diffKDE software is described below.

First we outline the possibility to display the diffKDE's evolution. By calling the *evol_plot* function, a plot opens that shows all temporal evolution stages of the solution of Eq. 22. The temporal progress is visualized by a sequential colorscheme progressing from light yellow over different shades of green to dark blue. On the x-axis, all used data points are drawn and in the background a cut-off part of the initial value in light yellow as the beginning of the temporal evolution. The final diffKDE is plotted as a bold blue line in front of the evolution process. This gives the user an insight in the distribution of the initial data and their influence on the shape of the estimate. As an example of the default setting, we created an evolution plot from 100 random samples of Eq. 35 visualized in Fig. 2. The second example shows the possibility of displaying the diffKDE together with the pilot estimate p by the function *pilot_plot*. This is the parameter function in Eq. 12 responsible for the adaptive smoothing. Where this function is larger, the smoothing is less intense and allows more structure in the estimate of the diffKDE. Contrarily where it is smaller, the smoothing becomes more pronounced and data gaps are better smoothed out. The result of the diffKDE

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions

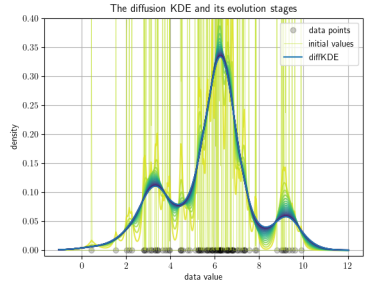

Figure 2. Pre-implemented direct visual output of the evolution process of the diffKDE. The input data are 100 samples randomly collected from Eq. 35. The individual data points are drawn on the x-axis. The y-axis represents the estimated probability density. The light yellow vertical lines in the background are the initial value of the the diffKDE. The temporal evolution of the solution of Eq. 22 is visualized by the sequent color scheme from light yellow over green to the bold blue graph in the front. The final diffKDE at the approximated optimal final iteration time represents as this graph the end of the time evolution.

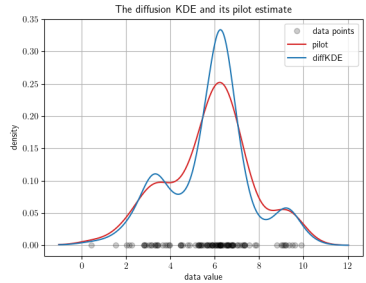

Figure 3. The diffKDE and its pilot estimate p. The input data are 100 samples randomly collected from Eq. 35. The data points are drawn on the x-axis. The y-axis represents the estimated density of the diffKDE in blue and the pilot estimate in red.

380 is shown together with its parameter function p in figure Fig. 3 on the same random sample of the distribution from Eq. 35 as before.

Lastly, we illustrate example snapshots of the interactive option to investigate different smoothing stages of the diffKDE by the function. We chose simpler and smaller example data for this demonstration, because these are better suited for visualization

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions

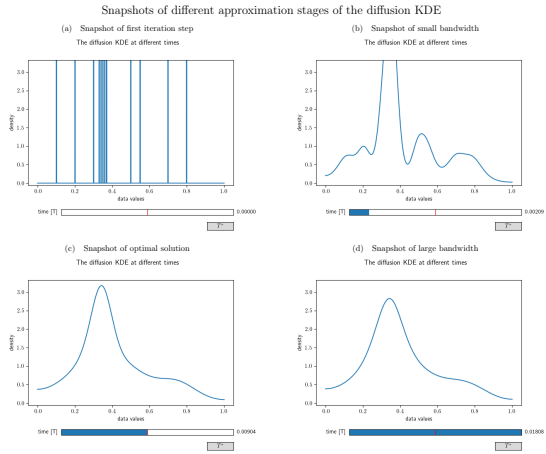

Figure 4. Different snapshots from the interactive visualization of the diffusion KDE generated from the artificial data set $(0.1, 0.2, 0.3, 0.33, 0.34, 0.35, 0.36, 0.37, 0.5, 0.55, 0.7, 0.8)$. (a) shows the output at $time = 0$ and hence the initial value. (b) shows an intermediate smoothing stage of the diffKDE. (c) shows the diffKDE of the input data at the approximated optimal iteration time T^*. This is the initial stage of the interactive graphic. By clicking the button on the lower right, the graphic can be reset to this stage. (d) shows an oversmoothed version of the diffKDE at the doubled approximated optimal iteration time.

of this tool's possibilities. The function *custom_plot* opens an interactive graphic, starting with a plot of the approximated optimal default solution of the diffKDE at T^*. In this graphic the user is able to individually choose, by a slider, the iteration time at which the desired approximation stage of the diffKDE can be seen. The time can be chosen from 0, where the initial value is shown, up until the doubled approximated optimal time ($2 \times T^*$). A reset button sets the graphic back to its initial stage of the diffKDE at T^*. Four snapshots of this interactive experience are drawn in Fig. 4.

3.2 Performance analyses on known distributions and in comparison to other KDEs

In this section we present results obtained by random samples of the trimodal distribution from Eq. 35 and lognormal distributions with differing parameters. Wherever suitable, the results are compared to other commonly used KDEs. These include the most common Gaussian KDE with the kernel function from Eq. 8 (Gommers et al., 2022), the Epanechnikov KDE with the kernel function from Eq. 7 (Pedregosa et al., 2012) and an improved implementation of the Gaussian KDE by Botev et al. (2010) in a Python implementation by Hennig (2021). We begin with an example of how the user may choose individually

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development
Discussions
Open Access
EGU

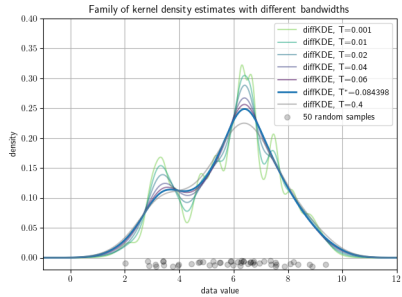

Figure 5. Family of diffKDEs evaluated at different bandwidths: A data set of 50 random samples drawn as grey circles on the x-axis serve to show the possibility to investigate a whole family of estimates by the diffDKE. The bold blue line represents the default solution of the diffKDE by solving the diffusion equation up to the approximated optimal final iteration time T^*. The other colors depict more detailed prior approximation stages with smaller bandwidth, i.e. earlier iteration times, and a smoother estimate with a far larger iteration time.

different smoothing grades of the diffKDE, then compare the different KDEs with the true distribution, followed by investigating the influence of noise on different KDEs, and finally show the convergence of different KDEs to the true distribution with increasing sample size.

We start with an individual selection of the approximation stages. This is one of the main benefits of the diffKDE compared to standard KDEs by providing naturally a family of approximations. This family can be observed by the function *custom_plot*. Individual members can be produced by setting the bandwidth parameter T in the function call of the diffKDE. This gives the user the chance to choose among more and less smooth approximations. A selection of such approximations along with the default solution are shown in Fig. 5 on a random sample of 50 data points from the trimodal distribution in Eq. 35. The plot shows how smaller iteration times resolve more structure in the estimate, while a substantially larger iteration time has only little influence on the increased smoothing of the diffKDE.

From now on we only work with the default solution of the diffKDE at T^*. We start with comparisons of the diffKDE and the three other popular KDEs directly to the underlying true distribution. The three other KDEs are the Gaussian KDE in an implementation from SciPy (Gommers et al., 2022), the Epanechnikov KDE in an implementation from Scikit-learn (Pedregosa et al., 2012) and an improved Gaussian KDE by Botev et al. (2010) in a Python implementation by Hennig (2021).

We use differently sized random samples of the known distribution from Eq. 35 and the standard lognormal distribution both over $[-1, 12]$, for a direct comparison of the accuracy of the KDEs. The random samples are 50 and 100 data points of each distribution and all four KDEs are calculated and plotted together in Fig. 6. The underlying true distribution is plotted in the background to visually assess the approximation accuracy. In general, the diffKDE resolves more of the details of the structure of the true distribution, while not being too sensitive to patterns introduced by the selection of the random sample and

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

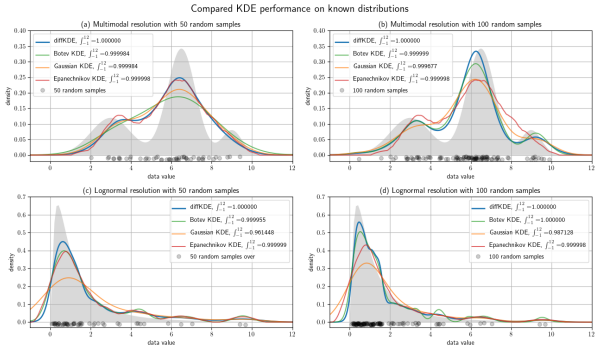

Figure 6. Test cases with known distributions: The plots (a) and (b) show KDEs of random samples of the trimodal distribution defined in Eq. 35, (c) and (d) the same for a lognormal distribution. The left figure column is constructed from 50 random samples, the right from 100. In all plots the true distribution is drawn in grey in the background and the random data sample as grey dots on the x-axis. Each subfigure shows four KDEs: the diffKDE, the Botev KDE, the Gaussian KDE and the Epanechnikov KDE. In the labels of the KDEs are also the integrals over the interval $[-1, 12]$ given for each of the KDEs

individual outliers. For the 50 random samples test of the trimodal distribution, all KDEs do not detect the third mode and only the diffKDE and the Epanechnikov KDE detect the second. The magnitude of the main mode is also best resolved by these two. In the 100 random samples test of the trimodal distribution, the diffKDE and the Botev KDE are able to detect all three modes. The main mode is best resolved by the diffKDE, whereas the third mode best by the Botev KDE. In both test cases for the trimodal distribution, the Gaussian KDE is the smoothest and the Epanechnikov KDE provides the least smooth graph. For 50 as well as for 100 random samples drawn from the lognormal distribution the magnitude and the steep decline to 0 is best reproduced by the diffKDE. The Gaussian KDE always performs the worst. The Botev KDE is generally also close to the diffKDE, but resolves in the tail of the distribution too much influence of individual outliers. An analysis of the the integral of the KDEs over the observed domain reveals that the diffKDE is the only one that integrates to 1 in all test cases.

We refined the test cases from Fig. 6 by investigating a lognormal distribution with different parameters and a restriction to the interval $[0, 12]$ in Fig. 7. We varied mean and variances of the normal distribution and used two different means and three different variances resulting in six test cases. All of them are run with 300 random samples and again with all four KDEs. The larger the variance becomes, the more structure of individual data points is resolved by the Botev KDE. The Gaussian KDE fails for increasing variance too, resulting in intense oversmoothing. The Epanechnikov KDE performs well for smaller variances and larger means, but also oversmoothes in the other cases. The diffKDE is generally one of the closest to the true distribution, while not resolving too much of the structure introduced by the choice of the random sample, especially for

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development
Discussions
Open Access
EGU

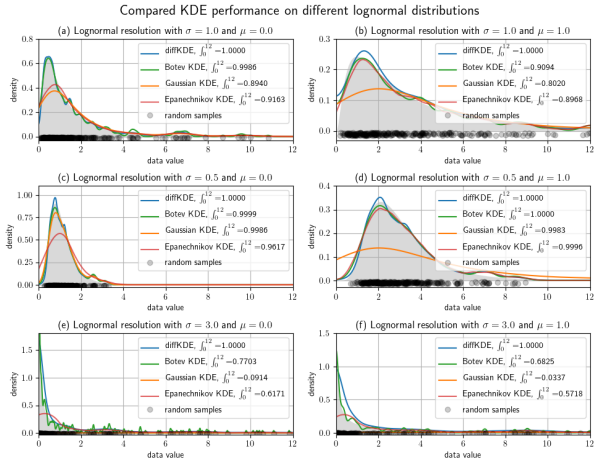

Figure 7. Lognormal test cases with different mean and variance parameters. Of each distribution 300 random samples were taken and the diffKDE, the Botev KDE, the Gaussian KDE and the Epanechnikov KDE calculated and plotted together with the true distribution. The random data sample is drawn as gray circles on the x-axis. (a) and (b) use $\sigma = 1$, (c) and (d) $\sigma = 0.5$ and (e) and (f) $\sigma = 3$ in their underlying normal distributions. The means are in the left column $\mu = 0$, the right column $\mu = 1$ of the underlying normal distribution.

increased variances. But this too tends to resolve too much structure in the vicinity of the mode for smaller variances. The integral of our implementation is again always exactly 1.

Now, we show the performance of the diffKDE on increasingly large data sets. We still use the trimodal distribution from Eq. 35. We start with four larger random data samples ranging from 100 to 10 million data points of the trimodal distribution and then being restricted to our core area of interest $[-1, 12]$. We calculate the diffKDE from all of them as well as the respective runtime on a consumer laptop from 2020. We compare the results again to the true distribution in Fig. 8. All of the estimates could be calculated in less than one minute. For 100 data points there is still an offset to the true distribution visible in the estimate. For the larger data samples the estimate only shows some minor uneven areas, which smooth out until the largest test case.

Furthermore, we investigated the convergence of the diffKDE to the true distribution, again in comparison to the three other KDEs. The error between the respective KDE and the true distribution is calculated by the Wasserstein distance (Panaretos and Zemel, 2019) with $p = 1$ by a SciPy function. We used increasingly large random samples from the trimodal distribution starting with 10 and reaching up to 1 million. The errors calculated for each of the KDEs on each of the random samples are

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

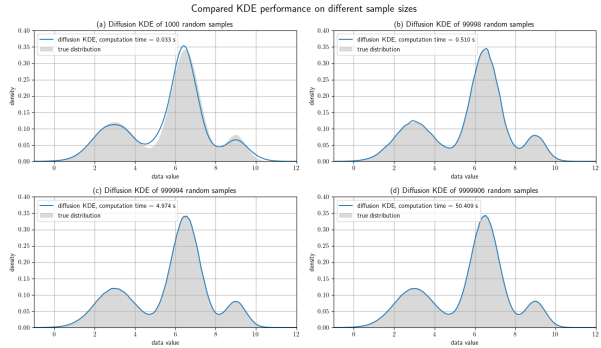

Figure 8. Test cases with different sample sizes: All four plots show the diffKDE of random samples of the known trimodal distribution defined in Eq. 35. (a) is calculated from a subsample of 100 data points, (b) $100,000$, (c) $1,000,000$ and (d) $10,000,000$, all cut to the interval $[-1,12]$ and hence lacking a few data points. The true numbers of incorporated data points in the four test cases are given in the respective sub-headings. The measured computing time on a 2020 MacBook Air is also drawn in the respective label.

listed in Tab. 3. The values from Tab. 3 are visualized in Fig. 9 on a log-scale and with a linear regression for each KDE's error values. The diffKDE, the Gaussian and the Botev show a similar steep decline, while the Epanechnikov KDE far slower decreases its error with increased sample size. The diffKDE and the Botev KDE generally show similar error values, the diffKDE relatively smaller ones on smaller data samples, the Botev KDE relatively smaller ones on data samples larger than around 5000.

Finally, we investigated the noise sensitivity of the diffKDE compared to the three other KDEs on data containing artificially introduced noise. We again used the trimodal distribution from Eq. 35 and 1000 random samples. From this, we created noised data $X_\theta \in \mathbb{R}^N$ by

$$(X_\theta)_i = (X)_i + (-1)^\tau \, rand 10^{-2} \theta \sigma \text{ for all } i \in \{1, ..., 1000\}, \tag{36}$$

where $\theta \in \{0, 1, 5, 15, 30\}$ defines the percentage of noise with respect to the standard deviation $\sigma \in \mathbb{R}$. $\tau \in \{1, 2\}$ was chosen randomly as well as $rand \in [0, 1]$. The error is again expressed by the Wasserstein distance between the original probability density and the respective KDE. The results are visualized in Fig. 10 with an individual panel for each KDE. The error of the Epanechnikov KDE is overall the largest and also increases to the largest. The Gaussian KDE produces the second largest error, but this even decreases with increased noise. The Botev KDE produces the smallest errors, but for increased noise this increases and approaches the magnitude of the one from the diffKDE. The error of the diffKDE only minimally responds to increased noise in the data. Visually, all four KDEs follow a similar pattern of a shift to the left of the graph. The Botev KDE additionally resolves more structure of the noised data as the noise increases.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions

Table 3. Error convergence

sample size	$error_{diffKDE}$	$error_{BKDE}$	$error_{GKDE}$	$error_{EKDE}$
10	0.02354	0.03618	0.02662	0.0273
50	0.01813	0.02484	0.02182	0.02017
100	0.00422	0.00724	0.01371	0.00702
150	0.00664	0.00937	0.01526	0.00933
200	0.00787	0.00894	0.01522	0.00967
300	0.0053	0.00621	0.01385	0.00849
400	0.00368	0.00484	0.01147	0.0081
500	0.0027	0.00324	0.01057	0.00761
750	0.00361	0.00343	0.00999	0.00807
1000	0.00321	0.00238	0.00933	0.00785
2000	0.00235	0.00171	0.00743	0.00771
5000	0.00154	0.00187	0.00578	0.00802
10000	0.00199	0.00188	0.00437	0.00791
50000	0.00113	0.00093	0.00234	0.00811
100000	0.00074	0.00059	0.00181	0.00822
500000	0.00048	0.00038	0.00108	0.00835
1000000	0.00046	0.00034	0.00084	0.00838

3.3 Performance analyses on biogeochemical data

In this final part, we show the diffKDE's performance on real marine biogeochemical field data. We chose two example data: A set of δ^{13}C in particulate organic carbon (POC) (Verwega et al., 2021) data and a set of plankton size spectra data (Lampe et al., 2021). Both data sets were already analyzed using KDEs in their original publications (Verwega et al., 2021; Lampe et al., 2021). Here we expand these analyses by a comparison of the KDEs used in the respective publications to the new implementation of the diffKDE. For the $\delta^{13}C_{POC}$ data, the Gaussian KDE was the one used in the data description publication. Since we have done this in the previous chapter, we furthermore added the Epanechnikov and the Botev KDE to these graphics. For the plankton size spectra data, we only compared the diffKDE to the two Gaussian KDEs used in the respective publication to preserve the clarity of the resulting figures.

The $\delta^{13}C_{POC}$ data (Verwega et al., 2021) was collected to serve for direct data analyses as well as for future model assessments (Verwega et al., 2021). We show here the Gaussian KDE as it was used in the data publication in a direct comparison to the diffKDE. Furthermore, we added the Epanechnikov and the Botev KDE. Since in this case no true known PDF is available, we have to compare the four estimates and subjectively judge their usefulness. In Fig. 11 we show the KDEs on four different

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development
Discussions
Open Access
EGU

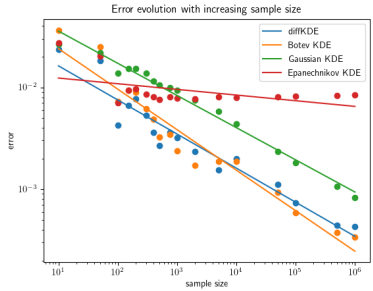

Figure 9. The evolution of the errors of the diffKDE, the Gaussian KDE, the Epanechnikov KDE and the Botev KDE are drawn on log-scale against the increasing sample size on the x-axis. The error has been calculated with the Wasserstein distance. A linear regression line on the log-scale is constructed from the discrete values of the individual errors for all four KDEs.

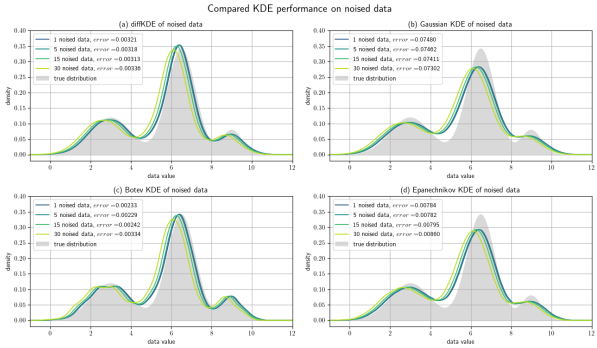

Figure 10. Noised data experiments: A random sample of 1000 data points of the trimodal distribution is artificially noised by differing amounts of the standard deviation. (a) shows the resulting diffKDEs of the differently noised data, (b) the Gaussian KDE, (c) the Botev KDE and (d) the Epanechnikov KDE. In all four panels the original true distribution is drawn in grey in the background. The values of the error between the KDEs and the original true distribution are also part of the respective labels.

subsets of the $\delta^{12}C_{POC}$ data: a) the full data set, b) a restriction to the core data interval of $[-35, -15]$, where $98.65\,\%$ of the data is located, and then even further restricted to c) the euphotic zone and d) only data sampled in the 1990s. In all three cases that

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development
Discussions
Open Access
EGU

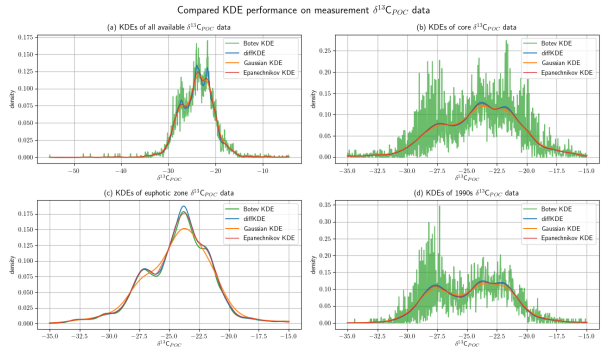

Figure 11. Comparison of KDE performance on marine biogeochemical field data: The $\delta^{13}C_{POC}$ data (Verwega et al., 2021) is in detail described in Verwega et al. (2021) and is covering all major world oceans, the 1960s to 2010s and reaches down into the deep ocean. In all four panels the diffKDE is plotted together with the Gaussian, the Epanechnikov and the Botev KDE. (a) Shows KDEs from all available data, (b) shows the KDEs of the data restricted to the core data values of $[-35, -15]$, (c) shows the KDEs from only euphotic zone data with values in $[-35, -15]$ and (d) the KDEs from all 1990s data with values in $[-35, -15]$.

involve deep ocean measurements, the Botev KDE produces strong oscillations while the Gaussian KDE strongly smoothes the dip between the modes at around $\delta^{13}C_{POC} = -24$ and $\delta^{13}C_{POC} = -22$ and mostly the one between around $\delta^{13}C_{POC} = -28$ and $\delta^{13}C_{POC} = -24$. The Epanechnikov KDE resolves more structure than the Gaussian, but still less pronounced than the diffKDE. Especially in the full data analysis, the diffKDE reveals the most structure while not resolving smaller data features of individual data points. The KDEs from the euphotic zone data are all reasonably smooth. The Gaussian KDE is again the smoothest and missing the mode at $\delta^{13}C_{POC} = -22$ completely. The other three KDEs resolve a similar amount of data structure. The Botev KDE reveals a better distinction between the modes at around $\delta^{13}C_{POC} = -24$ and $\delta^{12}C_{POC} = -22$ while the diffKDE shows the first one more pronounced. These observations are consistent with those from the experiments from Fig. 7 and Fig. Figure 10, where especially the Gaussian and the Botev KDE struggle with the resolution of data with increasing variances or noise. From the four here observed $\delta^{12}C_{POC}$ data sets the euphotic zone data shown in panel (c) in Fig. 11 has with 7.78 the smallest standard deviation. The other shown data has variances 13.91, 10.96 and 9.61 for panels (a), (b) and (d), respectively.

Another example demonstrates the performance of the diffKDE if applied to plankton size data (Lampe et al., 2021). The data of size, abundance of protist plankton was originally collected for resolving changes in plankton community size-structure, providing complementary insight for investigations of plankton dynamics and organic matter flux (e.g., Nöthig et al., 2015). In the study of Lampe et al. (2021) a KDE was applied for the derivation of continuous size spectra of phytoplankton and

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

microzooplankton that can potentially be used for the calibration and assessment of size-based plankton ecosystem models. In their study they used a Gaussian KDE, as proposed in Schartau et al. (2010), but with two different approaches for generating plankton size spectra. Uncertainties, also with respect to optimal bandwidth selection, were accounted for in both approaches by analyzing ensembles of pseudo-data resampled from original microscopic measurements. Smooth plankton spectra were obtained using the *combined* approach, where all phytoplankton and all zooplankton data were lumped together respectively and single bandwidths were calculated for every ensemble member (set of resampled data). This procedure avoided over-fitting but was also prone to over-smoothing. More structured size spectra were obtained with the *composite* approach, where individual size spectra were calculated for each species or genus and then pieced together. Since the variance within species or genus groups is smaller than within the large groups 'phytoplankton' or 'zooplankton', resulting bandwidths and therefore the degree of smoothing were considerably smaller than in the combined approach. This computationally expensive method revealed many details in the spectra, but at the same time tended to resolve narrow peaks that were either clearly insignificant or remained difficult to interpret (see supplemental material in Lampe et al. (2021)).The here proposed diffKDE is tested with resampled data used for the simpler *combined* approach. The objective is to identify details in the size spectra that remained previously unresolved while insignificant peaks, as found in the composite approach, become smoothed out. Figure 12 shows the performance of the diffKDE in comparison to the original combined and composite spectra that were derived as ensemble means of estimates obtained with a Gaussian KDE. The spatial discretization of the diffKDE was set to $n = 600$ to be comparable to the other already published KDEs in this case. The diffKDE seems to meaningfully combine the advantages of the two Gaussian KDE approaches in both spectra, of the phytoplankton and microzooplankton respectively. With the diffKDE it is possible to generate estimates that display more detailed structure of the composite KDE for cell sizes smaller than 10 μm, in particular in the microzooplankton spectrum. Concurrently, detailed variations, as caused by overfitting in the composite spectra, become suppressed for cell sizes larger than 10 μm. Thus, with the diffKDE it is possible to generate a single robust estimate that otherwise is only achieved by analyzing a series of estimates of a Gaussian KDE.

3.4 Future application to model calibration

The robustness of Earth system models is crucial for providing reliable climate projections for a sustainable development into Earth's future. Such models can assist the understanding of past and present and predict future conditions in the Earth system. Earth system models simulate the ocean's element cycling (e.g., Ilyina et al., 2013) and with this the ocean's carbon uptake capacity (e.g., Frölicher et al., 2015). They serve to assess the current and future state of our climate system and provide projections for different mitigation scenarios. This information can be used to support a sustainable development in our climate system (IPCC, 2022). As a consequence, political decisions depend on reliable projections to construct a safe pathway into Earth's future.

Calibration can increase the reliability of Earth system models (e.g., Oliver et al., 2022). For this purpose, a metric calculates the difference between simulated model output and measured field data. This metric defines the target or cost function in an optimization process, where unknown or uncertain model parameters are identified or estimated by numerical algorithms. This

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

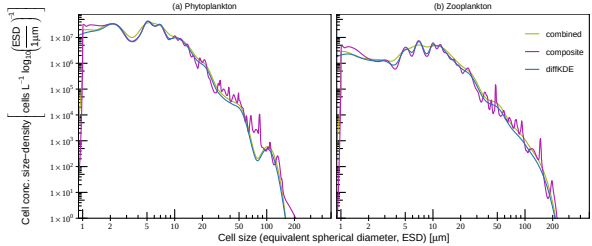

Figure 12. Comparison of KDE performance on (a) phytoplankton and (b) microzooplankton size spectra. The construction of composite and combined size spectra is described in Lampe et al. (2021) and based on Gaussian KDEs.Smoother combined spectra are the result of one KDE with a common bandwidth for all data. More structured composite spectra were assembled from taxon-specific spectra with individual, hence smaller, bandwidths.

process is sometimes also called "tuning" of the model. The result is usually a single or multiple sets of "optimal" parameters. They provide the model configuration with results closest to the incorporated field data.

Comparison of model and field data requires additional processing to account for spatial-temporal differences between collected samples and model resolution. Typically, simulation results are available at every single spatial grid point and in every time step. In comparison, field data are usually sparsely available only. Interpolating such sparse field data can introduce high uncertainty (e.g., Oliver et al., 2022). PDFs provide a useful approach to investigate data independent of the number of data points available (Thorarinsdottir et al., 2013). A comparison of two such functions can easily resolve the issue of non-equal field observations and simulation results. Histograms are commonly used as an approach to compare and ultimately constrain the distribution of model data to observations. However, many issues arrive including the subjective selection of intervals and histograms not being proper PDFs themselves.

The presented diffKDE provides a non-parametric approach to estimate PDFs with typical features of geoscientific data. Being able to resolve typical patterns such as multiple or boundary close modes, while being insensitive to noise and individual outliers makes the diffKDE a suitable tool for future work in the calibration and optimization of Earth system models.

4 Summary and conclusions

In this study we constructed and tested an estimator (KDE) of probability density functions (PDFs) that can be applied for analysing geoscientifc and ecological data. KDEs allow the investigation of data with respect to their probability distribution, and PDFs can be derived even for sparse data. To be well suited for geoscientific data, the KDE must work fast and reliably on differently sized data sets, while revealing multimodal details as well as features nearby data boundaries. A KDE should not

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

be overly sensitive to noise introduced by measurement errors or by numerical uncertainties. Such an estimator can be applied for direct data analyses or can be used to construct a target function for model assessment and calibration.

We presented a novel implementation of a KDE based on the diffusion heat process (diffKDE). This idea was originally proposed by Chaudhuri and Marron (2000) and its benefits in comparison to traditional KDE approaches were widely investigated by Botev et al. (2010). Our approach combines the solution of the diffusion equation with with two pilot estimation steps that correspond to the Gaussian KDE. We used an approximation of the optimal bandwidth for the diffKDE by a central differential quotient and plug-in of the pilot estimates. For their bandwidths we used variations of the *rule of thumb* by Silverman (1986). Our approach results in three subsequent estimations of the PDF, each of them chosen with a finer bandwidth approximation.

Finite differences build the fundamentals of our discretization. The spatial discretization are equidistant finite differences. The δ-distribution in the initial value is discretized by piecewise linear functions along the spatial discretization points constructing a Dirac-sequence. For the timestepping we applied an implicit Eulerian algorithm on an ordinary differential equation set up by a tridiagonal matrix corresponding to the diffusion equation on the spatial equidistant grid.

Our diffKDE implementation includes pre-implemented default output options. The first is the visualization of the diffusion time evolution showing the sequence of all solution steps from the initial values to the final diffKDE. This lets a user see the influence of individual data points and outlier accumulations on the final diffKDE and how this decreases over time. The second is the visualization of the pilot estimate that is also included in the partial differential equation to introduce adaptive smoothing properties. This provides the user an easy insight into the adaptive smoothing as well as the lower boundary of structure resolution given by this parameter function. Finally, an interactive plot provides a simple opportunity to explore all of these time iterations and look even beyond the optimal bandwidth and see smoother estimates.

Our implementation is fast and reliable on differently sized and multimodal data sets. We tested the implementation for up to 10 million data points and obtained acceptably fast results. A comparison of the diffKDE on known distributions together with classically employed KDEs showed reliable and often superior performance. For comparison we chose a SciPy implementation (Gommers et al., 2022) of the most classical Gaussian KDE (Sheather, 2004), an Scikit implementation (Pedregosa et al., 2012) of an Epanechnikov KDE (Scott, 1992) and a Python implementation (Hennig, 2021) of the improved Gaussian KDE developed by Botev et al. (2010). We designed multimodal and different boundary-close distributions and found our implementation to generate the most reliable estimates across a large range of sample sizes (Fig. 9). The diffKDE was neither prone to oversmoothing nor overfitting of the data, which we could observe in the other tested KDEs. A noise sensitivity test in comparison to the other KDEs also showed a good stability of the diffKDE against noise in the data.

An assessment of the diffKDE on real marine biogeochemical field data in comparison to usually employed KDEs reveals superior performance of the diffKDE. We used carbon isotope and plankton size spectra data and compared the diffKDE to the KDEs that were used to explore the data in the respective original data publications. On the carbon isotope data, we furthermore applied all previous KDEs for comparison. In both cases we were able to show that the diffKDE resolves relevant features of the data while not being sensitive to individual outliers or uncertainties (noise) in the data. We were able to obtain a best possible and reliable represantation of the true data distribution, better than those derived with other KDEs.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

In future studies the diffKDE may potentially be used for the assessment, calibration and optimization of marine biogeochemical- and Earth system models. Already a plot of PDFs, of field data and simulation results respetcively, may provide visual insight into some shortcomings of the applied model. A target function can be constructed by adding a distance like the Wasserstein distance (Panaretos and Zemel, 2019) or other useful metrics for the calibration of climate models that can be investigated (Thorarinsdottir et al., 2013). Thus, KDE applications such as our diffKDE can greatly simplify comparisons of differently sized field and simulation data sets.

Code availability. The exact version of the diffKDE implementation (Pelz and Slawig, 2023) used to produce the results used in this paper is archived on Zenodo: https://doi.org/10.5281/zenodo.7594915.

Appendix A

Here, we briefly give the proof of the integral property of the used Dirac sequence Φ_h defined in Equation 25. Let $h \in \mathbb{R}_{>0}$. Then we obtain

$$\begin{aligned}
\int \Phi_h(x)\,dx &= \int_{x_{i-2}}^{x_{i-1}} \Phi_h(x)\,dx + \int_{x_{i-1}}^{x_i} \Phi_h(x)\,dx + \int_{x_i}^{x_{i+1}} \Phi_h(x)\,dx \\
&= \frac{1}{2}(x_{i-2} - x_{i-1}) \frac{1}{x_{i-2}-x_{i-1}} \frac{x_i}{x_i - x_{i-1}} + \frac{1}{2}(x_i - x_{i-1}) \left(\frac{1}{x_{i-2}-x_{i-1}} \frac{x_i}{x_i - x_{i-1}} + \frac{1}{x_{i+1}-x_i} \frac{-x_{i-1}}{x_i - x_{i-1}} \right) \\
&\quad + \frac{1}{2}(x_{i+1} - x_i) \frac{1}{x_{i+1}-x_i} \frac{-x_{i-1}}{x_i - x_{i-1}} \\
&= \frac{1}{2} h \frac{1}{h} \frac{x_i}{h} + \frac{1}{2} h \left(\frac{1}{h} \frac{x_i}{h} + \frac{1}{h} \frac{-x_{i-1}}{h} \right) + \frac{1}{2} h \frac{1}{h} \frac{-x_{i-1}}{h} \\
&= \frac{1}{2} \frac{x_i}{h} + \frac{1}{2} \frac{x_i}{h} - \frac{1}{2} \frac{x_{i-1}}{h} - \frac{1}{2} \frac{x_{i-1}}{h} \\
&= \frac{x_i - x_{i-1}}{h} \\
&= 1.
\end{aligned} \tag{A1}$$

Author contributions. MTP set up the manuscript and developed the implementation of the diffusion-based kernel density estimator. VL conducted the comparison experiments with the plankton size spectra. CJS edited the manuscript. MS and TS edited the manuscript and supported the development of the implementation of the diffusion-based kernel density estimator.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

Acknowledgements. The first author is funded through the Helmholtz School for Marine Data Science (MarDATA), Grant No. HIDSS-0005.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

References

Abramson, I. S.: On bandwidth variation in kernel estimates-a square root law, The annals of Statistics, pp. 1217–1223, 1982.

Boccara, N.: Functional Analysis - An Introduction for Physicists, Academic Press, Inc., 1990.

Botev, Z. I., Grotowski, J. F., and Kroese, D. P.: Kernel density estimation via diffusion, The annals of Statistics, 38, 2916–2957, https://doi.org/10.1214/10-AOS799, 2010.

Breiman, L., Meisel, W., and Purcell, E.: Variable kernel estimates of multivariate densities, Technometrics, 19, 135–144, 1977.

Chaudhuri, P. and Marron, J.: Scale space view of curve estimation, ANNALS OF STATISTICS, 28, 408–428, https://doi.org/10.1214/aos/1016218224, 2000.

Chung, Y.-W., Khaki, B., Chu, C., and Gadh, R.: Electric Vehicle User Behavior Prediction Using Hybrid Kernel Density Estimator, in: 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), pp. 1–6, https://doi.org/10.1109/PMAPS.2018.8440360, 2018.

Deniz, T., Cardanobile, S., and Rotter, S.: A PYTHON Package for Kernel Smoothing via Diffusion: Estimation of Spike Train Firing Rate, Front. Comput. Neurosci. Conference Abstract: BC11 : Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting 2011, 5, https://doi.org/10.3389/conf.fncom.2011.53.00071, 2011.

Dirac, P. A. M.: The physical interpretation of the quantum dynamics, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 113, 621–641, https://doi.org/10.1098/rspa.1927.0012, 1927.

Farmer, J. and Jacobs, D. J.: MATLAB tool for probability density assessment and nonparametric estimation, SoftwareX, 18, 101 017, https://doi.org/10.1016/j.softx.2022.101017, 2022.

Frölicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., and Winton, M.: Dominance of the Southern Ocean in Anthropogenic Carbon and Heat Uptake in CMIP5 Models, Journal of Climate, 28, 862–886, https://doi.org/10.1175/jcli-d-14-00117.1, 2015.

Gommers, R., Virtanen, P., Burovski, E., Weckesser, W., Oliphant, T. E., Cournapeau, D., Haberland, M., Reddy, T., alexbrc, Peterson, P., Nelson, A., Wilson, J., endolith, Mayorov, N., Polat, I., van der Walt, S., Laxalde, D., Brett, M., Larson, E., Millman, J., Lars, peterbell10, Roy, P., van Mulbregt, P., Carey, C., eric jones, Sakai, A., Moore, E., Kai, and Kern, R.: scipy/scipy: SciPy 1.8.0, https://doi.org/10.5281/zenodo.5979747, 2022.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Heidenreich, N.-B., Schindler, A., and Sperlich, S.: Bandwidth selection for kernel density estimation: a review of fully automatic selectors, AStA Advances in Statistical Analysis, 97, 403–433, https://doi.org/10.1007/s10182-013-0216-y, 2013.

Hennig, J.: John-Hennig/KDE-diffusion: KDE-diffusion 1.0.3, https://doi.org/10.5281/zenodo.4663430, 2021.

Hirsch, F. and Lacombe, G.: Elements of Functional Analysis, Springer, 1999.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in science & engineering, 9, 90–95, https://doi.org/10.1109/mcse.2007.55, 2007.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, Journal of Advances in Modeling Earth Systems, 5, 287–315, https://doi.org/10.1029/2012ms000178, 2013.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Geoscientific Model Development Discussions
Open Access
EGU

IPCC: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)], https://doi.org/10.1017/9781009157926, 2022.

Jones, M. C., Marron, J. S., and Sheather, S. J.: A Brief Survey of Bandwidth Selection for Density Estimation, Journal of the American Statistical Association, 91, 401–407, https://doi.org/10.1080/01621459.1996.10476701, 1996.

Khorramdel, B., Chung, C. Y., Safari, N., and Price, G. C. D.: A Fuzzy Adaptive Probabilistic Wind Power Prediction Framework Using Diffusion Kernel Density Estimators, IEEE Transactions on Power Systems, 33, 7109–7121, https://doi.org/10.1109/tpwrs.2018.2848207, 2018.

Lampe, V., Nöthig, E.-M., and Schartau, M.: Spatio-Temporal Variations in Community Size Structure of Arctic Protist Plankton in the Fram Strait, Frontiers in Marine Science, 7, https://doi.org/10.3389/fmars.2020.579880, 2021.

Ma, S., Sun, S., Wang, B., and Wang, N.: Estimating load spectra probability distributions of train bogie frames by the diffusion-based kernel density method, International Journal of Fatigue, 132, 105 352, https://doi.org/10.1016/j.ijfatigue.2019.105352, 2019.

Marron, J. S. and Ruppert, D.: Transformations to reduce boundary bias in kernel density estimation, Journal of the Royal Statistical Society: Series B (Methodological), 56, 653–671, https://www.jstor.org/stable/2346189, 1994.

McSwiggan, G., Baddeley, A., and Nair, G.: Kernel Density Estimation on a Linear Network, Scandinavian Journal of Statistics, 44, 324–345, https://doi.org/10.1111/sjos.12255, 2016.

Nöthig, E.-M., Bracher, A., Engel, A., Metfies, K., Niehoff, B., Peeken, I., Bauerfeind, E., Cherkasheva, A., Gäbler-Schwarz, S., Hardge, K., Kilias, E., Kraft, A., Mebrahtom Kidane, Y., Lalande, C., Piontek, J., Thomisch, K., and Wurst, M.: Summertime plankton ecology in Fram Strait - a compilation of long- and short-term observations, Polar Research, 34, 23 349, https://doi.org/10.3402/polar.v34.23349, 2015.

Oliver, S., Cartis, C., Kriest, I., Tett, S. F. B., and Khatiwala, S.: A derivative-free optimisation method for global ocean biogeochemical models, Geoscientific Model Development, 15, 3537–3554, https://doi.org/10.5194/gmd-15-3537-2022, 2022.

Panaretos, V. M. and Zemel, Y.: Statistical Aspects of Wasserstein Distances, Annual Review of Statistics and Its Application, 6, 405–431, https://doi.org/10.1146/annurev-statistics-030718-104938, 2019.

Parzen, E.: On estimation of a probability density function and mode, The annals of mathematical statistics, 33, 1065–1076, 1962.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, https://doi.org/10.48550/ARXIV.1201.0490, 2012.

Pedretti, D. and Fernàndez-Garcia, D.: An automatic locally-adaptive method to estimate heavily-tailed breakthrough curves from particle distributions, Advances in Water Resources, 59, 52–65, https://doi.org/10.1016/j.advwatres.2013.05.006, 2013.

Pelz, M.-T. and Slawig, T.: Diffusion-based kernel density estimator (diffKDE), https://doi.org/10.5281/ZENODO.7594915, 2023.

Qin, B. and Xiao, F.: A Non-Parametric Method to Determine Basic Probability Assignment Based on Kernel Density Estimation, IEEE Access, 6, 73 509–73 519, https://doi.org/10.1109/ACCESS.2018.2883513, 2018.

Schartau, M., Landry, M. R., and Armstrong, R. A.: Density estimation of plankton size spectra: a reanalysis of IronEx II data, Journal of Plankton Research, 32, 1167–1184, https://doi.org/10.1093/plankt/fbq072, iSBN: 0142-7873, 2010.

Schmittner, A. and Somes, C. J.: Complementary constraints from carbon (^{13}C) and nitrogen (^{15}N) isotopes on the glacial ocean's soft-tissue biological pump, Paleoceanography, pp. 669–693, https://doi.org/10.1002/2015PA002905, 2016.

https://doi.org/10.5194/gmd-2023-17
Preprint. Discussion started: 13 February 2023

Scott, D. W.: Multivariate density estimation: theory, practice, and visualization, John Wiley & Sons, 1992.

Scott, D. W.: Multivariate density estimation and visualization, in: Handbook of computational statistics, pp. 549–569, Springer, https://doi.org/10.1007/978-3-642-21551-3-19, 2012.

Sheather, S. J.: Density Estimation, Statistical Science, 19, 588–597, https://doi.org/10.1214/088342304000000297, 2004.

Sheather, S. J. and Jones, M. C.: A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society: Series B (Methodological), 53, 683–690, 1991.

Silverman, B.: Density estimation, Monographs on Statistics and Applied Probability, 1986.

Terrell, G. R. and Scott, D. W.: Variable kernel density estimation, The Annals of Statistics, pp. 1236–1265, https://www.jstor.org/stable/2242011, 1992.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N.: Using Proper Divergence Functions to Evaluate Climate Models, SIAM/ASA Journal on Uncertainty Quantification, 1, 522–534, https://doi.org/10.1137/130907550, 2013.

Van Rossum, G.: The Python Library Reference, release 3.8.2, Python Software Foundation, 2020.

Verwega, M.-T., Somes, C. J., Schartau, M., Tuerena, R. E., Lorrain, A., Oschlies, A., and Slawig, T.: Description of a global marine particulate organic carbon-13 isotope data set, Earth System Science Data, 13, 4861–4880, https://doi.org/10.5194/essd-13-4861-2021, 2021.

Verwega, M.-T., Somes, C. J., Tuerena, R. E., and Lorrain, A.: A global marine particulate organic carbon-13 isotope data product, https://doi.org/10.1594/PANGAEA.929931, 2021.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Chapter 5

Other approaches to the diffusion kernel density estimator

In this chapter, I explore possibilities to implement a diffusion KDE that differ from my published version [PS23]. These include my first (discontinued) version with a discretization based on finite elements, a possible finite elements based implementation in FEniCS [ABH+15] and a version proposed by Botev et al. [BGK10] and implemented by Deniz et al. [DCR11] in Python.

5.1 A finite element approach to the diffusion kernel density estimator

Finite elements (FEM) were my first approach to discretize the diffusion KDE. This follows a common solution strategy for time dependent partial differential equations in two steps: first, discretize the time derivation by finite differences and second, derive the variational formulation in each of these time step. The result is a sequence of variational problems [LMW12].

In the following, I will show the FEM discretization on the specific example of the diffusion equation from Eq. 3.0.3. Let $n \in \mathbb{N}$ and u^n denote the solution $u \in C^{2,1}\left(\Omega \times \mathbb{R}_{>0}, \mathbb{R}_{\geqslant 0}\right)$ of Eq. 3.0.3 at the n-th time step. Now, we use a backward Euler to discretize the time stepping. By this, we receive a sequence of time stationary differential equations:

$$\frac{u^{n+1} - u^n}{\Delta t} = \frac{\partial}{\partial t} u^{n+1} = \frac{1}{2} \nabla \left(\nabla \left(\frac{u^{n+1}}{p} \right) \right) \tag{5.1.1}$$

We use this backward differential quotient to approximate the time deriva-

tive on the left hand side in Eq. 3.0.3.

Now, we identify the solution of the following time step $u^{n+1} =: u$ to be the unknown and the current u^n as the known initial value from the previous time step. This delivers

$$u - \Delta t \frac{1}{2} \nabla \left(\nabla \left(\frac{u}{p} \right) \right) = u^n \tag{5.1.2}$$

The solution u for Eq. 5.1.2 shall from now on be searched within a trial space defined as

$$V = \{ v \in H^1(\Omega); \frac{\partial}{\partial \nu} \left(\frac{v(x)}{p(x)} \right) = 0, x \in \partial\Omega \} \tag{5.1.3}$$

The restriction to V ensures the solution satisfying the Neumann boundary condition from Eq. 3.0.4 [LMW12].

Next, we define a test space as

$$\hat{V} = \{ v \in H^1(\Omega); v(x) = 0, x \in \partial\Omega \} \tag{5.1.4}$$

and use elements from $\hat{V}$ for multiplication of Eq. 5.1.2 to receive

$$u\,v - \Delta t \frac{1}{2} \nabla \left(\nabla \left(\frac{u}{p} \right) \right) v = u^n v. \tag{5.1.5}$$

Integration by parts and the property that $v(x) = 0$ for all $v \in \hat{V}$ and $x \in \partial\Omega$ delivers for all $v \in \hat{V}$

$$\begin{aligned} & \int_\Omega u\,v\,dx + \int_\Omega \Delta t \frac{1}{2} \nabla \left(\nabla \left(\frac{u}{p} \right) \right) v\,dx = \int_\Omega u^n v\,dx \\ \Rightarrow & \int_\Omega u\,v\,dx + \Delta t \frac{1}{2} \int_\Omega \nabla \left(\nabla \left(\frac{u}{p} \right) \right) v\,dx = \int_\Omega u^n v\,dx \\ \Rightarrow & \int_\Omega u\,v\,dx + \Delta t \frac{1}{2} \int_\Omega \left(\nabla \left(\frac{u}{p} \right) \right) \nabla v\,dx = \int_\Omega u^n v\,dx \end{aligned} \tag{5.1.6}$$

The last line (Eq. 5.1.6) is the so called *weak formulation* of the stationary diffusion partial differential equation Eq. 3.0.3.

The diffusion KDE can now be derived by solving Eq. 5.1.6 in every time step $n \in \mathbb{N}$ up to a final iteration time $T \in \mathbb{R}_{>0}$. For the first time step, u_0 is the initial condition given in Eq. 3.0.5. In every following time step$n \in \mathbb{N}$, the initial value u^n denotes the solution from the previous time

step and Eq. 5.1.6 is solved for the next one denoted as u. The solution u from the final time step is the diffusion KDE.

5.2 FEniCS implementation of the diffusion kernel density estimator

A simple and fast way to implement a solver for a FEM approach discussed in Sec. 5.1 is given by the software framework FEniCS [ABH+15; LMW12; LL16]. FEniCS is a free Python tool for the calculation of FEM solutions of partial differential equations. In this section, I present a possible FEM implementation to calculate the diffusion KDE as proposed in Eq. 5.1.6. This was my first attempt to the diffusion KDE realized in FEniCS. I discarded it later on due to the demanding prerequisites for running the FEniCS software.

As before, we use an input data vector $X \in \Omega^N$, where $\Omega \subseteq \mathbb{R}$ is a domain and $N \in \mathbb{N}$ the number of data points.

The calculations are carried out over an equidistant spatial discretization of $nel \in \mathbb{N}$ discretization points between the interval boundaries $x_{min} < x_{max} \in \Omega$. Boundaries of the spatial domain are minimum and maximum value of the input data. The trial and test space from Equation 5.1.3 and Equation 5.1.4 are set equal, since FEniCS does not account for boundary conditions in its definition of function spaces. The Neumann boundary conditions for the trial function u are default in FEniCS [ABH+15; LW10]. As FEM, I chose the linear Lagrange (P_1) functions. These are triangles with vertices at every edge. The solution $u \in V$ is continuous over Ω and linearly within each FEM. The δ-distribution in the initial value can be approximated by point sources at each data point. To ensure that the estimate integrates to 1, I weigh the point sources by a mass of

$$point\_mass = \frac{nel}{N\,(x_{max} - x_{min})}.$$

For comparability, I set the number of spatial discretization points, time step size and number equal those in the finite differences approach presented in Chap. 4. As well are the approximation of the optimal final iteration times for u and the two pilot estimates, all described in detail for

the finite differences implementation. The solution is delivered as a vector of the diffusion KDE values on the spatial grid together with this. The full implementation is described in Alg. 1.

Algorithm 1: Algorithm for FEM solver for the diffusion KDE

Require: $X \in \mathbb{R}^N$

$\Omega \leftarrow (x_{min}, x_{min} + h, ..., x_{max} - h, x_{max}) \in \mathbb{R}^{n+1}$

$V, \hat{V} \leftarrow \{v \in H^1(\Omega); v(x;t) = 0, x \in \partial\Omega\}$

$p^0, f^0, u^0 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i)$

$T_p \leftarrow \sigma^2 \left(\frac{4}{3}N\right)^{-\frac{2}{5}}$

$T_f \leftarrow \left(0.9min\left(\sigma, \frac{iqr(data)}{1.34}\right)\right)^2 N^{-\frac{2}{5}}$

$t \leftarrow 0$, $timesteps \leftarrow 20$, $\Delta_p \leftarrow T_p/timesteps$

while $t < T_p$ **do**

  solve: $\int_\Omega p\, v\, dx + \Delta_p t \frac{1}{2}\int_\Omega \nabla p \nabla v\, dx = \int_\Omega p^n v\, dx \; \forall v \in \hat{V}$

  $t \leftarrow t + \Delta_p$

end while

$t \leftarrow 0$, $\Delta_f \leftarrow T_f/timesteps$

while $t < T_p$ **do**

  solve: $\int_\Omega p\, v\, dx + \Delta_f t \frac{1}{2}\int_\Omega \nabla f \nabla v\, dx = \int_\Omega f^n v\, dx \; \forall v \in \hat{V}$

  $t \leftarrow t + \Delta_f$

end while

$q \leftarrow \sqrt{\int_\Omega \left(\left(\frac{f}{p}\right)''\right)^2 dh}$

$E_\sigma \leftarrow \frac{1}{n+1}\sum_{i=0}^{n+1}\sqrt{p(x_i)}$

$T \leftarrow \left(\frac{E_\sigma}{2N\sqrt{\pi}q^2}\right)^{\frac{2}{5}}$

$t \leftarrow 0$, $\Delta \leftarrow T/timesteps$

while $t < T$ **do**

  solve: $\int_\Omega u\, v\, dx + \Delta t \frac{1}{2}\int_\Omega \nabla\left(\frac{u}{p}\right)\nabla v\, dx = \int_\Omega u^n v\, dx \; \forall v \in \hat{V}$

  $t \leftarrow t + \Delta$

end while

return Ω, u

The main differences from Alg. 1 to the finite differences implementation are the solvers of the differential equations for p, f and u in line 8, 13 and 21.

Next, I compare the performance of the FEniCS implementation following Alg. 1 to the published finite differences approach [PS23]. The first test case is shown in Fig. 5.1 and build on known distributions: a trimodal and a lognormal one. From both a random sample of 50 and 100 data points are collected and both KDEs calculated an drawn together. In all four cases the KDEs align well. In the main modes, the KDEs generally differ the most. The finite differences approach resolves the true structure better, but this is also true for the structure of individual outliers, mostly prominent in the lognormal test cases.

A second test case shows the performance of the two different diffusion KDE implementations on the $\delta^{13}C_{POC}$ data [VST+21] in Fig. 5.2. The chosen test cases correspond to Fig. 11 from the diffusion KDE description in Chap. 4. Here, the true underlying dirstibutions are unknown and therefore not availabe for comparison. Furthermore, the two estimators are so closely aligned that I decided to evaluate their difference by the Wasserstein distance [PZ19]. Their difference is in the magnitude of 0.0001 to 0.0002 and nearly undetectable in visual examination of the graphs.

5.3 The approach by Botev et al. (2010)

In their publication about the diffusion KDE [BGK10], Botev et al. provide a discussion of the benefits of this approach and their own idea on how to design a possible implementation. Their idea is built on a fixed point iteration to solve optimal bandwidths implicit dependency on the true distribution. In this section, I want to briefly discuss their approach and show results of an implementation of this in comparison to my own algorithm.

The algorithm by [BGK10] is based on a new bandwidth approximation algorithm refining ideas by [SJ91]. The idea is a fixed point iteration for simultaneously solving the optimal bandwidth T^* and the squared L_2-norm of the second derivative of the PDF $||f^{(2)}||^2$. The iteration starts backwards with an initial guess for T^* at machine precision and a guess

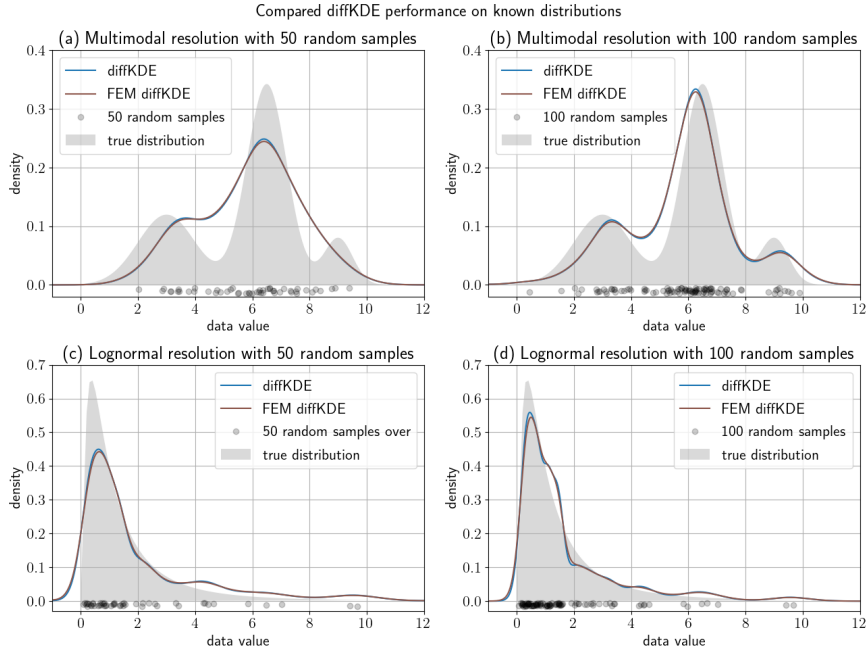

Figure 5.1. Diffusion KDE performance on known distributions: All four plots show the diffKDE and the FEM diffusion KDE on differently sized random samples of known distributions. (a) and (b) used a trimodal distribution, (c) and (d) a lognormal distribution. The true distribution is drawn as the grey shaded area and the random data samples as grey circles on the x-axis.

for $||f^{(8)}||^2$. The latter guess can then be used to estimate a better approximation for T^* and this for an approximation of $||f^{(7)}||^2$ and so on.

The full algorithm by Botev et al. is illustrated in Alg. 2 and corresponds to ALG1 in [BGK10]. It uses the fixed point iteration to calculate the pilot p evaluated at T_p and $||f''||^2$ to calculate the optimal bandwidth T^* for the diffusion KDE.

The first author of [BGK10] (Botev) provided an implementation of the pilot step (Alg. 2 up to line 9) of their idea. I will refer to this KDE as the Botev KDE. This implementation relies on the invariance of the diffusion equation under the Fourier transformation [LT05]. The

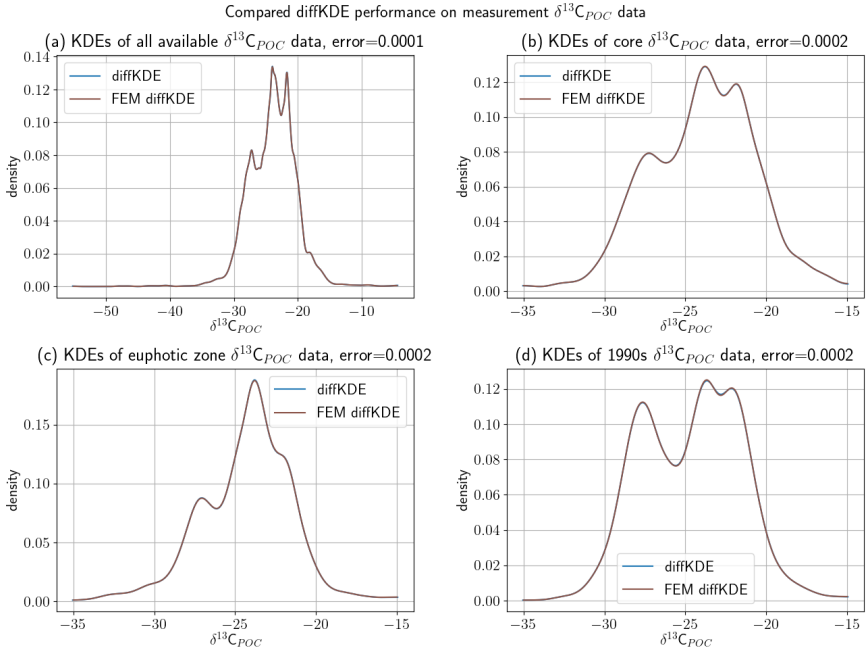

Figure 5.2. Compared FEM diffusion KDE and diffKDE performance on $\delta^{13}C_{POC}$ data: (a) shows all available $\delta^{13}C_{POC}$ data, (b) a restriction to the data core value area of $[-35, -15]$, (b) only euphotic zone data over the core interval and (d) only 1990s data over the core interval.

implementation is provided by Botev in Matlab and re-implemented in Phython by [Hen21]. This approach provides a fast solution in between of Fourier transformation and back transformation. Approximation of the optimal bandwidth and data smoothing are performed in the Fourier space. The full implemented algorithm for the calculation of the Botev KDE is given in Alg. 3. Lines 1 and 2 of Alg. describe the grid set up and the interpolation of the data onto this. Lines 3 to 6 transform the data into the Fourier space and lines 7 and 8 prepare the initial value for the fixed point iteration. Lines 9 to 13 describe the fixed point algorithm for the solution of the optimal bandwidth T_p corresponding to lines 3 to 7

Algorithm 2: The algorithm by Botev et al. (numbers in brackets are referrin to the equation numbers in ALG1 [BGK10])

Require: $X_1, ..., X_N \in \mathbb{R}$

1: $t_{l+1} = eps$

2: $||f^{(l+1)}||^2 \leftarrow \frac{(-1)^l}{N^2} \Sigma_{k=1}^N \Sigma_{m=1}^N \Phi^{(2l)} \left(X_k, X_m; 2t_{l+1}\right) |_{x=X_j}$ {(27), [BGK10]}

3: **for** $l = 7$ **to** 2 : **do**

4: $\quad t_l \leftarrow \left(\frac{1+\left(\frac{1}{2}\right)^{l+\frac{1}{2}}}{3} \frac{1 \times 3 \times ... \times (2l-1)}{N\sqrt{\frac{\pi}{2}} \left\| f^{l+1} \right\|^2} \right)$ {(27), [BGK10]}

5: $\quad ||f^{(l)}||^2 \leftarrow \frac{(-1)^l}{N^2} \Sigma_{k=1}^N \Sigma_{m=1}^N \Phi^{(2l)} \left(X_k, X_m; 2t_l\right)$ {(26), [BGK10]}

6: $\quad l = l - 1$

7: **end for**

8: $T_p \leftarrow t_l$

9: $p \leftarrow \frac{1}{N} \Sigma_{j=1}^N \Phi^{(2l)} \left(\cdot, X_j; T_p\right)$

10: $a \leftarrow p^\alpha$ $\{\alpha \in [0,1]\}$

11: $||f''||^2 \leftarrow \frac{(-1)^l}{N^2} \Sigma_{k=1}^N \Sigma_{m=1}^N \Phi^{(2l)} \left(X_k, X_m; 2t_l\right)$

12: $L \leftarrow \frac{1}{2} \frac{d}{dx} \left(a \frac{d}{dx} \frac{f}{p} \right)$

13: $||Lf||^2 \leftarrow \frac{1}{N^2} \Sigma_{i=1}^N \Sigma_{j=1}^N L^* L\kappa \left(x, X_i; 2t_l\right) |_{x=X_j}$ {(32), [BGK10]}

14: $T^* \leftarrow \left(\frac{\frac{1}{N} \Sigma_{i=1}^N \sigma^{-1}(X_i)}{2N\sqrt{\pi}||Lf||^2} \right)^{\frac{2}{5}}$ {(23), [BGK10]}

15: smooth the diffusion estimator until T^* as the diffKDE

in Alg. 2. Lines 14 to 16 use the optimal bandwidth T_p to calculate the KDE of the input data in the Fourier space. In lines 17 to 21, this KDE is backtransformed with an inverse Fourier transform and the final two lines provide the re-scaled bandwidth of the Botev KDE.

A further extension of Alg. 3 to Alg. 2 was realized by [DCR11] in Python and can be obtained by the authors upon request. I will refer to this KDE as the Deniz KDE. I received the implementation by Stefan Rotter from the Freiburg University (personal communication) and prepared a comparison on the $\delta^{13}C_{POC}$ data base [VST+21] and artificial data generated from known distributions in Figure 5.3. The known distributions

Algorithm 3: Algorithm for the calculation of the Botev KDE

1: $x_{mesh} \leftarrow range(MIN, MAX, 2^{14})$ {set up equidistant grid}
2: $X_{init} \leftarrow \frac{1}{N} histc(X, x_{mesh})$; $X_{init} \leftarrow \frac{X_{init}}{\sum X_{init}}$ {interpolate X in x_{mesh}}

3: $[n_{rows}, n_{cols}] \leftarrow size(X_{init})$
4: $weight \leftarrow \left[1; 2e^{-i(1:n_{rows}-1)\frac{\pi}{2*n_{rows}}}\right]$
5: $X_{init} \leftarrow [X_{init}(1:2:end,:); X_{init}(end:-2:2,:)]$ {re-order columns}
6: $\tilde{X} \leftarrow real(weight\mathbf{fft}(X_{init})$ {DFT}

7: $I \leftarrow [1:n-1]^2$; $\tilde{X_2} \leftarrow \left(\frac{\tilde{X}(2:end)}{2}\right)^2$
8: $||\widehat{f^{(l+1)}}||^2 \leftarrow 2\pi^{2l} \sum (I^l \tilde{X_2} e^{-I\pi^2 t_l})$
9: **for** $l = 7$ **to** 2 : {fixed-point-Alg. for $t = \zeta * \gamma^{[5]}(t)$} **do**
10: $\quad t_l \leftarrow \left(\frac{1+\left(\frac{1}{2}\right)^{l+\frac{1}{2}}}{3} \frac{1 \times 3 \times \ldots \times (2l-1)}{N\sqrt{\frac{\pi}{2}}||\widehat{f^{(l+1)}}||^2}\right)^{\frac{2}{3+2l}}$
11: $\quad ||\widehat{f^{(l)}}||^2 \leftarrow 2\pi^{2l} \sum \left(I^l \tilde{X_2} e^{-I\pi^2 t_l}\right)$
12: $\quad l \leftarrow l - 1$
13: **end for**
14: $T_p \leftarrow t_l - \left(\frac{1}{2N\sqrt{\pi}||\widehat{f^{(l)}}||^2}\right)^{\frac{2}{5}}$ {return of fixed-point-iteration}
15: {turn T_p into smallest root by Matlab minimizer searcher}
16: $\tilde{X_p} \leftarrow \tilde{X} e^{-[0:n-1]^2 \frac{\pi^2 T_p}{2}}$ {smooth discrete cosine transform of X using T_p}

17: $[n_{rows}, n_{cols}] \leftarrow size\left(\tilde{X_p}\right)$
18: $weights \leftarrow n_{rows} e^{i(0:n_{rows}-1)\frac{\pi}{2n_{rows}}}$
19: temp $\leftarrow real\left(\mathbf{ifft}\left(weights\tilde{X_p}\right)\right)$ {using equation (5.93) in Jain}
20: $\text{temp}_2 = zeros(n_{rows}, 1)$; $\text{temp}_2(1:2:n_{rows}) \leftarrow \text{temp}(1:\frac{n_{rows}}{2})$; $\text{temp}_2(2:2:n_{rows}) \leftarrow \text{temp}(n_{rows}:-1:\frac{n_{rows}}{2+1})$ {re-order elements}
21: KDE $\leftarrow \frac{\text{temp}_2}{MAX-MIN}$ {inverse DFT}

22: $T_p \leftarrow \sqrt{T_p}\,(MAX - MIN)$ {re-scaled bandwidth}
23: {remove negatives (round-off errors) from KDE}

are the same that I have used in the diffKDE paper in Chap. 4. They are a trimodal and a standard lognormal distribution. The comparisons between diffKDE and Botev KDE on these data can already be seen in Chap. 4, therefore it is not dicussed in further detail here. On the $\delta^{13}C_{POC}$ data, the Deniz KDE provides the smoothest estimate. Especially, on the full data set many of data details get smoothed out. On the euphotic zone $\delta^{13}C_{POC}$ data, as well as on the trimodal data, all three KDEs are closely aligned. On the trimodal distribution, the diffKDE and the Deniz KDE detect the main mode most accurately, while the two smaller modes are best resolved by the Botev and the Deniz KDEs. On the lognormal data only the diffKDE resolves the decline left of the mode, which is generally best resolved under the diffKDE. Furthermore, the diffKDE is the one being most sensitive to structure introduced by the choice of the random sample.

5.4 Fourier transform of the diffusion equation

The first author of [BGK10] provided a Matlab implementation of the pilot step of Alg. 2 described in Alg. 3 relying on the Fourier transform [LT05], therefore I will provide a closer look on this measure. The Fourier transform is defined for partial differential equations including L_1-functions and delivers a simple solution for the diffusion equation [LT05] as we defined it in Eq. 3.0.3. Our specifiic case is particularly difficult, because the δ-distribution as part of the initial value in Eq. 3.0.5 is not a L_1-function and therefore needed special consideration before application of the transform.

The Fourier transform is a functional on $L^1([0,1],\mathbb{C})$ mapping functions $f \in L^1([0,1],\mathbb{C})$ to

$$\hat{f}(k) = \int_0^1 f(x)\, e^{-2\pi i k x} dx \text{ fot all } k \in \mathbb{Z} \text{ and } f \in L^1([0,1],\mathbb{C}) \qquad (5.4.1)$$

Its limit as the limit of its symmetric partial sums converges to the input function f [Con16]. The map $f \mapsto \hat{f}$ is called the Fourier transform and denoted as $\mathcal{F}$. The functional $\mathcal{F}$ defines a bijection on the vector space (and algebra) of Schwartz functions, the space of rapid decreasing functions

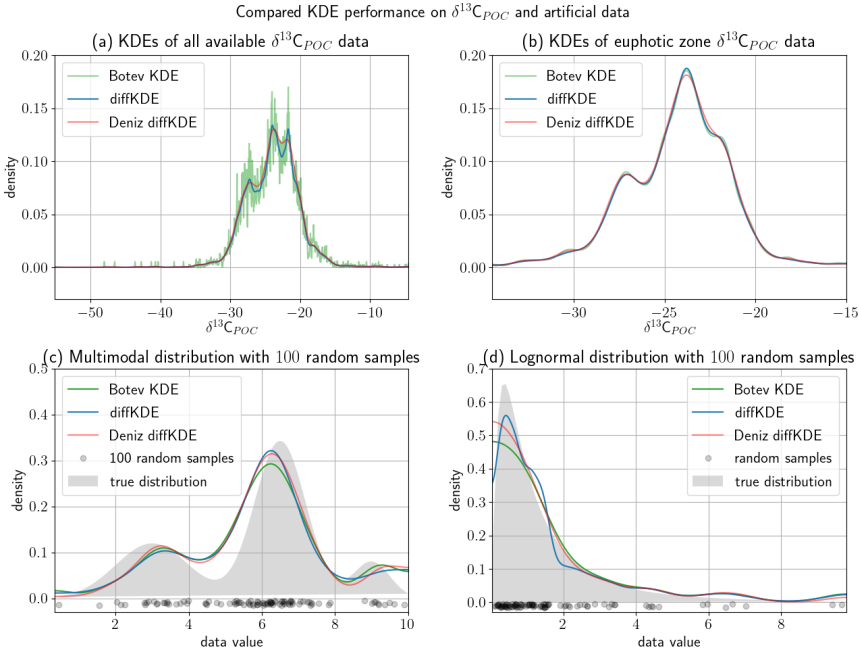

Figure 5.3. The diffKDE in comparison to the pilot by [BGK10] implemented by [Hen21] and full implementation of the [BGK10] algorithm by [DCR11]: (a) and (b) show $\delta^{13}C_{POC}$ data by [VST+21]. (a) shows all available data, (b) only euphotic zone data restricted to a core area of $[-34.5, -15]$. (c) and (d) show the KDEs of 100 random samples from known distributions that are drawn as grey shades in the background.

defined as

$$S\left(\mathbb{R}^m\right) := \{\varphi : \mathbb{R}^m \mapsto \mathbb{C} | \varphi \in C^\infty\left(\mathbb{R}^m\right), sup_{x\in\mathbb{R}^m} | x^\alpha \left(D^\beta \varphi\right)(x) | < \infty\}. \tag{5.4.2}$$

$S\left(\mathbb{R}^m\right)$ is dense in $L^2\left(\mathbb{R}^m\right)$, hence $\mathcal{F}$ can be extended to $L^2\left(\mathbb{R}^m\right)$ [Con16] as

$$\hat{f}\left(\zeta\right) = \int f\left(x\right) e^{-2\pi i \zeta x} dx \text{ for all } \zeta \in \mathbb{R}^m \text{ and } f \in L^2\left(\mathbb{R}^m\right) \tag{5.4.3}$$

with an inverse

$$f(\zeta) = \int \hat{f}(x) e^{2\pi i \zeta x} d\zeta \text{ fot all } x \in \mathbb{R}^m \text{ and } f \in L^2(\mathbb{R}^m) \tag{5.4.4}$$

and so $\mathcal{F}$ is a linear bijection on $L^2(\mathbb{R}^m)$. The applicability to our diffusion follows [Con16], still apart of the initial value.

The δ-distribution, which is neither continuous nor bounded, is neither in S nor in L^2. The δ-distribution is part of the dual space of S, the class of tempered distributions [Con16] defined as

$$S'(\mathbb{R}^m) := \{L : S(\mathbb{R}^m) \mapsto \mathbb{C} | L \, linear \, functional\}. \tag{5.4.5}$$

$S'(\mathbb{R}^m)$ forms a complex vector space and differentiation and convergence definitions can be extended to them [GRT16]. The δ-distribution for any point $a \in \mathbb{R}$ is part of the tempered distributions $S'(\mathbb{R}^m)$ by

$$< \delta_a, \varphi >:= \varphi(a) \text{ for all } \varphi \in S(\mathbb{R}^m) \tag{5.4.6}$$

[GRT16]. The Fourier transform of a tempered distribution $u \in S'(\mathbb{R}^m)$ can be defined by giving its value at a Schwartz function $\varphi \in S(\mathbb{R}^m)$ as

$$< \hat{u}, \varphi >:=< u, \hat{\varphi} > . \tag{5.4.7}$$

This defines a linear bijection from $S'(\mathbb{R}^m)$ on itself and is the unique weakly continuous extension of the Fourier transform of Schwartz functions [Con16]. Now, it is straightforward to calculate the Fourier transform of δ_0 by solving

$$< \hat{\delta_0}, \varphi >=< \delta_0, \hat{\varphi} >= \hat{\varphi}(0) = \int e^{-ix*0} \varphi(x) \, dx = \int \varphi(x) \, dx =< 1, \varphi > \tag{5.4.8}$$

and receiving

$$\widehat{\delta_0} = 1 \tag{5.4.9}$$

as the Fourier transform of the δ-distribution distribution.

Chapter 6

Assessing Earth system models supported by the diffusion-based kernel density estimator

The overall goal of my research was to develop a kernel density estimator for the evaluation of marine data. Achieving that goal can substantially support the calibration of Earth system models. To use common target functions like an Euklidean metric, it is necessary to make field and simulation data of comparable size. A possibility for this, is to only incorporate data from grid cells where both – field and simulation data – are available. By this, a lot of simulation data has to be discarded. Furthermore, spatial biases are highly resolved by a comparison of individual grid cells. The emphasized approach here is to construct the target function not from the data themselves, but from their PDFs. This provides the possibility to incorporate all available data into the calibration process. The construction of a target function from PDFs also disregards the influence of spatial biases and directly resolves differences between the two entire data sets. In this section, I discuss, how the diffKDE can serve for the construction of such a target function that can be used for model calibration. I have shown it to be fast and reliably resolving the typical multimodal data structure in biogeochemical data while being insensitive to noise. A target function can be constructed from diffKDEs of simulation and the respective field data. A metric for PDFs can measure the distance between the two KDEs. Typically, during model calibration this difference is sought to be minimized to obtain optimal model fit. There is a variety of metrics applicable to measure the distance between PDFs and their choice in climate modeling non-trivial [TGG13].

As an example metric, I chose the Wasserstein distance [PZ19]. If X_f and X_s denote the vectors including the field and simulation data, respectively, d_W the Wasserstein distance and diffKDE the here presented diffusion-based KDE, the error I want to discuss in the following is defined as

$$error\left(X_f, X_s\right) := d_W\left(\text{diffKDE}\left(X_f\right), \text{diffKDE}\left(X_s\right)\right) \tag{6.0.1}$$

In the following, I use $\delta^{13}C_{POC}$ field data [VST+21] in comparison with simulation results from [SDW+21] to construct example model-data comparisons. The simulation data is averaged data from the year 2000. The field data is available as decadal averages over either the 1990s or 2000s. All diffKDEs are calculated over the data interval $[-15, -35]$.

In my first example in Fig. 6.1, I show the traditional approach of data reduction to a comparable amount. I applied a mask to the data, that kept only data from locations, where both data types were available and calculated the diffKDEs of both resulting subset data sets. From these diffKDEs, I calculated the error according to Eq. 6.0.1 and drew both diffKDEs together providing a visual insight. This example includes four different data samples: a comparison of all euphotic zone data to the 1990s and 2000s and a split-up of the euphotic zone data into only Southern Ocean data and all oceans excluding the Southern Ocean, both using data from the 1990s.

The general shape of the diffKDEs are well matching between simulation and field data, apart from the averaged 2000s data. In this case, the data had to be reduced to only 39 data points. Furthermore, the 1990s field data are more evenly geographically distributed [VSS+21]. I focus my following discussion on the 1990s field data comparisons. In all graphs it is visible that in general the model tends to overestimate the pronunciation of $\delta^{13}C_{POC}$ modes, especially in the Southern Ocean. Furthermore, the model values are generally lower and all modes estimated at lower $\delta^{13}C_{POC}$ values than observable in the field data.

In my second example in Fig. 6.2, I present a model data comparison built on all available data. The conducted experiments are the exact same apart from not restricting the data to shared grid cells before KDE calculation.

First of all, we see how the 2000s field data is now far more comparable

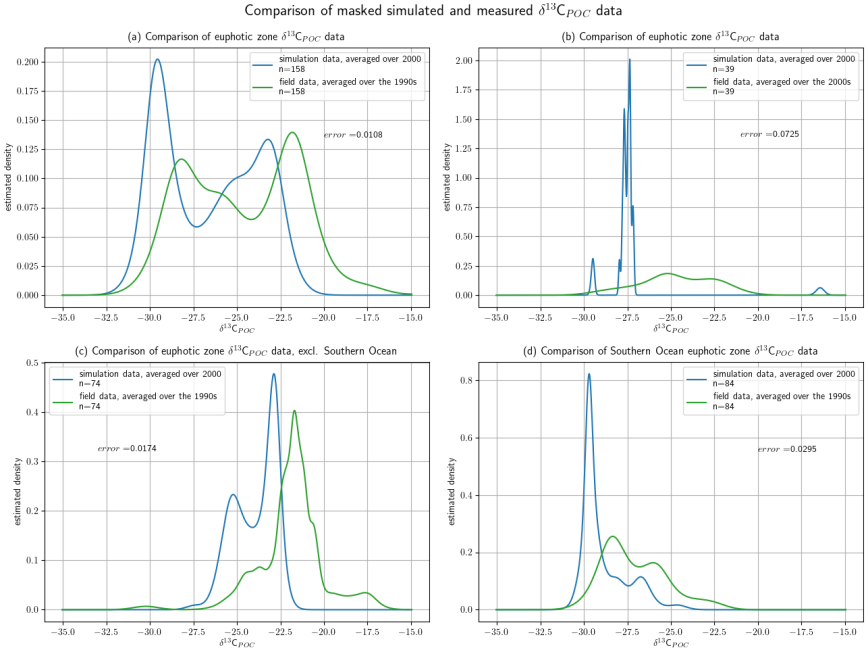

Figure 6.1. Model data comparison on masked data: Simulation and field data are compared using only data points from grid cells, where both data types exist. The simulation data is averaged from the year 2000. The field data are decadal averages from the 1990s in panel (a), (c) and (d) and from the 2000s in panel (b). (a) and (b) show a comparison of the KDEs of the simulation and field data taken from the euphotic zone. (c) shows euphotic zone data excluding the Southern Ocean and (d) euphotic zone data from exclusively the Southern Ocean.

to the year 2000 simulation results. The overall data range is well met. The diffKDE structure is similar with the biggest mode being the one with the highest $\delta^{13}C_{POC}$ values at around $\delta^{13}C_{POC} = -22.5$ ‰in the field data and around $\delta^{13}C_{POC} = -21$ ‰in the simulation data. The lowest mode is also comparable at around $\delta^{13}C_{POC} = -29$ ‰in the field data and around $\delta^{13}C_{POC} = -30$ ‰in the simulation data. In between them, there are three more modes visible in the simulation data and only one in the field data, where none of them seem directly correlated.

The 1990s field data diffKDE even better fits the simulation data diffKDE. There are two main modes detectable in the field data at around $\delta^{13}C_{POC} = -22$ ‰and $\delta^{13}C_{POC} = -28$ ‰that are both located directly in between the two outer modes of the simulation data. The mode at $\delta^{13}C_{POC} = -22$ ‰is more pronounced in the simulation than in the field data. Again, the general data range is well met.

In the exclusion of the Southern Ocean, the main mode well correlates. In the field data this is located at around $\delta^{13}C_{POC} = -22$ ‰in the simulation data this is again split up into two neighboring modes. A lower mode at around $\delta^{13}C_{POC} = -27.5$ ‰in the simulation data is barely detectable in the field data.

The main mode of the Southern Ocean data is still overpronunced in the simulation data and again underestimated in its value. In the field data diffKDE this is located at around $\delta^{13}C_{POC} = -28$ ‰, in the simulation data at around $\delta^{13}C_{POC} = -30$ ‰. A second smaller mode in the field data is located at around $\delta^{13}C_{POC} = -26$ ‰. The simulation data shows four small and similarly pronounced modes close to this.

In all four cases, the incorporation of all available data reduced the error by a magnitude of at least 13.89 % up to 89.38 %. Furthermore, the general data range and locations of modes are far better met.

Overall, I am able to show in this experiment how my diffKDE can highly increase the amount of incorporated data into model data comparisons and provide benefits for model calibration. In the classical masked approach in Fig. 6.1, I had to reduce the simulation and field data to only 39 to 158 data points each. In the unmasked experiment in Fig. 6.2, I included all available data points from simulation results as well as from field data and ended up with 2958 to 12772 data points from simulations and 58 to 172 data points from field measurements. This second approach allows to discard potential influence by local biases and already reveals a smaller error in three out of the four sub-examples. This fourth example is using field data from the 2000s decadal average and the data is nearly exclusively sampled in the Arctic Ocean [VSS+21], which generally complicates drawing useful conclusion from these in global comparisons.

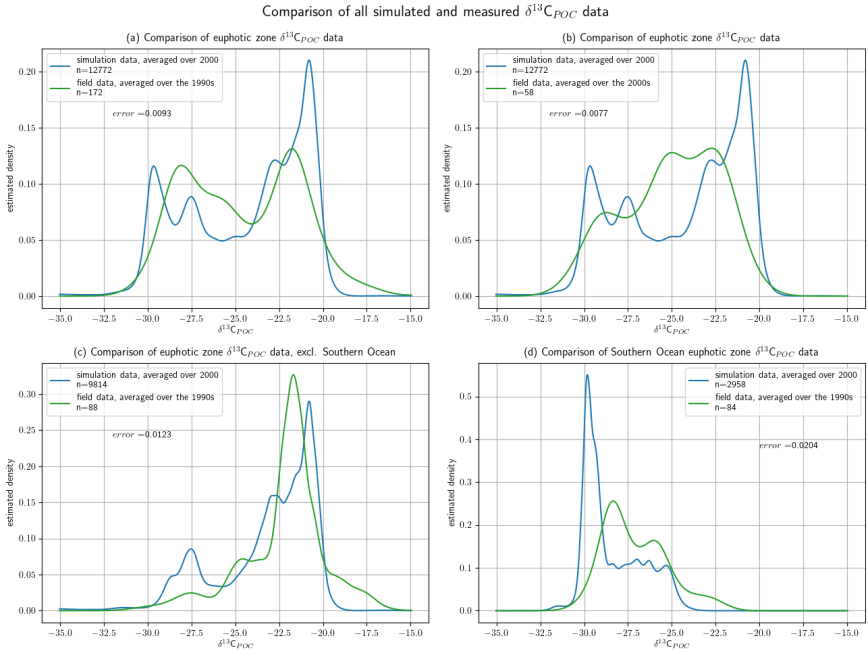

Figure 6.2. Model data comparison on all data: Simulation and field data are compared using all available data points in the respective areas. The simulation data is averaged from the year 2000. The field data are decadal averages from the 1990s in panel (a), (c) and (d) and from the 2000s in panel (b). (a) and (b) show a comparison of the KDEs of the simulation and field data taken from the euphotic zone. (c) shows euphotic zone data excluding the Soutehrn Ocean and (d) euphotic zone data from exclusively the Southern Ocean.

Part III

Conclusion

Chapter 1

Successfully becoming a marine data scientist

Marine data science is an emerging discipline extending data science directly into marine sciences to develop data science measures directly within specific marine research environments. This can improve the understanding of marine data and the ability to draw knowledge from these. Still, marine data science is not an established field on its own. But every scientist conducting research in marine data science is forming the future of this field.

In Tab. 1.1, I present the necessary skills and knowledge a marine data scientist needs to possess according to [VTA+21]. In-depth data science knowledge is especially fundamental for the selection, development, application and improvement of suitable methods. This comprises knowledge of all basic statistics and probability theory, computer science and scientific computing as well as pure and applied mathematics. Knowledge from marine sciences is necessary to understand data origins and their typical features, applications and uncertainties. Marine science knowledge should cover a broad understanding of all parts of marine sciences and their dependencies and interconnections. Skills from marine and data sciences are again needed for development and application of data science tools on marine data as well as for the interpretation of the results. Skills from both parental sciences focus on data handling and programming. Additionally, soft and interface skills are of significant importance for conducting research at the boundary of such different domains. These comprise communication and adaptability, but also a strong sense of oneselfs value as a researcher.

Overall, marine data scientists need to be curious and with a drive to

1. Successfully becoming a marine data scientist

Table 1.1. A marine data scientists necessary skills and knowledges.

	Marine sciences	Data sciences	Soft and interface
Knowledge	physics, geology biology, chemistry society data characteristics	statistics, probability theory algorithms, data bases ML, data mining numerics, optimization differential equations pure mathematics	
Skills	methods, software selection preprocessing transformation pattern mining evaluation	programming version control application of knowledge	boundary communication specific language necessary depth limitations of results questioning entrepreneurial mindset resilience, enthusiasm confidence, determination

constantly extend their knowledge and broaden their perspective. They need to be able to adapt to changing scientific environments, languages and demands. But still, they need also to be able to focus on and identify what is essential for them to solve their individual research questions. By this, marine data scientists are able to develop state of the art solutions and transfer them across scientific domain boundaries.

Chapter 2

A global $\delta^{13}C_{POC}$ data set

I created a new global marine carbon isotope $\delta^{13}C_{POC}$ data set. Up until then, the biggest available global $\delta^{13}C_{POC}$ data set was by [GF94] and comprising only around 500 data samples. I chose the $\delta^{13}C_{POC}$ data as a marine biogeochemical data example to provide them for future model calibration [SDW+21].

Starting from the existing data base by [GF94], I extended the data to a relational data base including different meta-information. In the first data base version [VST+21], I was able to increase the data to 4732 samples. The additional data originates from the data platform PANGAEA and data provided by other marine researchers by personal communication. The meta-information consist of sample time, location, method and original source. From the first data set version I provide interpolations onto two different global grids: one for model future calibration [SDW+21] and one for general data analysis on a well known and widely used grid of the World Ocean Atlas [GWP+18].

I designed the data base for dynamic growth, so that all additional $\delta^{13}C_{POC}$ data can easily be added to extend the data. A second extended version is already available including additional 2220 data points [PST+22]. This version is again available as a spreadsheed file and onto the grid of the World Ocean Atlas.

Along with the first version of the data base, I published a data description paper [VSS+21]. In this, I was able to show that my data base is able to reproduce well known features of $\delta^{13}C_{POC}$ such as the Suess effect as a long term decadal trend. Furthermore, I gave some insights into what kinds of analyses are possible with the provided data and meta-information. These include separation of the data by sample kind and comparison of regions by biomes and sampling method. The second data

base version is already able to support these observations and provides the possibility to explore even better covered geographical and temporal scales.

By being the first well-covered global data base, my $\delta^{13}C_{POC}$ data base can substantially support investigations of the carbon cycle. It can be used to reconstruct pathways of anthropogenic carbon through the Earth system and by this assist to understand human impacts on our climate system. Its ability to dynamically grow makes it well adaptable to insights provided by new data.

Chapter 3

A new approach to a diffusion kernel density estimator for the exploration of marine data

My overall and final marine data science task was the development of a kernel density estimator (KDE) to explore data with typical features of marine data. For this, the KDE must work fast and reliable on large data sets disturbed by different kinds of noise. Typical marine data structures are multimodal and also near boundary, which must be well resolved by the KDE. Noise is introduced numerically in simulations or by measurement errors in field data. A short computation time is useful to provide several evaluations for data analyses and model calibration.

I developed and published a new algorithm for the calculation of a diffusion-based KDE (diffKDE) [PS23; PSS+23]. The idea to use the diffusion equation for a KDE calculation was first proposed by [CM00] and proven to be well suited for multimodal and boundary-close data by [BGK10]. My new algorithm is built on two pilot estimation steps and a direct approximation of the the analytical optimal smoothing parameter by finite differences. The solution of the diffusion equation is discretized by finite differences in space and equidistant timesteps. The temporal solution is approximated by an implicit Euler. The initial value is set to the weighed sum of the δ-distribution of the input data as proposed by [BGK10]. I defined a Dirac sequence on the spatial grid to approximate the δ-distribution.

I tested my implementation on different marine biogeochemical data and artificial data of known distributions in comparison to other state of the art KDEs and approaches to the diffusion KDE. As real marine data,

I used my previously collected $\delta^{13}C_{POC}$ field data [VST+21; PST+22] as well as simulation results [SDW+21] and plankton size data [LNS21]. Overall, the diffKDE is well able to reproduce known multimodal and boundary close distributions and often outperforms other state of the art KDEs on them. It furthermore turned out to be insensitive to noise. The diffKDE shows a similar error convergence rate as the most common Gaussian KDE, but with a generally smaller error. Especially on small data sets with up to a few thousand data points, the diffKDE produces the smallest observed error. On the real marine data, the diffKDE produces detailed data structures, but does not account for individual outliers and smoothes out uncertain data structures. Furthermore, the diffKDE turned out to be faster than previously adapted implementations to the demands of marine biogeochemical data [SLA10]. The software is published as a Python package [PS23].

Finally, I provided an outlook into possibilities for KDE-based calibration of Earth system models. This can be done by calculating the diffKDE of field and simulation data and afterwards using a metric for density comparison such as the Wasserstein distance. The resulting value can be used to determine the model error. Comparing KDEs instead of the data themselves allows to make use of all available data and disregard spatial biases inside the selected observation domain. Furthermore, the visualization of the two KDEs next to each other gives a direct impression into how well the model is able to reproduce specific data values. Building such model calibration on a suitable KDE is crucial to construct reliable Earth system models. My new diffKDE is well able to approximate PDFs of marine data and thus can substantially improve calibration of Earth system models.

Overall, my diffKDE well resolves structures of differently sized and distributed data. This makes it well applicable to marine data, it supports their exploration in general as well as in future specific applications like model calibration. Beyond this, my diffKDE can provide substantial benefits for general data exploration in many other research fields dealing with data of unknown distributions.

Bibliography

[ABH+15] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. "The FEniCS project version 1.5". In: *Archive of Numerical Software* 3 (2015). DOI: 10.11588/ans.2015.100.20553.

[Abr82] Ian S Abramson. "On bandwidth variation in kernel estimates-a square root law". In: *The annals of Statistics* (1982), pp. 1217–1223.

[Alf] Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (AWI) and Center for Marine Environmental Sciences, University of Bremen (MARUM). *PANGAEA. Data Publisher for Earth & Environmental Science*. URL: https://www.pangaea.de (visited on 11/16/2022).

[BGK10] Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. "Kernel density estimation via diffusion". In: *The annals of Statistics* 38.5 (2010), pp. 2916–2957. DOI: 10.1214/10-AOS799.

[BMP77] Leo Breiman, William Meisel, and Edward Purcell. "Variable kernel estimates of multivariate densities". In: *Technometrics* 19.2 (1977), pp. 135–144.

[BWR+17] Benjamin Bronselaer, Michael Winton, Joellen Russell, Christopher L. Sabine, and Samar Khatiwala. "Agreement of CMIP5 simulated and observed ocean anthropogenic CO_2 uptake". In: *Geophysical Research Letters* 44.24 (Dec. 2017). DOI: 10.1002/2017gl074435. URL: https://doi.org/10.1002/2017gl074435.

[CH20] Hilary G. Close and Lillian C. Henderson. "Open-ocean minima in δ13c values of particulate organic carbon in the lower euphotic zone". In: *Frontiers in Marine Science* 7 (Sept. 2020). DOI: 10.3389/fmars.2020.540165. URL: https://doi.org/10.3389/fmars.2020.540165.

[CM00] P Chaudhuri and JS Marron. "Scale space view of curve estimation". In: *ANNALS OF STATISTICS* 28.2 (Apr. 2000), pp. 408–428. ISSN: 0090-5364. DOI: 10.1214/aos/1016218224.

[Con16] Adrian Constantin. *Fourier Analysis*. London Mathematical Society, 2016. ISBN: 978-1-107-62035-3.

[DCR11] T Deniz, S Cardanobile, and S Rotter. "A PYTHON package for kernel smoothing via diffusion: estimation of spike train firing rate". In: *Front. Comput. Neurosci. Conference Abstract: BC11 : Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting 2011* 5 (2011). DOI: 10.3389/conf.fncom.2011.53.00071. URL: https://doi.org/10.3389/conf.fncom.2011.53.00071.

[EPH+19] B Espinasse, EA Pakhomov, BPV Hunt, and SJ Bury. "Latitudinal gradient consistency in carbon and nitrogen stable isotopes of particulate organic matter in the southern ocean". In: *Marine Ecology Progress Series* 631 (Nov. 2019), pp. 19–30. DOI: 10.3354/meps13137. URL: https://doi.org/10.3354/meps13137.

[GEH+21] Katie St John Glew et al. "Isoscape models of the southern ocean: predicting spatial and temporal variability in carbon and nitrogen isotope compositions of particulate organic matter". In: *Global Biogeochemical Cycles* 35.9 (Sept. 2021). DOI: 10.1029/2020gb006901. URL: https://doi.org/10.1029/2020gb006901.

[GF94] Ralf Göricke and Brian Fry. "Variations of marine plankton δ^{13}c with latitude, temperature, and dissolved co_2 in the world ocean". In: *Global Biogeochemical Cycles* 8.1 (Mar. 1994), pp. 85–90. DOI: 10.1029/93gb03272. URL: https://doi.org/10.1029/93gb03272.

[GRT16] Alfred Göpert, Thomas Riedrich, and Christiane Tammer. *Approximation und Nichtlineare Optimierung in Praxisaufgaben*. Springer, 2016. ISBN: 978-3-658-14760-0.

[GVB+22] Ralf Gommers et al. *Scipy/scipy: scipy 1.8.0*. Version v1.8.0. Feb. 2022. DOI: 10.5281/zenodo.5979747. URL: https://doi.org/10.5281/zenodo.5979747.

[GWP+18] H. E. Garcia et al. "Dissolved inorganic nutrients (phosphate, nitrate and nitrate+nitrite, silicate)". In: *World Ocean Atlas 2018* 4 (2018). NOAA ATLAS NESDIS 84, 35pp.

[Hay04] John M Hayes. "An introduction to isotopic calculations". In: (2004). URL: http://www.whoi.edu/cms/files/jhayes/2005%20/9/IsoCalcs30Sept04_5183.pdf.

[Hen21] John Hennig. *John-hennig/kde-diffusion: kde-diffusion 1.0.3*. Version v1.0.3. Apr. 2021. DOI: 10.5281/zenodo.4663430. URL: https://doi.org/10.5281/zenodo.4663430.

[HMW+20] Charles R. Harris et al. *Array programming with NumPy*. 2020. DOI: 10.1038/s41586-020-2649-2.

[HRA+14] F. M. Hoffman et al. "Causes and implications of persistent atmospheric carbon dioxide biases in earth system models". In: *Journal of Geophysical Research: Biogeosciences* 119.2 (Feb. 2014), pp. 141–162. DOI: 10.1002/2013jg002381. URL: https://doi.org/10.1002/2013jg002381.

[Hun07] John D Hunter. "Matplotlib: a 2d graphics environment". In: *Computing in science & engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/mcse.2007.55. URL: https://doi.org/10.1109/mcse.2007.55.

[Inc20] Anaconda Inc. *Anaconda software distribution*. Version Vers. 2-2.4.0. 2020. URL: https://docs.anaconda.com/.

[IPC13] IPCC. "Summary for policymakers". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley. Cambridge, UK: Cambridge University Press, 2013, pp. 3–29. DOI: 10.1017/CBO9781107415324.004.

[IPC14] IPCC. *Climate change 2014: synthesis report. contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change [core writing team, r.k. pachauri and l.a. meyer (eds.)]* Ed. by Allen et al. Geneva, Switzerland, Jan. 2014, 151 pp.

[Irl10] A. Irle. *Wahrscheinlichkeitstheorie und Statistik: Grundlagen - Resultate - Anwendun- gen. 2. Auflage*. Wiesbaden: Springer, 2010. ISBN: 978-3-519-12395-8.

[ISS+13] Tatiana Ilyina, Katharina D. Six, Joachim Segschneider, Ernst Maier-Reimer, Hongmei Li, and Ismael Núñez-Riboni. "Global ocean biogeochemistry model HAMOCC: model architecture and performance as component of the MPI-earth system model in different CMIP5 experimental realizations". In: *Journal of Advances in Modeling Earth Systems* 5.2 (May 2013), pp. 287–315. DOI: 10.1029/2012ms000178. URL: https://doi.org/10.1029/2012ms000178.

[LL16] Hans Petter Langtangen and Anders Logg. "A gallery of finite element solvers". In: *Solving PDEs in Python*. Springer International Publishing, 2016, pp. 37–81. DOI: 10.1007/978-3-319-52462-7_3. URL: https://doi.org/10.1007/978-3-319-52462-7_3.

[LMW12] Anders Logg, Kent-Andre Mardal, and Garth Wells, eds. *Automated solution of differential equations by the finite element method*. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-23099-8. URL: https://doi.org/10.1007/978-3-642-23099-8.

[LNS21] Vanessa Lampe, Eva-Maria Nöthig, and Markus Schartau. "Spatio-temporal variations in community size structure of arctic protist plankton in the fram strait". In: *Frontiers in Marine Science* 7 (Jan. 2021). DOI: 10.3389/fmars.2020.579880. URL: https://doi.org/10.3389/fmars.2020.579880.

[LPC+19] Anne Lorrain et al. "Trends in tuna carbon isotopes suggest global changes in pelagic phytoplankton communities". In: *Global Change Biology* 26.2 (Nov. 2019), pp. 458–470. DOI: 10.1111/gcb.14858. URL: https://doi.org/10.1111/gcb.14858.

[LT05] Stig Larsson and Vidar Thomée. *Partielle Differentialgleichungen und numerische Methoden*. Springer, 2005. ISBN: 3-540-01772-0.

[LW10] Anders Logg and Garth N. Wells. "DOLFIN". In: *ACM Transactions on Mathematical Software* 37.2 (Apr. 2010), pp. 1–28. DOI: 10.1145/1731022.1731030. URL: https://doi.org/10.1145/1731022.1731030.

[MBN16] Greg McSwiggan, Adrian Baddeley, and Gopalan Nair. "Kernel density estimation on a linear network". In: *Scandinavian Journal of Statistics* 44.2 (Nov. 2016), pp. 324–345. DOI: 10.1111/sjos.12255. URL: https://doi.org/10.1111/sjos.12255.

[Mer14] Dirk Merkel. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux journal* 2014.239 (2014), p. 2.

[MR94] James Stephen Marron and David Ruppert. "Transformations to reduce boundary bias in kernel density estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.4 (1994), pp. 653–671. URL: https://www.jstor.org/stable/2346189.

[NMK+16] C. D. Nevison et al. "Evaluating CMIP5 ocean biogeochemistry and southern ocean carbon uptake using atmospheric potential oxygen: present-day performance and future projection". In: *Geophysical Research Letters* 43.5 (Mar. 2016), pp. 2077–2085. DOI: 10.1002/2015gl067584. URL: https://doi.org/10.1002/2015gl067584.

[OCK+22] Sophy Oliver, Coralia Cartis, Iris Kriest, Simon F. B Tett, and Samar Khatiwala. "A derivative-free optimisation method for global ocean biogeochemical models". In: *Geoscientific Model Development* 15.9 (May 2022), pp. 3537–3554. DOI: 10.5194/gmd-15-3537-2022. URL: https://doi.org/10.5194/gmd-15-3537-2022.

[Par62] Emanuel Parzen. "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[PS23] Maria-Theresia Pelz and Thomas Slawig. *Diffusion-based kernel density estimator (diffkde)*. en. 2023. DOI: 10.5281/ZENODO.7594915. URL: https://zenodo.org/record/7594915.

[PSS+23] Maria-Theresia Pelz, Markus Schartau, Christopher J. Somes, Vanessa Lampe, and Thomas Slawig. "A diffusion-based kernel density estimator (diffkde, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research". In: *Geosci. Model Dev. Dis-*

cuss. [preprint] (2023). in review. DOI: 10.5194/gmd-2023-17. URL: https://gmd.copernicus.org/preprints/gmd-2023-17/.

[PST+22] Maria-Theresia Pelz, Christopher J Somes, Robyn E Tuerena, Anne Lorrain, Hilary G Close, Lillian C Henderson, Katie St John Glew, Boris Espinasse, and Clive N Trueman. *A global marine particulate organic carbon-13 isotope data product (version2)*. en. 2022. DOI: 10.1594/PANGAEA.946915. URL: https://doi.pangaea.de/10.1594/PANGAEA.946915.

[PZ19] Victor M. Panaretos and Yoav Zemel. "Statistical aspects of wasserstein distances". In: *Annual Review of Statistics and Its Application* 6.1 (Mar. 2019), pp. 405–431. DOI: 10.1146/annurev-statistics-030718-104938. URL: https://doi.org/10.1146/annurev-statistics-030718-104938.

[RW86] J S Rounick and M J Winterbourn. *Stable carbon isotopes and carbon flow in ecosystems - measuring ^{13}c to ^{12}c ratios can help to trace carbon pathways*. Mar. 1986. DOI: 10.2307/1310304.

[Sco12] David W Scott. "Multivariate density estimation and visualization". In: *Handbook of computational statistics*. Springer, 2012, pp. 549–569. DOI: 10.1007/978-3-642-21551-3_19.

[SDW+21] Christopher J. Somes, Andrew W. Dale, Klaus Wallmann, Florian Scholz, Wanxuan Yao, Andreas Oschlies, Juan Muglia, Andreas Schmittner, and Eric P. Achterberg. "Constraining global marine iron sources and ligand-mediated scavenging fluxes with GEOTRACES dissolved iron measurements in an ocean biogeochemical model". In: *Global Biogeochemical Cycles* 35.8 (Aug. 2021). DOI: 10.1029/2021gb006948. URL: https://doi.org/10.1029/2021gb006948.

[SGC+13] A Schmittner, N Gruber, A C, Mix R M Key, A Tagliabue, and T K Westberry. "Biology and air-sea gas exchange controls on the distribution of carbon isotope ratios (^{13}c) in the ocean". In: *Biogeosciences* (2013), pp. 5793–5816. DOI: 10.5194/bg-10-5793-2013.

[She04] Simon J. Sheather. "Density estimation". In: *Statistical Science* 19.4 (Nov. 2004), pp. 588–597. DOI: 10.1214/088342304000000297. URL: https://doi.org/10.1214/088342304000000297.

[Sil86] BW Silverman. "Density estimation". In: *Monographs on Statistics and Applied Probability* (1986).

[SJ91] Simon J Sheather and Michael C Jones. "A reliable data-based bandwidth selection method for kernel density estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 683–690.

[SLA10] Markus Schartau, Michael R. Landry, and Robert A. Armstrong. "Density estimation of plankton size spectra: a reanalysis of IronEx II data". In: *Journal of Plankton Research* 32.8 (Aug. 2010). ISBN: 0142-7873, pp. 1167–1184. ISSN: 1464-3774. DOI: 10.1093/plankt/fbq072. URL: https://academic.oup.com/plankt/article-lookup/doi/10.1093/plankt/fbq072.

[Sla15] Thomas Slawig. *Klimamodelle und Klimasimulationen*. Berlin Heidelberg: Springer Spektrum, 2015. ISBN: 978-3-662-47063-3. DOI: 10.1007/978-3-662-47064-0.

[SS16] A Schmittner and C J Somes. "Complementary constraints from carbon (^{13}c) and nitrogen (^{15}n) isotopes on the glacial ocean's soft-tissue biological pump". In: *Paleoceanography* (2016), pp. 669–693. DOI: 10.1002/2015PA002905.

[TB08] Alessandro Tagliabue and Laurent Bopp. "Towards understanding global variability in ocean carbon-13". In: *Global Biogeochemical Cycles* 22.1 (Mar. 2008), n/a–n/a. DOI: 10.1029/2007gb003037. URL: https://doi.org/10.1029/2007gb003037.

[TGG13] Thordis L. Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. "Using proper divergence functions to evaluate climate models". In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1 (Jan. 2013), pp. 522–534. DOI: 10.1137/130907550. URL: https://doi.org/10.1137/130907550.

[TGH+19] Robyn E. Tuerena, Raja S. Ganeshram, Matthew P. Humphreys, Thomas J. Browning, Heather Bouman, and Alexander P. Piotrowski. "Isotopic fractionation of carbon during uptake by phytoplankton across the south atlantic subtropical convergence". In: *Biogeosciences* 16.18 (Sept. 2019), pp. 3621–3635. DOI: 10.5194/bg-16-3621-2019. URL: https://doi.org/10.5194/bg-16-3621-2019.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009. ISBN: 978-0-378-79051-0. DOI: 10.1007/978-0-378-79052-7.

[Van20] Guido Van Rossum. *The python library reference, release 3.8.2*. Python Software Foundation, 2020.

[VD09] Guido Van Rossum and Fred L. Drake. *Python 3 reference manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[VSS+21] Maria-Theresia Verwega, Christopher J. Somes, Markus Schartau, Robyn Elizabeth Tuerena, Anne Lorrain, Andreas Oschlies, and Thomas Slawig. "Description of a global marine particulate organic carbon-13 isotope data set". In: *Earth System Science Data* 13.10 (Oct. 2021), pp. 4861–4880. DOI: 10.5194/essd-13-4861-2021. URL: https://doi.org/10.5194/essd-13-4861-2021.

[VST+21] Maria-Theresia Verwega, Christopher J Somes, Robyn E Tuerena, and Anne Lorrain. *A global marine particulate organic carbon-13 isotope data product*. data set. 2021. DOI: 10.1594/PANGAEA.929931. URL: https://doi.org/10.1594/PANGAEA.929931.

[VTA+21] Maria-Theresia Verwega et al. "Perspectives on marine data science as a blueprint for emerging data science disciplines". In: *Frontiers in Marine Science* 8 (Dec. 2021). DOI: 10.3389/fmars.2021.678404. URL: https://doi.org/10.3389/fmars.2021.678404.

[WCW+99] Jinping Wu, S.E. Calvert, C.S. Wong, and F.A. Whitney. "Carbon and nitrogen isotopic composition of sedimenting particulate material at station papa in the subarctic northeast pacific". In: *Deep Sea Research Part II: Topical Studies in Oceanog-*

raphy 46.11-12 (Nov. 1999), pp. 2793–2832. DOI: 10.1016/s0967-0645(99)00084-3. URL: https://doi.org/10.1016/s0967-0645(99)00084-3.

[ZZC+14] Run Zhang, Minfang Zheng, Min Chen, Qiang Ma, Jianping Cao, and Yusheng Qiu. "An isotopic perspective on the correlation of surface ocean carbon dynamics and sea ice melting in prydz bay (antarctica) during austral summer". In: *Deep Sea Research Part I: Oceanographic Research Papers* 83 (Jan. 2014), pp. 24–33. DOI: 10.1016/j.dsr.2013.08.006. URL: https://doi.org/10.1016/j.dsr.2013.08.006.