Marine Data Fusion for Analyzing Spatio-Temporal Ocean Region Connectivity

Carola Trahms

geboren in Berlin

Dissertation
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel
eingereicht im Jahr 2022

Kiel Computer Science Series (KCSS) 2023/3 dated 2023-06-22

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 mail@carola-trahms.de

Published by the Department of Computer Science, Kiel University

Archaeoinformatics – Data Science

Please cite as:

▻ Carola Trahms. *Marine Data Fusion for Analyzing Spatio-Temporal Ocean Region* Number 2023/3 in Kiel Computer Science Series. Department of Computer Science, 2023. Dissertation, Faculty of Engineering, Kiel University.

```
@book{Trahms2023,
  author    = {Carola Trahms},
  title     = {Marine Data Fusion for
               Analyzing Spatio-Temporal Ocean Region Connectivity},
  publisher = {Department of Computer Science, Kiel University},
  year      = {2023},
  number    = {2023/3},
  doi       = {10.21941/kcss/2023/3},
  series    = {Kiel Computer Science Series},
  note      = {Dissertation, Faculty of Engineering,
               Kiel University.}
}
```

© 2023 by Carola Trahms. The author has obtained the right to include the published and accepted articles in the thesis, with a condition that they are cited and/or copyright/credits are placed prominently in the references.

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. Matthias Renz Christian-Albrechts-Universität Kiel
2. Gutachter:	Prof. Dr. Martin Visbeck Christian-Albrechts-Universität Kiel

Datum der mündlichen Prüfung: 07.12.2022

Zusammenfassung

Die vorliegende Arbeit entwickelt Methoden um die Konnektivitätsanalyse zwischen Meeresregionen zu automatisieren und zu objektivieren.

Die bisherigen Methoden zur Konnektivitätsanalyse stützen sich oft auf manuelles Einflechten von Expertenwissen, was die Verarbeitung großer Datenmengen langwierig gestaltet.

Diese Arbeit stellt ein neues Framework zur Datenfusion vor, welches viele Automatisierungs- und Objektivierungsansätze für den gesamten Analyseprozess bereit hält. Sie identifiziert verschiedene Komplexitäten der Konnektivitätsanalyse und zeigt, wie sich das Datenfusionsframework darauf anpassen und anwenden lässt.

Das Framework wird in dieser Arbeit verwendet um geo-referenzierte Trajektorien von Fischlarven im westlichen Mittelmeer zu analysieren, die Ausbreitungspfade neu gebildeten Wassers im subpolaren Nordatlantik anhand ihrer hydrographischen Eigenschaften zu verfolgen und ihre zeitliche Änderung zu erfassen.

Diese Beispiele geben den verwendeten und neu kombinierten informationstechnischen Methoden ein neues und im Zuge des fortschreitenden Klimawandels hoch relevantes Anwendungsfeld. Für die Weiterentwicklung dieser Methoden ergeben sich damit neue Richtungen, die vor dem Hintergrund der Meereswissenschaften über das Optimieren vorhandener Methoden hinaus gehen. Die Meereswissenschaft, genauer die Physikalische Ozeanographie, profitiert von den neuen Möglichkeiten große Datenmengen schnell und objektiv auf ihre exakten Fragestellungen hin zu analysieren.

Diese Arbeit ist ein Vorstoß in das neue Gebiet der Marinen Datenwissenschaften. Sie erforscht praktisch und theoretisch die Möglichkeiten Daten- und Meereswissenschaften gewinnbringend für beide Seiten zu vereinen. Sie zeigt beispielhaft anhand der Konnektivitätsanalyse zwischen Meeresregionen den Mehrwert der Kombination der Daten- und Meereswissenschaften. Außerdem zeigt sie erste Erkenntnisse und Ideen auf, wie

sich Forschende aus beiden Disziplinen aufstellen können um erfolgreiche Marine Datenwissenschaft zu betreiben und damit das Verständnis des Ozeans voran zu treiben.

Abstract

This thesis develops methods to automate and objectify the connectivity analysis between ocean regions.

Existing methods for connectivity analysis often rely on manual integration of expert knowledge, which renders the processing of large amounts of data tedious.

This thesis presents a new framework for Data Fusion that provides several approaches for automation and objectification of the entire analysis process. It identifies different complexities of connectivity analysis and shows how the Data Fusion framework can be applied and adapted to them.

The framework is used in this thesis to analyze geo-referenced trajectories of fish larvae in the western Mediterranean Sea, to trace the spreading pathways of newly formed water in the subpolar North Atlantic based on their hydrographic properties, and to gauge their temporal change.

These examples introduce a new, and - as climate change progresses - highly relevant field of application for the established Data Science methods that were used and innovatively combined in the framework. New directions for further development of these methods are opened up, which, against the background of Marine Science, go beyond optimization of existing methods. The Marine Science, more precisely Physical Oceanography, benefits from the new possibilities to analyze large amounts of data quickly and objectively for its exact research questions.

This thesis is a foray into the new field of Marine Data Science. It practically and theoretically explores the possibilities of combining Data Science and the Marine Sciences advantageously for both sides. The example of automating and objectifying connectivity analysis between marine regions in this thesis shows the added value of combining Data Science and Marine Science. This thesis also presents initial insights and ideas on how researchers from both disciplines can position themselves to thrive as Marine Data Scientists and simultaneously advance our understanding of the ocean.

Acknowledgements

It takes a village to raise a PhD.

— approximate English proverb

I want to thank my village for supporting me not only scientifically, but also mentally. Without your patience and constant encouragement this thesis would have never existed. Thank you all!

Contents

List of Figures

Part I

Thesis Summary

Chapter 1

Introduction

A changing climate affects the Earth and with it the global ocean. As consequence of human activity and the respective emission of greenhouse gas to the atmosphere, the Earth's climate system is changing. This climate emergency manifests not only in the atmosphere but also in the global ocean [MZP+21]. The ocean has a high heat capacity in comparison to the atmosphere. Thus, the hydrographic structure and with it the ocean dynamics might change slower than the atmosphere. Changes in the dynamics of a warmer ocean might only start to show now.

To observe and understand the global ocean and its changes the Marine Sciences collect and produce huge amounts of data every year. These data are spatially and temporally sparse and they are of very different type and accuracy [VTA+21]. Handling and making collective sense of these huge amounts of heterogeneous data is not part of the traditional education of Marine Scientists. Data Science, an emerging branch of the Computer Sciences, aims at developing and using methods to be able to handle these large data sets and infer relevant insights from them.

Combining Data Science and Marine Sciences promises to propel both sciences forward by learning from each other. Integrating Data Science methods into the marine analytic workflow has the potential to help tremendously in interpreting and generating objective insights. The Data Science community, in turn, benefits from the introduction of novel real world (i.e., messy) data sets for testing existing and established methods. As a consequence, new algorithms may emerge that aim at tackling the details where a mere combination of state-of-the-art methods falls short. Having worked at this exciting interface between Data and Marine Science for three years, we provide hands-on insights into synergistic effects between the sciences that occur on one hand, and pitfalls where the differ-

ences in terminology and publication culture create a lot of friction and loss of momentum on the other hand.

In this thesis we developed a Data Fusion framework that integrates deep knowledge and understanding of the various areas of research inside Data and Computer Science and applied it on connectivity analysis problems stemming from Physical Oceanography, a branch of Marine Sciences. We investigated connectivity research questions of increasing complexity and specifically tailored our approaches to solve the exact domain science problems. We concentrated on automating and objectifying the analytical methods for connectivity analysis between ocean regions with simulated Lagrangian Particle Release Experiments. Lagrangian Particle Release Experiments are a standard method from Physical Oceanography to generate data for connectivity analysis. Understanding how different ocean regions are connected by the underlying dynamics is a current topic of research in Physical Oceanography, e.g., in the subpolar North Atlantic [PTM+17; FKO+18; HVK21].

1.1 Thesis Structure

This thesis is divided into two parts. Part I contains four chapters and gives an overview over the research that has been done during this thesis. The papers and manuscript written during this work are attached in Part II.

In the remainder of this chapter we introduce how we define the increasing complexity of different aspects of connectivity analysis (Section 1.2).

Chapter 2 gives a short introduction to connectivity analysis in Physical Oceanography for readers stemming from Data or Computer Science with no background knowledge on Marine Sciences, Physical Oceanography and their methods for connectivity analysis.

Chapter 3 briefly summarizes the contributions and further implications of the research in this work.

Chapter 4 concludes the findings stemming from this interdisciplinary research for both Data and Marine Science and gives an outlook into further work.

1.2 Dimensions of Complexity in Connectivity Analysis

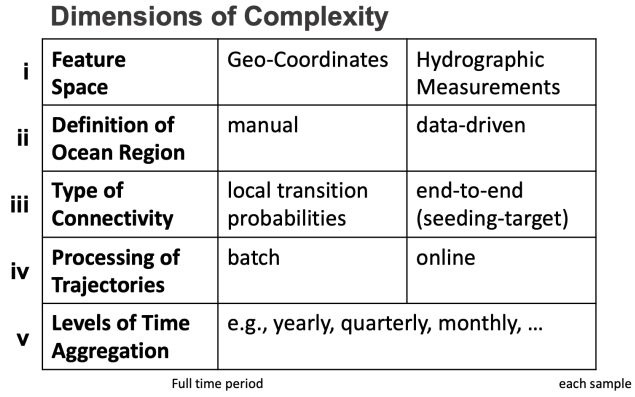

Figure 1.1. Overview on the dimensions of complexity in connectivity between ocean regions. The complexity depends on what research questions are asked and what kind of data are available. From a data methodical point of view, there are five main dimensions influencing the complexity: i) feature space ii) definition method of the ocean regions iii) type of connectivity iv) way of processing the trajectories v) level of time aggregation.

We identified five dimensions that influence the complexity for machine learning aided connectivity analysis between ocean regions: i) The feature space that describes the regions, ii) the way of defining an ocean region, iii) the way of defining what connectivity means, iv) the way how the trajectories are processed, and v) the aggregation levels with regard to the time span. Figure 1.1 shows a visual overview of the complexity dimensions and their expressions investigated in this thesis. The combination of the different dimensions leads to configurations that have different interacting levels of complexity. The right hand expressions of each dimension are either more complex on their own for each dimension, e.g., online segmentation of trajectories is more complex than batch segmentation, or combining both expressions leads to more complex problems, e.g., com-

bining geo-references with hydrographic measures is more complex than each on their own. We developed approaches for increasingly complex configurations of connectivity analysis. We use this schema to introduce the research contributions of this work in Chapter 3.

Feature Space Ocean regions can be described through spatial information, i.e. geo-coordinates or hydrographic information, such as measurements of temperature and salinity.

The complexity of the feature space depends on how much spatial and hydrographic information is recorded in the trajectories of the particles. The minimal complexity contains merely the longitude and latitude as geo-coordinates along with the timestamp. The next level of complexity – especially when trying to generate a probability density map – is to add the depth as an additional geo-coordinate. In even higher complexity, hydrographic measurements are recorded along the way.

Definition of Ocean Regions The ocean regions can be defined manually in an expert guided manner or automatically in a data driven way, e.g., through clustering.

The complexity of defining ocean regions is an interaction between the feature space that is used and the modus operandi (manual vs. data-driven). One of the lowest complexities for connectivity analysis is to have no hydrographic and depth information in the trajectories while the ocean regions of interest are few and manually defined (see section 3.2). When tracing three-dimensional movement in many ocean regions that are defined by their hydrographic properties, the complexity of the analysis increases (see sections 3.3 and 3.4).

Type of Connectivity The type of connectivity, whether it is end-to-end from a seeding area to a designated target area or whether the local connectivities underway are regarded.

The complexity in the types of connectivity follows the length of pathways. For the local connectivity, short pathways need to be identified and evaluated. For the end-to-end connectivity, the whole pathway and all possible variations and detours need to be considered. These longer

pathways are much harder to find and the run times of the detecting algorithms are also much increased. When they rely on distances between data points or trajectory segments, the distance matrices are huge, i.e., computatively and memorywise very costly.

Processing of Trajectories Trajectories can be processed as a complete construct (batch) or regarded as a time series where the next position is fed online to the analysis.

Depending on the availability of the data source used in the research question, this influences the possible methods massively. When using simulated data, usually there is no need to treat trajectories in an online manner. There is no real time constraint, so the data can be processed in a batch. When using real measurement data, such as the readings of Argo floats [Arg22], it might be more sensible to apply an online approach. In this work we used simulation data and treated them as batch. Recent methods from Physical Oceanography, however, treat the trajectories as online sequences [FBC+22].

Levels of Time Aggregation Oceanic processes happen on vastly different time spans [VTA+21].

For the complexity of the connectivity analysis it has a huge impact whether the data is aggregated over years to highlight differences or trends in certain years or decades, or if the aggregation is done over quarters or months to observe trends on shorter time scales – such as seasonal influences. A special case is the aggregation of all quarters or months over the course of several years as we introduced in Section 3.4. This emphasizes even further the seasonal or inter-quartal differences in a certain set of years. The smaller the aggregation level, the more data needs to be analyzed and interpreted. To find a way or measure to efficiently compare on these different time scales helps mitigate this complexity.

1.3 List of Publications

The following papers and presentations were assembled during this thesis.

Papers

Trahms, C., Handmann, P., Visbeck, M. and Renz, M. (Manuscript). **Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions**.

Trahms, C., Wölker, Y., Handmann, P., Visbeck, M. and Renz, M. (in print 2022). **Data Fusion for Connectivity Analysis between Ocean Regions**. In (eScience 2022).

Tavares de Sousa, N., **Trahms, C.**, Kröger, P., Renz, M., Schubert, R., Biastoch, A. (2022). **Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters**. The 23rd IEEE International Conference on Mobile Data Management (MDM 2022).

Trahms, C., Handmann, P., Rath, W., Visbeck, M. and Renz, M. (2021). **Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories**. In 17th International Symposium on Spatial and Temporal Databases (SSTD 2021).

Verwega, M.T. and **Trahms, C.**, Antia, A.N., Dickhaus, T., Prigge, E., Prinzler, M.H.U., Renz, M., Schartau, M., Slawig, T., Somes, C.J. and Biastoch, A. (2021) **Perspectives on Marine Data Science as a Blueprint for Emerging Data Science Disciplines. Frontiers in Marine Science**, 8:678404. https://doi.org/10.3389/fmars.2021.678404.

Abstracts and Presentations

Trahms, C., Handmann, P., Rath, W., Renz, M., and Visbeck, M. (2022): **Autonomous Assessment of Source Area Distributions for Sections in Lagrangian Particle Release Experiments**, EGU General Assembly (EGU 2022). https://doi.org/10.5194/egusphere-egu22-5631.

Trahms, C., Handmann, P., Renz, M., and Visbeck, M. (2022) **Grid-free detection of relevant moving directions in marine particle trajectories using state-of-the-art clustering**, Ocean Sciences Meeting, OSM 2022

Rath, W., **Trahms, C.**, Handmann, P., Wölker, Y. (2022): **Unsupervised learning on large collections of high-resolution trajectories**, 7th Data Science Symposium, Hereon 2022

Trahms, C., Handmann, P., Rath, W., Renz, M., and Visbeck, M. (2021) **Efficient Pathway Identification from Geospatial Trajectories, European Geosciences Union**, EGU 2021

Trahms, C., Renz, M., and Visbeck, M. (2021) **A Whole New World: Leveraging the Power of Data Links with Heterogeneous Information Networks**, 5th Data Science Symposium, Geomar 2021

Chapter 2

Connectivity Analysis between Ocean Regions

πάντα ῥεῖ - *everything flows*

— Heraklit

This chapter gives a brief overview over the terminology and standard methods used in Physical Oceanography and connectivity analysis. It is meant for Data and Computer Scientists that are interested in diving deeper into the marine research behind the methods that are presented in this thesis.

Physical Oceanography investigates and strives to explain ocean dynamics. For this, data on hydrographic and dynamic properties are collected by observing the ocean through in-situ measurements, satellite observations, and by simulating the ocean or parts of it in Ocean General Circulation Models (OGCMs). Hydrographic properties are physical descriptions of the water such as temperature, salinity and density. Dynamic properties include longitudinal and latitudinal flow velocities.

Connectivity analysis focuses on investigating how the water moves. This is studied by releasing particles in a seeding region and tracking them for a fixed time or until they reach a target region.

To understand connectivity analysis in Physical Oceanography it is essential to understand Lagrangian Particle Release Experiments (Section 2.1), how the connectivity between ocean regions is commonly visualized and analyzed (Section 2.2), and how ocean regions can be defined (Section 2.3).

2.1 Lagrangian Particle Release Experiments

Lagrangian Particle Release Experiments release particles and track them over time. The name *Lagrangian* relates to the reference frame of the particles. The reference frame is on the moving particle (Lagrangian) instead of on a fixed outside grid (Eulerian). Lagrangian Particle Release Experiments are broadly used to simulate the spreading of water from one region to one or more other regions [GBB+09; Ben06]. A huge ensemble of single Lagrangian particle trajectories is needed to map out the underlying velocity field [SGA+17]. The analysis of these experiments enhances the understanding of a great number of different marine phenomena. Examples are the distribution of hatchling sea turtles [SBR+14], the dispersion of fish larvae [SP14], the transport of plastics or marine debris [DBN+19; LKK+21] or the spreading of water masses [RFS+02; ZBF+20; FHB22].

There exist two commonly used methods to estimate water pathways between two ocean regions with Lagrangian Particle Release Experiments: i) using *physical tracers* and ii) simulating particle movement with *virtual particles* [Ben06; BLB+19] in an Ocean General Circulation Model (OGCM) [RDB+13; FBC+22].

Physical tracers are introduced into the water body of interest and examined in different temporal intervals. This is done through on-site measurements such as water samples taken at elaborate ship surveys [RFS+02] or Argo float data [GYB+21] which are very costly and temporally and spatially sparse. Physical tracers can be salt [CH95; RG92], CFCs [SFP+00; RFS+02] or tritium-helium [TCJ87].

Virtual particles are computed with Ocean General Circulation Models (OGCMs). OGCMs use physical knowledge on fluid dynamics to calculate the movement of the ocean. For this, the ocean is gridded into cells. The size of these cells, i.e., the resolution of the model, can differ. OGCMs give information about the changes in the hydrographic properties of the water for each cell. The virtual particles have zero spatial extend and are moved by the advective three dimensional velocity field and, if defined, by the diffusion in the OGCM.

Research on connectivity focuses on two main sources of information from the Lagrangian Particle Release Experiments: spatial and hydrographic data. "Spatial" connectivity analysis only concentrates on the

spatial information and how two or more regions are connected. When tracking particles for this kind of analysis only information on time and location is important. "Hydrographic" connectivity analysis concentrates on the differences of hydrographic properties between regions that are themselves defined by similar hydrographic properties. It investigates how the regions are connected or how their volume is changing over space or time or both. Although the spatial information is not neglectable here, it does not play a role as important as in solely spatial analyses. In addition to the time and location information, hydrographic properties are recorded in particle release experiments for this kind of analysis.

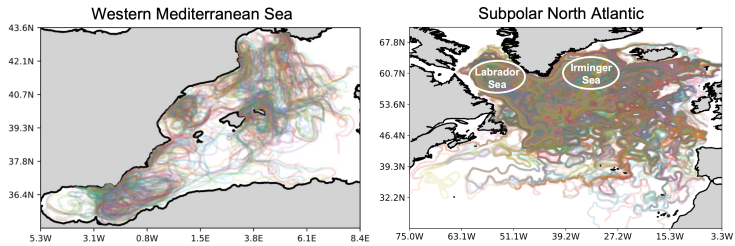

Figure 2.1. Trajectory data from simulated Lagrangian Particle Release Experiments in the Western Mediterranean Sea and the subpolar North Atlantic. The colors indicate different single trajectories. Left: Simulated migration trajectories of fish larvae [THR+21]. Right: Spreading of simulated particles seeded in the central Labrador Sea [TWH+22] (© 2022 IEEE).

Figure 2.1 shows a subsample of the trajectories that we analyzed in this work. For analyzing spatial connectivity, where the seeding region is manually defined by an area with coordinates, we used a Lagrangian Particle Release Experiment that aims at finding out how fish larvae spread in the Mediterranean Sea (see Section 3.2 and Paper B). For analyzing the more complex hydrographic connectivity, where the seeding region is defined by a certain density range in a certain area, e.g., the central Labrador Sea, we used Lagrangian Particle Release Experiments that aim at tracking how a certain water type of a certain density spreads in the subpolar North Atlantic over different time spans (see Sections 3.3, 3.4 and Papers C, D, E).

2.2 Connectivity Analysis

Traditional methods for analyzing connectivity with Lagrangian Particle Release Experiment data rely on manual and subjective input.

Complete pathways can only be described in a qualitative way with probability density maps up to now [SGA+17]. Probability density maps are heat maps that grid the complete area of interest and count the number of particle positions in each cell. This count can be normalized by the total number of position entries, data points or normalized by the number of single particles whose trajectories' positions were counted for the map. Probability density maps disregard temporal aspects completely as only the positions in the trajectory data are used. As with all heat maps, the fixed grid is another major drawback of probability density maps. The different resolutions of this grid might lead to maps that show substantially different areas with high traffic. Figure 2.2 shows the probability density maps of the trajectories that we analyzed in this work. When comparing them to the trajectories in Figure 2.1, the main pathways are clearly visible in the probability density maps. This is, however, only a qualitative description of the overall positions without any regard to movement and its direction.

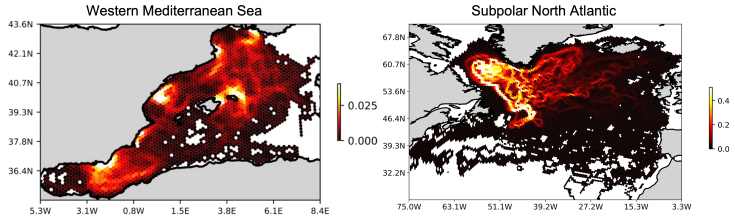

Figure 2.2. Probability density maps of the trajectory data from simulated Lagrangian Particle Release Experiments in the Western Mediterranean Sea and the subpolar North Atlantic. Lighter colors indicate higher density. See Figure 2.1 for a subsample of the original trajectories. Left: Probability density map of simulated migration trajectories of fish larvae. Right: Probability density map of the spreading of simulated particles seeded in the central Labrador Sea [TWH+22] (© 2022 IEEE).

Particles reaching a certain region are detected by manually defining a border for this region and then backtracking the particles' full trajectories until they hit this border. When analyzing areas of potential interesting

underway hydrographic landmarks, literature based assumptions about these areas need to be available before the analysis for their borders to be inserted manually. Novel, potentially interesting changing points in the hydrographic properties might stay undetected.

These methods are not applicable on a very large scale. They constrain the analysis to predefined regions and time spans which might lead to ignoring regions that were not believed to be interesting enough for being examined with these costly methods. To thoroughly analyze any region for any time span in a quantitative, data-driven, and objective way the analysis process needs to become less subjective and manual.

2.3 Defining Ocean Regions

Ocean regions are defined by a range of geo-coordinates and optional hydrographic properties. The spatial defined regions are usually based on bathymetric[1] information or political decisions, e.g., the geo-coordinates of marine protected areas. Hydrographic property ranges are defined to describe bodies of water with similar hydrographic properties.

Spatial definiton of ocean regions can be done guided by an expert or solely data driven. An expert indicates the exact coordinates of the area or sections when defining ocean regions based on bathymetry or special regions of interest [FBC+22]. A data driven approach needs to learn about the interpretation of bathymetry or politics to be able to find these ranges. Another data driven approach is to grid the whole area of interest into equi-sized cells and use these cells as generic spatial ocean regions [THR+21]. In both cases, the defined regions serve as a form of positive space, i.e., the ocean region of interest as opposed to the "negative" space of the rest of the ocean.

Hydrographic definition of ocean regions identifies coherent volumes of water with similar hydrographic properties which are called water masses. The ocean consists of water, but it is not mixed into a homogeneous volume. Water at the surface of the ocean interacts with the atmosphere and thus changes its hydrographic properties. It is then spreading through the ocean basins along regions of the same or similar density, following isopycnals.

[1] topology of the ocean floor

Different factors, such as the water's inertia, the rotation of the Earth, climate influences such as sun rays and winds move the water and change its hydrographic properties through mixing. Physical Oceanographers have identified different water masses throughout the oceans of the Earth [Tal11]. Hydrographic definiton of ocean regions has been done manually up to now [HVK21; FBC+22]. Contrary to spatially defining ocean regions, the problem with hydrographic definition is to agree on the exact ranges of the hydrographic properties. An objective way could be using data driven methods such as clustering. However, clustering approaches for oceanic hydrography come with its own problems, such as accommodating the strong difference in the data's variance between the higher and deeper depths in the water column.

Chapter 3

Research Contributions

He came nobody knows whence, and he has gone nobody knows where.
— Washington Irving, *Adventures of the Black Fisherman - 1824*

This chapter summarizes the insights generated during the interdisciplinary work between the Marine Sciences and Computer Science. Figure 3.1 visualizes the contributions of this thesis towards objectifying and quantifying connectivity analysis in terms of the complexity dimensions introduced in Section 1.2.

Section 3.1 introduces the Data Fusion framework that we developed during this work.

The following three sections give an overview on how this Data Fusion framework can be applied to different complexities of connectivity analysis with trajectories from Lagrangian Particle Release Experiments.

Section 3.2 applies the trajectory and network part of the Data Fusion framework to find the hot spots of transitional movement in fish larvae migration trajectories. This approach generates a network based on binning positional data into hexagonal cells and counting the connections between the cells to identify the transition frequencies, and thus, transition probabilities, between the cells (see Paper B).

Sections 3.3 and 3.4 apply the complete Data Fusion framework to objectively identify the changes and spreading pathways of hydrographic properties in the subpolar North Atlantic. Geo-referenced gridded data are anlyzed for clusters of similar hydrographic properties and then this information is fused with the trajectory data.

Section 3.3 fuses the geo-referenced hydrographic data while disregarding temporal aspects. The gridded data are averaged over the complete time period of interest and the trajectories are viewed as spatial point

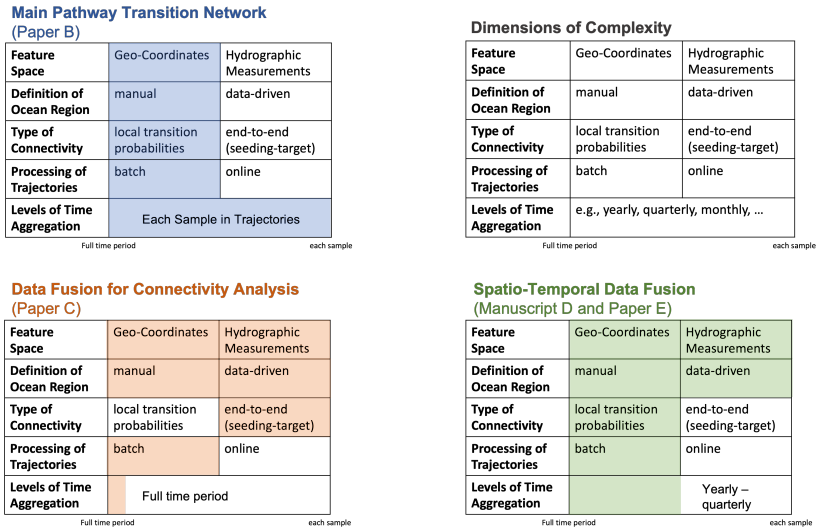

Figure 3.1. Contributions of the research on automating connectivity analysis in this thesis in terms of complexity of the connectivity analysis. The respective complexity configurations that are investigated are highlighted in blue (Section 3.2), orange (Section 3.3) and green (Section 3.4). The dimensions of complexity are explained in Section 1.2.

clouds. These two assumptions disregard temporal aspects in the analysis (see Paper C).

Section 3.4 integrates temporal aspects into the connectivity analysis. It is the most complex type of connectivity analysis regarded in this work. The hydrographic input data for the clustering are averaged over quarterly or yearly time slices yielding more time-sensitive clustering results. The trajectory segmentation is also conducted into the hydrographic space. This way, temporal aspects are regarded implicitly (see Paper D). For comparing clusters found in the different data sets, we worked on developing efficient methods to match the clusters across data sets or time steps (see Paper E).

Simultaneously to the connectivity research, we explored ways as to how it might be possible to establish a new field, Marine Data Science, as a sensible combination of Data Science and Marine Sciences.

Section 3.5 summarizes the insights of an expert workshop on this topic (see Paper A).

3.1 A Data Fusion Framework for Connectivity Analysis

We developed a Data Fusion framework to automate and objectify connectivity analysis between ocean regions. The Data Fusion framework fuses geo-referenced hydrographic data (*gridded data*) with trajectories recorded in Lagrangian Particle Release Experiments (*trajectory data*) to infer insights about the connectivity of ocean regions in a data-driven way. The research contributions summarized in Sections 3.2 to 3.4 show how this framework can be applied on connectivity analysis of different complexities (see Section 1.2). The description of the Data Fusion framework in this section has been taken from Paper C [TWH+22] (© 2022 IEEE).

We established a stage-based cross-domain fusion framework [Zhe15] consisting of a Data Enhancement step and a subsequent Data Reduction step. Figure 3.2 shows the complete process. We first enhance the data available by combining gridded and trajectory data. Then we reduce the amount of data and distill the information of interest. Through the combination of the Data Enhancement and Data Reduction steps, it is possible to exclude manual definition and investigation from the analysis.

The *Data Enhancement* combines gridded and trajectory data (Figure 3.2 left). We do a *cluster analysis* on the gridded data to then *segment* the trajectory data by assigning these clusters to them. In contrast to [ZLY+11], who segment a city map into areas along the road network and fuse this with taxi cab trajectories, our Data Fusion approach uses information from different feature spaces. The clustering is done in the space of the physical properties of seawater, *temperature* and *salinity*. The segmentation of the trajectories takes place in the longitude-latitude-depth-space, i.e. *coordinate space*.

After having enhanced the available data, they are reduced by focusing on the analysis aspects at hand (Figure 3.2, right). First, to investigate where the trajectories change the respective water cluster, the corresponding positions of the trajectories are extracted. Then, to get a view on the transition probabilities between the water clusters, the cluster changes of all trajectories are statistically summarized into a Markov Model [MTG09]. It is possible to create this Markov Model for each complete trajectory, or

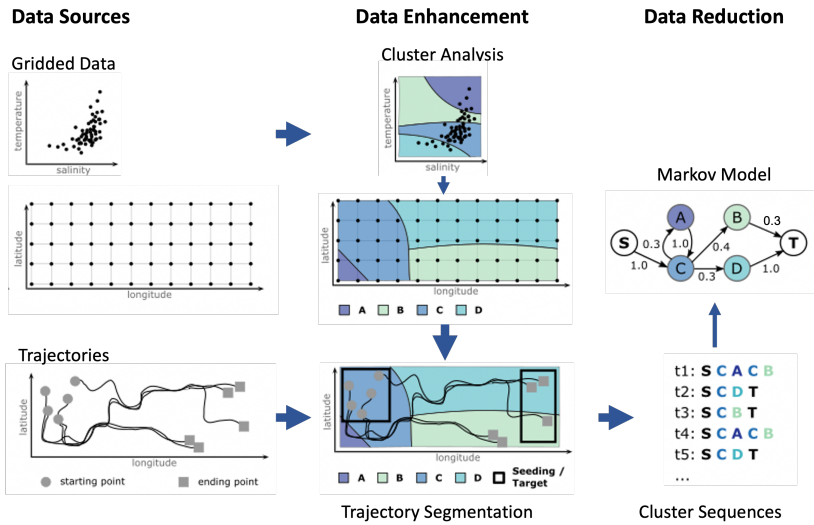

Figure 3.2. This schema shows the complete Data Fusion process in this work. Two Data Sources are combined in the Data Enhancement step. Data Reduction then reduces the total amount of data by focusing on the information relevant to the domain research aspects. The *gridded data* are temperature and salinity measures (top) that are equidistantly sampled in the coordinate space (bottom). The *trajectories* are seemingly chaotically sampled positions following the underlying velocity fields. Data Enhancement combines *water clusters* obtained in the temperature-salinity space of the *gridded data* with the particle *trajectories* into water cluster *segmented trajectories*. Data Reduction condenses the trajectories into *cluster sequence* strings by removing repeating water cluster labels to focus only on the transitions. Then it aggregates the transition probabilities between the water clusters in a *Markov Model*.
[TWH+22], © 2022 IEEE.

to select only (sub)trajectories that start at the seeding area S and end in the target area T.

For more details see Paper C.

3.2 Finding Hot Spots of Transitional Movement with a Transition Network

We investigated spatial connectivity analysis on the spread of fish larvae in the Western Mediterranean Sea (see Figure 2.1). The problem of quantitatively finding main pathways of fish larvae is partially solved in the approach presented here. We do not need to fuse data for this analysis, so the approach follows the lower row of the Data Fusion framework (Figure 3.2) concentrating completely on trajectories. Paper B [THR+21] was published containing the research presented in this section. The explanations in this section are partly taken from Paper B [THR+21].

Main Pathway Transition Network
(Paper B)

Feature Space	Geo-Coordinates	Hydrographic Measurements
Definition of Ocean Region	manual	data-driven
Type of Connectivity	local transition probabilities	end-to-end (seeding-target)
Processing of Trajectories	batch	online
Levels of Time Aggregation	Each Sample in Trajectories	

Full time period — each sample

Figure 3.3. Contributions of Paper B [THR+21] in the optimization of connectivity analysis.

The Mediterranean Sea is a high traffic sea that is nearly completely enclosed. The population of fish is an economically important factor and to keep it healthy it is crucial to protect the cradles of their larvae. Some of these areas are protected from over-fishing as marine protected areas. The fish, however, do not stay in their hatching area, but move with the ocean currents until reaching adolescence. Where exactly these movements are and how many baby fish wander along is unclear up until now. Biologists catch fish larvae and count them in several places, but this is a very expensive approach. Using simulated Lagrangian Particle Release

experiments promises to be a much less expensive approach that can cover much more space and give probable hints about the fish movement.

Problem Definition The goal is to identify regions where a high transition traffic of fishlarvae is prevalent.

Figure 3.3 shows the complexity of this connectivity analysis. The Mediterranean Sea is comparatively small with water movement mostly contained in this ocean basin. The connectivity research question focuses solely on the spatial connectivity between regions that are manually defined by coordinates. By aiming to find high traffic fish larvae migration routes it focuses on local transition probabilities. The data are completely available at the start of the analysis, so the method can assume batch processing. Time information is disregarded in this analysis. The measurements in the trajectories are not aggregated.

Transition Network from Trajectory Data

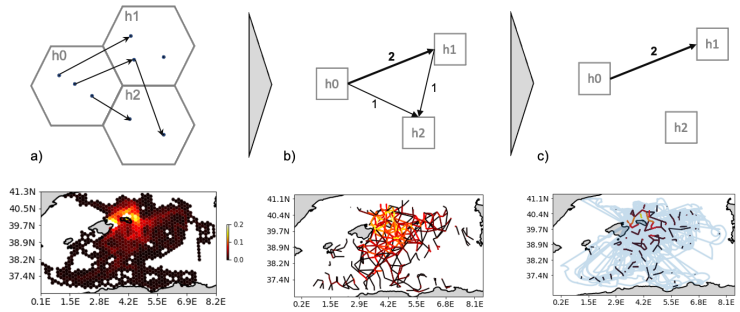

Figure 3.4. Transition network approach [THR+21]. Top: Schematic on how the algorithm works. Bottom: Exemplary Results of the processing steps for one seeding region. a) The data points of the original trajectory (pieces) are binned into hexagonal cells h0 to h2. b) The nodes denote cells h0 to h2 in the transition network. The edge weights are the sum of trajectories crossing between the cells. c) The first-degree-neighbourhood filtered network is obtained as follows: For each node, outgoing edges are kept if their weight is greater than the mean weight of all of the node's outgoing edges.

Figure 3.4 shows a schematic of the proposed method for extracting a transition network from geo-referenced trajectory data (top row) and exemplary results of the processing steps for one seeding region of the fish larvae (bottom row). The trajectory data points are binned into hexagonal bins using Uber's h3 framework [Bro18]. The transitions between these hexagonal cells per trajectory are counted and filtered by the mean transition frequency in the first-degree-neighbourhood of each cell. In the corresponding results for the fish larvae migration patterns, areas with a lot of cross over movement emerge clearly in the right most plot. The light blue background of this plot indicates the original trajectories. As in the probability density maps explained in Section 2.2, lighter colors mean higher transition probabilities. The data set [RSR21] used in this analysis can be found online[1].

The method performs well in comparison to the trajectory clustering algorithm TraClus [Jae07] and scales much better with increasing number of trajectories. The results show clearly where the local transition activity is high and indicates possible high-lavae-traffic areas.

For more details see Paper B.

Conclusion

We developed a fast approach that detects the regions where the most fish larvae cross, i.e. transitions between h3-cells, compared to the transitions in the local neighborhood. It helps to detect regions where in consequence is a lot of crossing-over movement and where heavy fishing traffic might need to be reduced. Compared to TraClus [Jae07] the presented approach scales much better with the number of trajectories.

Furthermore, we introduced a new type of and resource for trajectory data. Up to now, generally very small data sets with recorded map-independent, "organic" movement are freely available. This kind of movement is usually animal movement and as such very expensive to record. We introduced simulated animal movement trajectories with an abundance of data available. The data sets are huge and can be generated according to the research question that triggered the analysis.

[1] All data used in this study are published in: https://doi.org/10.5281/zenodo.4644862

The proposed analysis completely ignores depth information in the trajectories. The h3-cells are designed for navigating cars that move – obviously – only *along* the Earth's surface. Depth could be segmented below the h3 cells. It would make sense to do this in a data driven approach depending on how much data is available in different depths. Fish larvae stay in the first hundred meters of depth (euphotic zone) to be able to eat and get some light. Cells created equidistantly below this, would always be almost empty.

The transition network approach is a promising approach to get a fast overview on areas with intense traffic in a particle trajectory data set. Opposed to probability density maps, this approach focuses on actual movement of the particles instead of their static positions. It is also aware of the relative frequency of transitions and shows transitions that have a high relative transition frequency between cells in this local area.

3.3 Data Fusion for Detecting Dominant Spatial Movement Patterns

We investigated hydrographic and spatial connectivity analysis on the spread of newly formed water from the Labrador Sea to the Irminger Sea in the subpolar North Atlantic (see Figure 2.1, right). This type of connectivity analysis is not only concerned with the spatial pathways of the particles, but also with the change in their hydrographic properties. This approach aims at solving the limitations of traditional analysis methods and it follows the complete Data Fusion framework (Figure 3.2). Paper C [TWH+22] was published containing the research presented in this section. The explanations in this section are partly taken from Paper C [TWH+22] (©2022 IEEE).

Data Fusion for Connectivity Analysis
(Paper C)

Feature Space	Geo-Coordinates	Hydrographic Measurements
Definition of Ocean Region	manual	data-driven
Type of Connectivity	local transition probabilities	end-to-end (seeding-target)
Processing of Trajectories	batch	online
Levels of Time Aggregation	Full time period	

Full time period — each sample

Figure 3.5. Contributions of Paper C [TWH+22] in the optimization of connectivity analysis.

Understanding the connectivity between bodies of water is important for understanding the changes in ocean dynamics that influence the spreading pathways of nutrients and thus, marine ecosystems [TWH+22]. The spreading and formation of North Atlantic Deep Water and thus, the connectivity between Labrador and Irminger Sea, are a part of the Atlantic Meridional Overturning Circulation. The cold Deep Water southward flow is generally well documented, but the dynamics of formation and

subsequent spreading within the subpolar North Atlantic are not so well understood. Since the 1980's and up until now the spreading of the North Atlantic Deep Water component within the subpolar North Atlantic was subject to multiple observational [TCJ87; RFS+02; FSS07; FKO+18] and modeling studies [SPL03; GYB+20]. However, there exists no standardized manual or automated way to analyze these data. It is thus important to find a reliable and interpretable way of objectifiying connectivity analysis between ocean regions.

Problem Definition The goal is to analyze the connectivity in hydrographic properties between two manually defined regions, the seeding and target regions (e.g., Labrador Sea and Irminger Sea) in a quantitative and objective way. Of particular interest for this connectivity analysis are i) the amount of particles that started in the seeding regions and that reach the target region, ii) where the particles have been right before entering the target region, and iii) the major pathways and underway changes in hydrographic properties of the particle trajectories.

Figure 3.5 shows the complexity of this connectivity analysis. The connectivity analysis mostly focuses on the hydrographic properties and three dimensional geo-coordinates (longitude, latitude and depth). The seeding and target regions are given manually with ranges of geo-coordinates. The underway landmarks are not defined and should be found in a data-driven way. End-to-end connectivity is the main concern. Local transitions are only relevant for transitions from and to seeding and target areas. The data are completely available at the start of the analysis, so the method can assume batch processing. The time span that is regarded encompasses 10 years. The gridded data used to identify ocean regions are aggregated over the whole time span. When analyzing the trajectories, temporal information is disregarded.

Data Fusion for Connectivity Analysis between Ocean Regions

This approach combines temporally aggregated gridded data with georeferenced trajectory data while disregarding their time stamps. The data enhancement yields clusters that are found from gridded data that is

averaged over 10 years. The trajectories are segmented by these clusters. The data reduction then identifies the changing points between clusters along the complete trajectories and derives a Markov Model [MTG09] with the transition probabilities between these clusters. Figure 3.6 shows the hydrographic clustering at 855 m depth and a resulting Markov Model depicting the transition probabilities between the clusters.

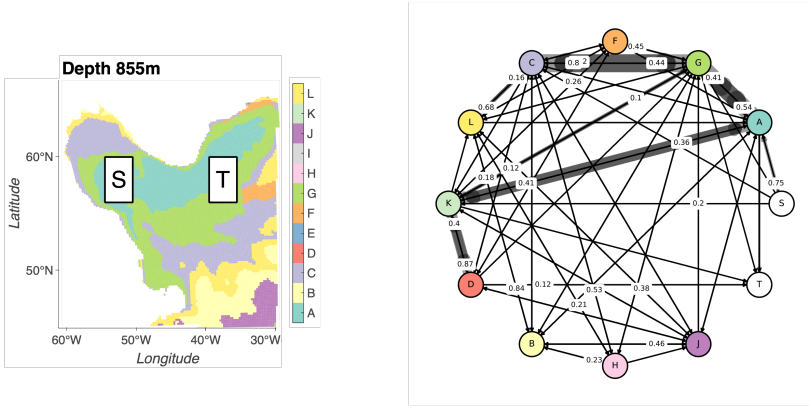

Figure 3.6. Hydrographic water clusters (A-L) at 855 m depth and a Markov Model showing the transition probabilities between the identified clusters and seeding and target regions (S and T) [TWH+22]. Note that some of the clusters that were identified across all depths (735-1655 m) are not depicted in this map in 855 m depth.
© 2022 IEEE.

For more details see Paper C.

Conclusion

We introduced a Data Fusion framework that combines several well established methods to objectify the connectivity analysis between ocean regions in the subpolar North Atlantic. The single methods, e.g., clustering, segmentation and Markov Models for transition statistics, were selected to optimally fit to the constraints and assumptions underlying from the Marine Science research question. This was done by carefully matching the

data requirements for the methods, such as normally distributed features, with the actual physically based assumptions on the research data.

This is a very important step towards automating connectivity analysis that also regards hydrographic properties, however, there are some points where further research is needed. The application in the domain still needs to be integrated into a domain scientific concrete research question and be compared to previous knowledge. This is important to validate the method in the application domain. The connectivity between ocean regions is highly influenced by local and temporal variability in the hydrographic data. Thus, disregarding temporal aspects is very likely to miss changing processes and thus distributions of water clusters. Approaches to integrate time into this framework should be investigated.

Furthermore, the optimal number of clusters in the Data Enhancement stage was obtained through an expert interview. An objective way to decide on this should be found. This is no simple task as the clusters of hydrographic properties are not well defined. They do not have distinct borders. The clusters are regions in the hydrographic space that border onto one another. They do not have space between them. The clustering should help drawing the borders between them in an objective way. An example of defining these regions manually can be seen in [LT21].

The approach presented here shows how it might be possible to custom tailor an automated analytic approach by combining sensibly selected established methods into a workflow that can be applied in various regions for connectivity analysis, not only the subpolar North Atlantic.

3.4 Integrating Temporal Aspects into Data Fusion

We investigated integrating temporal aspects into the hydrographic and spatial connectivity analysis on the differences in water formation processes in the Labrador Sea (see Figure 2.1) in the years 1993-1997. We extended the Data Fusion framework to mitigate two of its main concerns. The temporal aspects have been completely disregarded in the analysis and in the clustering in the *Data Enhancement* stage (Figure 3.2), and the optimal number of clusters was obtained through an expert interview. Manuscript D and Paper E [TTK+22] were written or published containing the research presented in this section. The explanations in this section are partly taken from manuscript D.

Spatio-Temporal Data Fusion
(Manuscript D and Paper E)

Feature Space	Geo-Coordinates	Hydrographic Measurements
Definition of Ocean Region	manual	data-driven
Type of Connectivity	local transition probabilities	end-to-end (seeding-target)
Processing of Trajectories	batch	online
Levels of Time Aggregation		Yearly – quarterly
	Full time period	each sample

Figure 3.7. Contributions of Manuscript D and Paper E [TTK+22] in the optimization of connectivity analysis.

Water formation processes in the Labrador Sea can vary between different time periods. The superficial water in the subpolar North Atlantic cools down and starts to sink into deeper layers of the ocean each winter. This water formation process is called *deep convection*. The processes related to deep convection can be stronger or weaker in more or less regular intervals. Part of ongoing research is to understand how different intensities of deep convection influence the spreading of newly formed water throughout the

Atlantic Ocean. Physical Oceanographers found several main pathways of this spreading, one of which is the spreading from the Labrador Sea to the Irminger Sea (see Figure 2.1, right).

Problem Definition The goal is to include time-sensitive clustering into the Data Fusion framework to integrate seasonal or inter-annual differences in the strength of deep convection into the connectivity analysis. These changes in deep convection and thus the changes in the positions of water clusters affect the segmentation of particle trajectories in the Data Fusion framework.

Figure 3.7 shows how incorporating time sensitivity and different periods of time affects the complexity of the connectivity analysis. The domain research still focuses on the hydrographic properties and three dimensional geo-coordinates (longitude, latitude and depth). The seeding region is given by ranges of geo-coordinates and hydrographic properties. When tracking and analyzing the changing water clusters over time, the focus lies on the local transitions between them. The data is completely available at the start of the analysis, so the method can assume batch processing. The time period that is regarded encompasses 5 years with a clear change in the physical process that is observed.

Spatio-Temporal Data Fusion

Figure 3.8 shows the extension of the Data Fusion framework to being time-sensitive by incorporating several gridded time slices $t_0 - t_n$ as input for the cluster analysis. The time slices are either averaged over all quarters (January-March, etc.) of all observed years (*quarter-yearly*) or for each year separately (*quarterly*). The clustering (and later the trajectory segmentation) is done in the hydrographic space using all data points in all time slices.

The special case of marine hydrographic data significantly simplifies the tasks of spatio-temporal clustering and segmentation. Due to the spatial relative unambiguousness of hydrographic properties in the subpolar North Atlantic and the temporal averaging according to relevant time spans, the spatio-temporal information does not need to be explicitely included in the clustering and segmentation process. Both clustering and

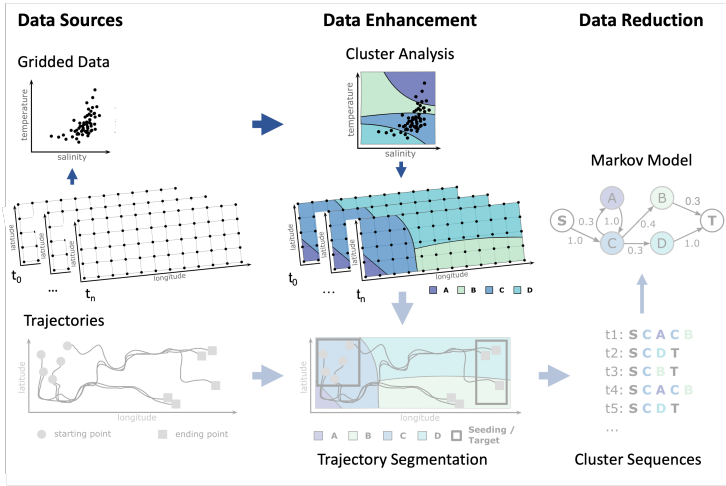

Figure 3.8. Extension of the Data Fusion framework shown in Figure 3.2 to accommodate temporal information in the clustering step.

segmentation can be done in the hydrographic space without loosing any information on the evolution of the clusters over time or their position.

For more details see Paper D.

Deciding on the Optimal Number of Clusters

The optimal number of clusters is obtained by analyzing clustering performance measures for the clustering results on the different time slices. We selected the Davies-Bouldin Score [DB79] and the Calinski-Harabasz Index [CH74] as performance measures. Figure 3.9 shows how this approach yields a helpful overview on the different hydrographic structures in the different time periods (green, red and blue) that were analyzed. The black lines indicate a good pick for the optimal number of clusters based on these plots. The lower line at twelve clusters is the same number that was picked through an expert interview [TWH+22].

For more details see Paper D.

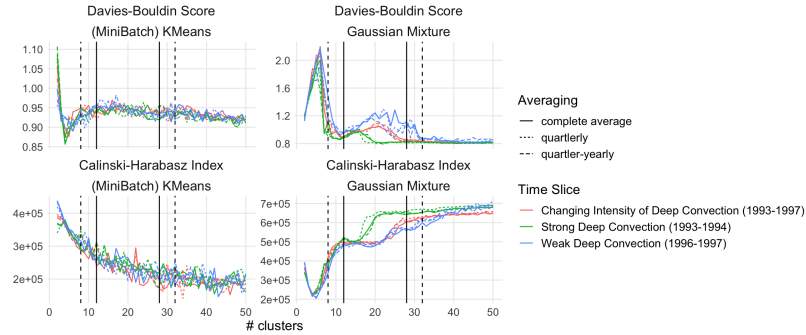

Figure 3.9. Comparison of measures for cluster separation for two clustering algorithms on data subsets of different size and different temporal aggregation. The y-axis indicates the respective scores of the measures in the title of the plots. The dotted black lines indicate the span of numbers of clusters that is suggested by the measures. The black lines indicate a good number of clusters for Gaussian Mixture models.

Cluster Tracking over Different Data Sets

To be able to compare the results of the clustering across data sets, the cluster labels need to be mapped, so that cluster A representing certain hydrographic properties is called cluster A in all data sets. This is no trivial task. The most straight forward way is to compare all entries in one cluster in data set A with the entries in all clusters in data set B [KMB05]. This is computatively very costly and we explored several ways of optimizing mapping clusters over time steps (Paper E [TTK+22] (© 2022 IEEE)). To analyze the proposed extension to the Data Fusion framework, we used pairwise cluster comparison [KMB05] implemented by [TTK+22] for mapping the clusters across the different input data sets.

For more details on optimizing the cluster mapping see Paper E.

Conclusion

Extending the Data Fusion framework to being time-sensitive by changing the way of averaging the input for clustering is another step towards a fully

automated framework for analyzing spatial and temporal connectivity in the ocean. It generates, however, the problem of matching the clusters found on different data sets. We could develop algorithms that do this in an efficient way. As a side effect, the time-sensitive clustering approach yields guiding information for deciding on the optimal number of clusters for the ocean region of interest in a purely data-driven way.

Integrating time into the Data Fusion framework is more complex than only changing the input data. We identified three ways for integrating temporal aspects into the Data Fusion framework: i) clustering on time slices instead of data averaged for the whole period of interest (*Inner time*), ii) comparing the Transition Markov Models for two different periods of time (*Outer time*), and iii) integrating traveling durations of particles, i.e. integrating how long particles stay in one cluster before moving on to the next (*Transition time*). Integrating the *inner time* into the framework serves best to analyze the influence of the changing intensity of deep convection between different time periods, which is why we focused on this in our approach. The other ways of integrating temporal aspects still need investigation. For this, a deeper understanding of the mechanics from clustering to segmenting the trajectories and then statistically aggregating the Markov Models has to be established, especially regarding the time-sensitive clustering that has been introduced in this section. Clustering over several time slices adds the number of gridded data points per time slice to the data set. Depending on the number of time slices, the data for clustering gets very large. More optimized clustering approaches for this type of data should be investigated. For example, if the hydrographic properties in a certain region stays the same between two time slices, it should be examined whether it is sensible not to add these data to the clustering data set.

The findings from this time-sensitive approach need to be integrated in ongoing domain research on the variability of deep convection in the subpolar North Atlantic. While the single clustering results catch the dynamics of the hydrographic structures, the statistics on how many particles move between the clusters and between seeding and target region need still to be evaluated and contextualized.

The temporal extension of the Data Fusion framework by changing the way of averaging the input for clustering is a highly effective and elegant

approach that gives additionally deeper insights into the hydrographic structure of the region of interest.

3.5 Marine Data Science: Combining Data Science and Marine Sciences

We investigated theoretically and hands-on how to best combine Data Science and the Marine Sciences into a new field, Marine Data Science. In addition to working in this interface between Data Science and Physical Oceanography for the time of this work, we conducted an expert workshop discussing the possibilities of Marine Data Science. In a consortium of four Professors, four PostDocs and three PhD students from both paternal sciences we identified opportunities and challenges for combining Data Science with the Marine Sciences. Figure 3.10 shows elements of Marine Data Science that were identified by this consortium and published in Paper A. The explanations in this section are partly taken from Paper A [VTA+21].

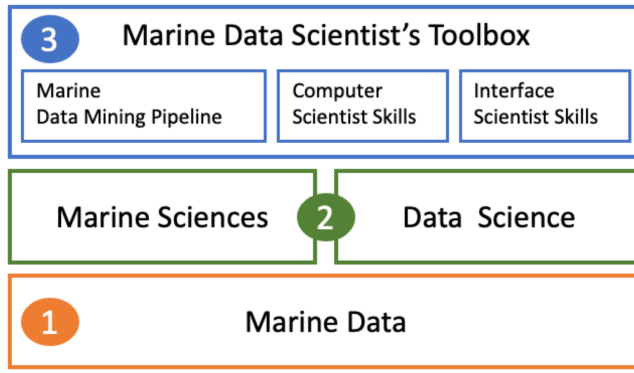

Figure 3.10. Elements of Marine Data Science (MDS) [VTA+21]. The object of MDS research is marine data (1). Knowledge involved in MDS bases on the parental sciences, Marine and Data Sciences (2). Key methods of MDS are collected in the Marine Data Scientist's Toolbox (3). This includes the Marine Data Mining Pipeline as well as Computer and Interface Scientist Skills.

In the expert consortium, we mapped out the relevant knowledge and skills in order to be a successful scientist in Marine Data Science. We identified possible pitfalls and designed a concept for a training program combining the two disciplines.

We identified these elements of Marine Data Science (Figure 3.10): 1) Marine Data, 2) Knowledge from both Data Science and the Marine Sciences, 3) Specific skills that are needed to be able to work in this interface science, the Marine Data Scientist's Toolbox.

Marine Data Marine data form the base of Marine Data Science research. Since Marine Sciences include the full range of natural sciences, Marine Data Scientists 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 Marine Data Scientists to understand the background of these data.

Knowledge from the Marine Sciences and Data Science It is neither possible nor necessary for any individual Marine Data Scientist to have a deep understanding of all Marine Sciences. As in any field, it is possible to understand the big questions and the areas of possible breakthrough that big data can mediate. Marine Data Scientists 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 Marine Data Science research questions.

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. Marine Data Scientists 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. Finally, it is crucial to transform data into information, and ultimately into knowledge. Specifically in Marine Data Science, 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.

The Marine Data Scientist's Toolbox To facilitate knowledge discovery, Marine Data Scientists 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*. To facilitate the steps of the marine data mining pipeline, Marine Data Scientists apply classical programming skills such as handling databases and UNIX platforms as well as different programming languages.

Marine Data Scientists 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 Marine Data Science project. At a personal level, Marine Data Scientists 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.

For more details see Paper A.

Chapter 4

Conclusion and Outlook

For the world is changing:
I feel it in the water, I feel it in the earth, and I smell it in the air.

— J.R.R. Tolkien, *The Return of the King - 1955*

This thesis contributes to automating and objectifying connectivity analysis between ocean regions with simulated Lagrangian Particle Release Experiments. We developed and applied a Data Fusion approach that integrates deep knowledge and understanding of the research inside Data and Computer Science with the peculiarities of the Marine Science research. We applied the approach to physical oceanographic connectivity analyses of different complexities.

We first worked to understand the Marine problems fully, to then set sensible constraints for the problem definition in Data Science. We carefully weighed strengths and weaknesses of various different methods from Data Science against the marine problems' assumptions to select the methods that promised to fit and solve the problems best. The approach for tracing Mediterranean fish larvae combined transition probability networks stemming from Data Science with Lagrangian fish larvae particles. When finding main pathways in the subpolar North Atlantic, several well established methods from Data Science were applied in this very Physical Oceanography specific field: hydrographic property clustering, trajectory segmentation, data compression and Markov Models. We showed that it is possible to depict the connectivity between ocean regions with Markov Model networks. Integrating temporal aspects into the Data Fusion framework was possible because we used the specific hydrographic properties in the subpolar North Atlantic to circumvent the problem of temporal clustering for geo-referenced time series via the input data.

All of the above are examples of where the combination of Data Science and Marine Science led to innovative new solutions for long existing problems in the Marine Sciences. Data Science benefits as well from these new fields of application with data sets that have profoundly different and more variable properties than the data sets that have been freely available so far. As soon as the established methods meet their limits, new methods that tackle exactly the problems arising from these very large data sets will need to be developed.

The following are the next steps in further enhancing the Data Fusion framework and could inspire further development of Data Science methods. We focused entirely on batch processing of the gridded and trajectory data in this thesis. However, e.g., Argo floats supply a stream of new trajectory bits every ten days. Thus, the Data Fusion framework should be extended to accommodate online data. The segmentation of the trajectories by previously found clusters was done with of-the-shelf classification algorithms that are not tailored to work with time series data. For our approach, these methods were sufficient as we averaged over certain time periods and within these periods we had a stable clustering structure. The clustering of hydrographic data still poses many challenges. The varying variance of the values between the surface and greater depths ($>$ 1000 m) pose a problem that is not yet solved. To circumvent this problem, we used data from between 700 and 1700 m depth where the variances are not too big for the clustering algorithms available.

Another important step is integrating and contextualizing the results back into the marine community, e.g., by publishing in Marine Science journals in addition to Computer Science conferences. When applying Data Science methods, it is essential to know the inner workings of the algorithms and have a deep and thorough understanding on how to profoundly evaluate and validate the outcomes of machine learning models, especially their results and predictions. This stretches from the statistical knowledge on how the data are selected and normalized to match the prerequisites and up-front assumptions of the algorithms that are used, to a knowledgeable evaluation of the "predictive power" of algorithms trained on a certain set of data and then tested on another.

When developing or choosing algorithms in an interdisciplinary context, it is not only important to choose algorithms that are fast and data-

efficient. The exact domain specific research questions need to be kept in mind as well. It would do little good to find a result with an algorithm that does not specifically answer the exact research question, but only some approximate of it. Furthermore, the selection needs to come up with a trustworthy model.

Trustworthiness encompasses the explainability of results, the intuitively, i.e. physical, interpretability of all parts of the algorithm, and the ability to compare the results with previously attained knowledge. During this work, we found that each of these three components is fundamental for results stemming from a machine learning driven method being accepted in the domain community. Explainability of the results is needed to communicate data driven findings to the domain community. The results might not be accepted as holding value to the community when performance measures are used that are solely based on the underlying data structure without an intuitive link to the domain research question. For building trust in the methods, easy interpretability of each of the building blocks of the final algorithms as "making sense" and "doing something sensible" is the most important part for furthering application of the algorithm in the community. Finally, being able to compare previous finding fast and without having to transform them into complex data structures that fit the outcome of the new algorithm helps building confidence in future findings derived from the algorithm.

In this thesis we shone a light on the opportunities and innovation that arise when combining Data Science and Marine Science. We successfully solved automating and objectifiying different aspects of connectivity analysis in Physical Oceanography and presented the methods and results in the Marine Science community. At the same time we published the methods in the Data Science community. This shows that the combination of Data Science and Marine Science is a - albeit sometimes frustrating - very prolific endeavor.

Bibliography

[Arg22] Argo. *Argo float data and metadata from global data assembly centre (argo gdac)*. Dataset. 2022. DOI: https://doi.org/10.17882/42182. URL: https://www.seanoe.org/data/00311/42182/.

[Ben06] Andrew Bennett. *Lagrangian fluid dynamics*. Cambridge University Press, 2006.

[BLB+19] A. Bower, S. Lozier, A. Biastoch, K. Drouin, N. Foukal, H. Furey, M. Lankhorst, S. Rühs, and S. Zou. "Lagrangian views of the pathways of the atlantic meridional overturning circulation". In: *J. Geophys. Res. Oceans* 124.8 (Aug. 2019), pp. 5313–5335. ISSN: 2169-9275. DOI: 10.1029/2019jc015014. URL: https://doi.org/10.1029/2019JC015014.

[Bro18] Isaac Brodsky. *H3: uber's hexagonal hierarchical spatial index*. 2018.

[CH74] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[CH95] Stuart A. Cunningham and Thomas W. N. Haine. "Labrador sea water in the eastern north atlantic. part ii: mixing dynamics and the advective-diffusive balance". In: *J. Phys. Oceanogr.* 25.4 (Apr. 1995), pp. 666–678. ISSN: 0022-3670. DOI: 10.1175/1520-0485(1995)025<0666:lswite>2.0.co;2. URL: https://doi.org/10.1175/1520-0485(1995)025%3C0666:LSWITE%3E2.0.CO;2.

[DB79] David L. Davies and Donald W. Bouldin. "A cluster separation measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[DBN+19] Jonathan V. Durgadoo, Arne Biastoch, Adrian L. New, Siren Rühs, Aylmer J. G. Nurser, Yann Drillet, and Jean-Raymond Bidlot. "Strategies for simulating the drift of marine debris". In: *Journal of Operational Oceanography* (2019), pp. 1–12.

[FBC+22] A. D. Fox et al. "Exceptional freshening and cooling in the eastern subpolar north atlantic caused by reduced labrador sea surface heat loss". In: *Ocean Science Discussions* 2022 (2022), pp. 1–35. DOI: 10.5194/os-2022-18. URL: https://os.copernicus.org/preprints/os-2022-18/.

[FHB22] Jörg Fröhle, Patricia Handmann, and Arne Biastoch. "Major sources of north atlantic deep water in the subpolar north atlantic from lagrangian analyses in a high–resolution ocean model". In: *Ocean Science* (2022).

[FKO+18] Jürgen Fischer, Johannes Karstensen, Marilena Oltmanns, and Sunke Schmidtko. "Mean circulation and eke distribution in the labrador sea water level of the subpolar north atlantic". In: *Ocean Science Discussions* (2018), pp. 1–27.

[FSS07] Anastasia Falina, Artem Sarafanov, and Alexey Sokov. "Variability and renewal of labrador sea water in the irminger basin in 1991–2004". In: *Journal of Geophysical Research: Oceans* 112.C1 (2007).

[GBB+09] Stephen M. Griffies, Arne Biastoch, Claus Böning, Frank Bryan, Gokhan Danabasoglu, Eric P. Chassignet, Matthew H. England, Rüdiger Gerdes, Helmuth Haak, Robert W. Hallberg, et al. "Coordinated ocean-ice reference experiments (cores)". In: *Ocean Modelling* 26.1 (2009), pp. 1–46.

[GYB+20] Sotiria Georgiou, Stefanie L. Ypma, Nils Brüggemann, Juan-Manuel Sayol, Julie D. Pietrzak, and Caroline A. Katsman. "Pathways of the water masses exiting the labrador sea: the importance of boundary-interior exchanges". In: *Ocean Modelling* (2020).

[GYB+21] Sotiria Georgiou, Stefanie L. Ypma, Nils Brüggemann, Juan-Manuel Sayol, Carine G. van der Boog, Paul Spence, Julie D. Pietrzak, and Caroline A. Katsman. "Direct and indirect pathways of convected water masses and their impacts on the overturning dynamics of the labrador sea". In: *J. Geophys. Res. Oceans* 126.1 (Jan. 2021). ISSN: 2169-9275. DOI: 10.1029/2020JC016654. URL: https://doi.org/10.1029/2020JC016654.

[HVK21] Patricia Handmann, Martin Visbeck, and Torsten Kanzow. "Ocean current changes". In: *Climate Change*. Elsevier, 2021, pp. 219–249.

[Jae07] Jiawei Han Jae-gil Lee. *Trajectory clustering: a partition-and-group framework*. 2007.

[KMB05] Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. "On discovering moving clusters in spatio-temporal data". In: *Advances in Spatial and Temporal Databases*. Springer Berlin Heidelberg, 2005, pp. 364–381. DOI: 10.1007/11535331_21.

[LKK+21] Delphine Lobelle, Merel Kooi, Albert A. Koelmans, Charlotte Laufkötter, Cleo E. Jongedijk, Christian Kehl, and Erik van Sebille. "Global modeled sinking characteristics of biofouled microplastic". In: *Journal of Geophysical Research: Oceans* 126.4 (2021), e2020JC017098. DOI: https://doi.org/10.1029/2020JC017098. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JC017098.

[LT21] Mian Liu and Toste Tanhua. "Water masses in the atlantic ocean: characteristics and distributions". In: *Ocean Science* 17.2 (2021), pp. 463–486. DOI: 10.5194/os-17-463-2021. URL: https://os.copernicus.org/articles/17/463/2021/.

[MTG09] Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov chains and stochastic stability*. 2nd ed. Cambridge Mathematical Library. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511626630.

[MZP+21] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, S Berger, N Caud, Y Chen, L Goldfarb, MI Gomis, et al. "Climate change 2021: the physical science basis. contribution of working group i to the sixth

assessment report of the intergovernmental panel on climate change". In: (2021).

[PTM+17] Anne Piron, Virginie Thierry, Herlé Mercier, and Guy Caniaux. "Gyre-scale deep convection in the subpolar north atlantic ocean during winter 2014–2015". In: *Geophysical Research Letters* 44.3 (2017), pp. 1439–1447.

[RDB+13] Siren Rühs, Jonathan V. Durgadoo, Erik Behrens, and Arne Biastoch. "Advective timescales and pathways of agulhas leakage". In: *Geophys. Res. Lett.* 40.15 (Aug. 2013), pp. 3997–4000. ISSN: 0094-8276. DOI: 10.1002/grl.50782. URL: https://doi.org/10.1002/grl.50782.

[RFS+02] Monika Rhein, Jürgen Fischer, William M. Smethie, Denise Smythe-Wright, Ray F. Weiss, Christian Mertens, D.-H. Min, Uli Fleischmann, and Alfred Putzka. "Labrador sea water: pathways, cfc inventory, and formation rates". In: *Journal of Physical Oceanography* 32.2 (2002), pp. 648–665. DOI: 10.1175/1520-0485(2002)032<0648:LSWPCI>2.0.CO;2.

[RG92] Jane F. Read and W. John Gould. "Cooling and freshening of the subpolar north atlantic ocean since the 1960s". In: *Nature* 360 (Nov. 1992), p. 55. URL: https://doi.org/10.1038/360055a0.

[RSR21] Willi Rath, Christina Schmidt, and Siren Rühs. *Mediterranean sea trajectory data examples*. Version v2020.03.31.1. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4650317. URL: https://doi.org/10.5281/zenodo.4650317.

[SBR+14] Rebecca Scott, Arne Biastoch, Christian Roder, Victor A. Stiebens, and Christophe Eizaguirre. "Nano-tags for neonates and ocean-mediated swimming behaviours linked to rapid dispersal of hatchling sea turtles". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1796 (2014), p. 20141209.

[SFP+00] William M. Smethie, Rana A. Fine, Alfred Putzka, and E. Peter Jones. "Tracing the flow of north atlantic deep water using chlorofluorocarbons". In: *J. Geophys. Res.* 105.C6 (June 2000), pp. 14297–14323. ISSN: 0148-0227. DOI: 10.1029/1999jc900274. URL: https://doi.org/10.1029/1999JC900274.

[SGA+17] Erik van Sebille, Stephen M Griffies, Ryan Abernathey, Thomas P Adams, Pavel Berloff, Arne Biastoch, Bruno Blanke, Eric P Chassignet, Yu Cheng, Colin J Cotter, et al. "Lagrangian ocean analysis: fundamentals and practices". In: *Ocean Modelling* (2017).

[SP14] Erica Staaterman and Claire B. Paris. "Modelling larval fish navigation: the way forward". In: *ICES Journal of Marine Science* 71.4 (2014), pp. 918–924.

[SPL03] Fiammetta Straneo, Robert S. Pickart, and Kara Lavender. "Spreading of labrador sea water: an advective-diffusive study based on lagrangian data". In: *Deep Sea Research Part I: Oceanographic Research Papers* 50.6 (2003), pp. 701–719.

[Tal11] Lynne D. Talley. *Descriptive physical oceanography: an introduction*. Academic press, 2011.

[TCJ87] Z. Top, W. B. Clarke, and W. J. Jenkins. "Tritium and primordial 3he in the north atlantic: a study in the region of charlie-gibbs fracture zone". In: *Deep Sea Research Part A. Oceanographic Research Papers* 34.2 (Feb. 1987), pp. 287–298. ISSN: 0198-0149. URL: http://www.sciencedirect.com/science/article/pii/0198014987900872.

[THR+21] Carola Trahms, Patricia Handmann, Willi Rath, Martin Visbeck, and Matthias Renz. "Where have all the larvae gone? towards fast main pathway identification from geospatial trajectories". In: *17th International Symposium on Spatial and Temporal Databases*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 126–129. ISBN: 9781450384254. DOI: 10.1145/3469830.3470896. URL: https://doi.org/10.1145/3469830.3470896.

[TTK+22] Nelson Tavares de Sousa, Carola Trahms, Peer Kröger, Matthias Renz, René Schubert, and Arne Biastoch. "Tracking the evolution of water flow patterns based on spatio-temporal particle flow clusters". In: *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2022, pp. 246–253. DOI:

10.1109/MDM55031.2022.00054. URL: https://doi.ieeecomputersociety.org/10.1109/MDM55031.2022.00054.

[TWH+22] Carola Trahms, Yannick Wölker, Patricia Handmann, Martin Visbeck, and Matthias Renz. "Data fusion for connectivity analysis between ocean regions". In: *2022 IEEE 18th International Conference on e-Science (e-Science)*. 2022, pp. 326–335. DOI: 10.1109/eScience55777.2022.00046.

[VTA+21] 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 (2021). ISSN: 2296-7745. DOI: 10.3389/fmars.2021.678404. URL: https://www.frontiersin.org/articles/10.3389/fmars.2021.678404.

[ZBF+20] Sijia Zou, Amy Bower, Heather Furey, M. Susan Lozier, and Xiaobiao Xu. "Redrawing the iceland-scotland overflow water pathways in the north atlantic". In: *Nature Communications* 11.1 (2020), p. 1890. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15513-4. URL: https://doi.org/10.1038/s41467-020-15513-4.

[Zhe15] Yu Zheng. "Methodologies for cross-domain data fusion: an overview". In: *IEEE transactions on big data* 1.1 (2015), pp. 16–34.

[ZLY+11] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. "Urban computing with taxicabs". In: *Proceedings of the 13th International Conference on Ubiquitous Computing*. UbiComp '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 89–98. ISBN: 9781450306300. DOI: 10.1145/2030112.2030126. URL: https://doi.org/10.1145/2030112.2030126.

Part II

Papers

Paper A

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

Maria-Theresia Verwega and Carola Trahms, Avan N. Antia, Thorsten Dickhaus, Enno Prigge, Martin Prinzler, Matthias Renz, Markus Schartau, Thomas Slawig, Christopher J. Somes, Arne Biastoch

The paper has been published in
Frontiers in Marine Science.

Trahms, C. and Verwega, M.T., Antia, A.N., Dickhaus, T., Prigge, E., Prinzler, M.H.U., Renz, M., Schartau, M., Slawig, T., Somes, C.J. and Biastoch, A. (2021) Perspectives on Marine Data Science as a Blueprint for Emerging Data Science Disciplines. Frontiers in Marine Science, 8:678404. DOI: https://doi.org/10.3389/fmars.2021.678404

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

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

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

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

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.

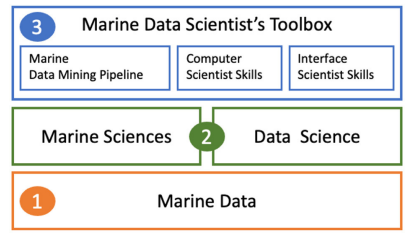

FIGURE 1 | Elements of Marine Data Science. The object of MDS research is marine data (1). Knowledge involved in MDS bases on the parental sciences, Marine and Data Sciences (2). Key methods of MDS are collected in the Marine Data Scientist's Toolbox (3). This includes the Marine Data Mining Pipeline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

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

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/

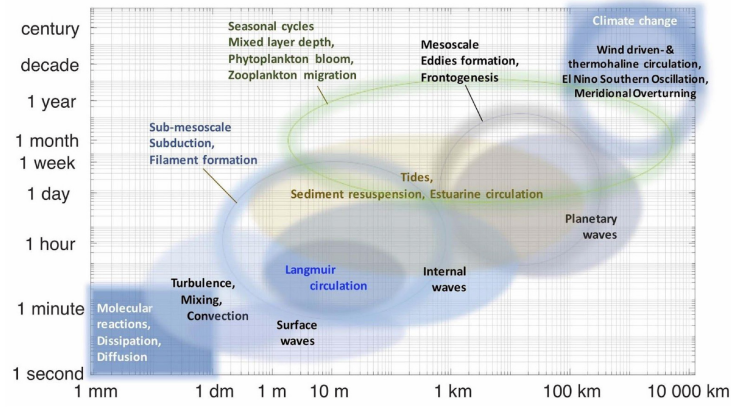

FIGURE 2 | Processes and features that generate variability in the ocean on various temporal and spatial scales. The sketch has been refined and redrawn, inspired by an illustration by Dickey (2001).

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.

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

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Verwega, Trahms, Antia, Dickhaus, Prigge, Prinzler, Renz, Schartau, Slawig, Somes and Biastoch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Paper B

Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories

Carola Trahms,
Patricia Handmann, Willi Rath, Martin Visbeck,
Matthias Renz

The paper has been published in
17th International Symposium on Spatial and Temporal Databases (SSTD),
2021.

Trahms, C., Handmann, P., Rath, W., Visbeck, M. and Renz, M. (2021). Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories. In 17th International Symposium on Spatial and Temporal Databases (SSTD 2021). Association for Computing Machinery (ACM), New York, NY, USA, 126–129. DOI: https://doi.org/10.1145/3469830.3470896

Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories

Carola Trahms
GEOMAR Helmholtz Centre for Ocean Research Kiel & Kiel University
Kiel, Germany
ctrahms@geomar.de

Patricia Handmann
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Germany
phandmann@geomar.de

Willi Rath
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Germany
wrath@geomar.de

Martin Visbeck
GEOMAR Helmholtz Centre for Ocean Research Kiel & Kiel University
Kiel, Germany
mvisbeck@geomar.de

Matthias Renz
Kiel University
Kiel, Germany
mr@informatik.uni-kiel.de

ABSTRACT

The distribution of passively drifting particles within highly turbulent flows is a classic problem in marine sciences. The use of trajectory clustering on huge amounts of simulated marine trajectory data to identify main pathways of drifting particles has not been widely investigated from a data science perspective yet. In this paper, we propose a fast and computationally light method to efficiently identify main pathways in large amounts of trajectory data. It aims at overcoming some of the issues of probabilistic maps and existing trajectory clustering approaches. Our approach is evaluated against simulated larvae dispersion data based on a real-world model that have been produced as part of work in the marine science domain.

CCS CONCEPTS

• **Applied computing** → *Earth and atmospheric sciences*; • **Theory of computation** → **Design and analysis of algorithms**.

KEYWORDS

pathway identification, transition network mining, trajectory clustering

ACM Reference Format:
Carola Trahms, Patricia Handmann, Willi Rath, Martin Visbeck, and Matthias Renz. 2021. Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories. In *17th International Symposium on Spatial and Temporal Databases (SSTD '21), August 23–25, 2021, virtual, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469830.3470896

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SSTD '21, August 23–25, 2021, virtual, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8425-4/21/08…$15.00
https://doi.org/10.1145/3469830.3470896

1 INTRODUCTION

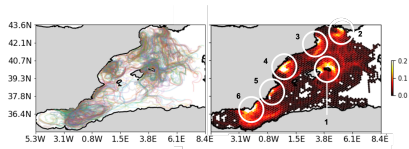

Figure 1: Left: Trajectories of fish larvae in the western Mediterranean Sea. Right: The six seeding regions protrude clearly in the probability density heatmap. The probability distribution of fish larvae is obtained by counting only the first visit of a fish larva to a hexagonal cell of ~ 36 km^2 [12].

Marine Protected Areas (MPA) are an important element for sustainable fish stock management. They are typically 'no take' zones and often are co-located with fish nurseries. Marine Biologists are interested to describe and understand the migration patterns of fish larvae and the connection to fishing grounds and protected areas. Understanding the dominant advective migration pathways of fish larvae helps to guide and optimize the location and size of MPAs. Fish larvae mainly drift with the currents while they mature. There is no economically feasible way to record the migration patterns of adolescent fish at scale, because even for a confined area, such as the western Mediterranean sea, there would still be thousands of square kilometers to monitor.

To address this challenge, simulations of fish larvae particles[1] [10] have been utilized. Lagrangian particle release experiments are a standard method to examine particle pathways [12]. Each particle's movement over time is described in trajectories. This includes its start and end points, the pathway, the length of the trajectory and the time it takes to follow it. To identify common or effective pathways, traditionally, Lagrangian trajectories are analysed using

[1] All data used in this study are published in: https://doi.org/10.5281/zenodo.4644862

probability density maps, aiming at visualizing pathways, or stream functions in order to infer the transported volume. Figure 1 (left) shows an example of a set of simulated trajectories in the western Mediterranean Sea. On the right, a heatmap denotes the probability density of simulated particles reaching a particular area. The lighter the color, the more particles are present. Six seeding regions encompassing nine MPAs are clearly visible; as are some of the main distribution ways of the larvae. When identifying pathways from Lagrangian trajectories, Marine Scientists are faced with a number of challenges. First, the amount of data, especially when using simulations, tends to be huge. Analyzing these data calls for fast and potentially parallelizable tools that minimize runtimes. Second, when utilizing one of the main advantages of Lagrangian data, the filtering of whole trajectories by their origin or destination, the whole analysis process needs to be repeated for each filtering variant. Third, the probability density maps that are state-of-the-art for visualizing pathways do not carry any information on the direction of movement and thereby neglect essential information provided by Lagrangian particle experiments.

We identified the following criteria and challenges for an algorithm for fast main pathway identification: 1) the algorithm needs to scale well to big numbers of trajectories. 2) there should be as few parameters as possible to be hand-tuned. 3) the algorithm should be able to assess the number and direction of particles that follow the pathway. In this paper we propose the transition network approach that meets these challenges by aiming at being fast, scaling well and enabling to regard the mass transport along the identified pathways.

2 RELATED WORK

There are two basic approaches to identifying main pathways in Lagrangian particle experiment data: Trajectory clustering and probability density maps.

Existing trajectory clustering approaches follow roughly the paradigm of trajectory data mining [14]. Raw spatial trajectories are preprocessed, e.g., by segmentation, alignment or compression, before the actual clustering task can be applied on the data. This leads to long runtimes mainly due to the expensive preprocessing step. Trajectories can be "organic" which means they do not follow (artificial) pathways such as roads or rails. A lot of work has been done on map-based trajectories, e.g., movement patterns of cars, trains and people in a city [3, 5, 8] or ships at sea [1, 6, 13]. There are, however, some approaches that tackle organic trajectories, such as TraClus [7]. TraClus segments trajectories and uses DBScan [4] to cluster the trajectory segments. As stated before, the segmentation preprocessing step is very expensive and does not scale well on bigger amounts of trajectories. Another drawback of this algorithm is the usage of DBScan, which requires the optimization of two parameters (epsilon and min number of neighbours). Visually finding these parameters takes a long time on huge trajectory data sets. For automatically optimizing them with, e.g., simulated annealing, an optimization measure for the clusters needs to be defined. A more recent approach transforms trajectories into transition graph networks [14]. This transformation is done by identifying *key locations* from the trajectories by, e.g., identifying stay points where particles lingered for a longer time, and regarding them as nodes in a graph. In a second step, the nodes, i.e., key locations are connected with edges that indicate transitions between them. A challenge in this approach is the identification of the key locations.

Probability density maps are a widely applied method in marine science to visualize main pathways. They subdivide space (and sometimes time) into cells in a longitude-latitude grid and count the number of particles in each cell. There are two ways of counting: counting all occurrences of a particle in a cell and counting only the first visit of a particle to a cell. The first count is normalized by the total number of trajectory data points. The second count is normalized by the number of individual particles that were released. The latter method yields sharper images of potential pathways [12] (see Fig. 1 (right) for an example). Probability density maps can either be generated on complete trajectories over the whole duration of the experiment or on a temporal "snapshot" analyzing the current state of the experiment. They qualitatively and visually identify main pathways. Extracting the exact pathways from these maps remains a challenge; as does the gridding and statistical analysis on big data sets. Additionally, these methods disregard temporal dependencies in trajectories and the direction of movement which, however, is a substantial requirement for the problem we are addressing in this paper.

3 TRANSITION NETWORKS FOR PATHWAY IDENTIFICATION

Our transition network approach is based on a space partitioning approach. The basic idea of our approach is to combine the transition graph network approach described in [14] and the ideas of probability density maps applied on a hexagonal grid system as illustrated in Figure 2. First, we discretize the trajectory space into hexagonal cells and count the number of data points within each cell (a). The bottom of Figure 2 a) shows the probability density map that results from normalizing this count by the overall number of particles that were released for the experiment, i.e., the overall number of trajectories. In a second step (b), we count the number of particle transitions between cells. These counts refer to the weights on the edges in the transition network graph, where the grid cells build the nodes of the graph and the particle transitions between adjacent cells refer to the edges of the graph. To ultimately identify the main pathways, edges with low weights are filtered by the mean of the first degree neighbourhood (c) as follows: For each node the average weight of all outgoing edges is calculated. In our example, for $h0$ this is 1.5 and for $h1$ it is 1.0. Then only the edges with a weight greater than this average are kept. $h0$ keeps the edge to $h1$ with the weight 2.0. $h1$ has only one edge with weight 1.0, which is not greater 1.0 and thus it is removed from the graph. For the spatial grid we used the H3 framework[2] introduced by the transportation network company Uber Technologies Inc. [2]. It allows to group geographical data into hexagonal cells of a selected size. 16 different resolutions are available[3] which have a hierarchical relationship. The smaller child cells have approximately 1/7 of the area of the corresponding parent cell [11]. The advantage of using this hierarchical spatial grid for our approach is the ability to adjust the spatial resolution of the main pathways one wants to retrieve.

[2] H3: Uber's Hexagonal Hierarchical Spatial Index (https://github.com/uber/H3)
[3] see https://h3geo.org/docs/core-library/restable for an overview

B. Where have all the larvae gone? Towards Fast Main Pathway Identification from Geospatial Trajectories

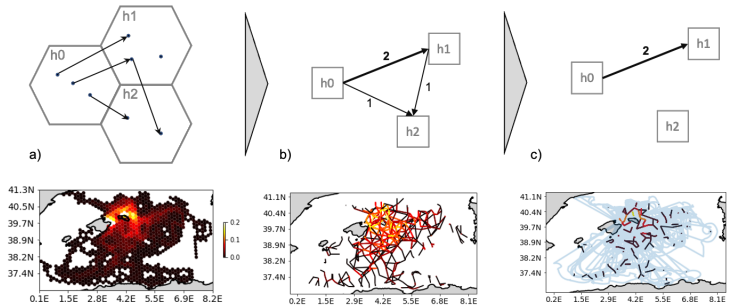

Figure 2: Transition network approach. Top: Schematic on how the algorithm works. Bottom: Exemplary Results of the processing steps for seeding region 1. a) The data points of the original trajectory (pieces) are binned into hexagonal cells h0 to h2. b) The nodes denote cells h0 to h2 in the transition network. The edge weights are the sum of trajectories crossing between the cells. c) The first-degree-neighbourhood filtered network is obtained as follows: For each node, outgoing edges are kept if their weight is greater than the mean weight of all of the node's outgoing edges.

4 EXPERIMENTAL EVALUATION

To evaluate our transition network approach against the baseline (TraClus [7]), in our experiments we focus on two seeding areas near enough to each other so that migration patterns from one region to another region appear more clearly which helps for the qualitative comparison of the two competitors. We selected seeding region 1 and 3 (see Figure 1). For our baseline TraClus, we used the implementation provided in [9]. The experiments were carried out on a MacBook Pro with a 2.3 GHz Quad-Core Intel Core i7 processor.

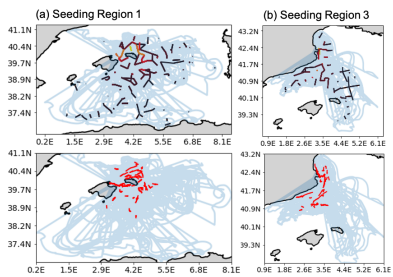

Figure 3: Top: transition network approach results for region 1 (left) and 3 (right). Bottom: TraClus results

4.1 Effectivity

For a qualitative comparison of the two competitors, Figure 3 shows the results of the transition network approach with a grid of ~ 1770 km^2 cells (h3-resolution 4) (top) and the TraClus algorithm with optimized parameters epsilon 0.01 and minimum number neighbours 3 (bottom). It shows the main pathways found by both approaches for seeding regions 1 (left) and 3 (right).

One can see that the pathways retrieved by both approaches overlap. However, the pathways identified by the transition network approach have stronger structural information, i.e., our approach is more sensitive to areas where there exist only few data points without being too sensitive in trajectory-dense areas. This effect is due to the first degree neighbourhood filtering by the mean of the edge weights around the single nodes. When the weights around a node are small, smaller weights will be accepted than in regions with a lot of traffic, i.e., where the weights are generally higher.

The main pathways in the transition network approach not only protrude more clearly when filtered by the mean edge weight for the first degree neighbourhood of each node (top), but also visualize the strength of the corresponding pathways (shown by the color, lighter colors refer to stronger pathways while weaker pathways are shown in darker colors) – a feature that is of essential interest for Marine research questions. In seeding region 1 (Figure 3 left, also cf. Figure 1), most of the larvae stay near the seeding region. There are, however, some main pathway bits near the African coast, identified by our method while lost by the TraClus approach. The fish larvae from seeding region 3 (Figure 3 right) mostly leave the seeding area and reach seeding region 1 by crossing over from Europe to the Balearic Islands. Here again, one can see that due to the ability of our approach to locally adapt the sensitivity of the main pathway identification, main pathways close to the Balearic

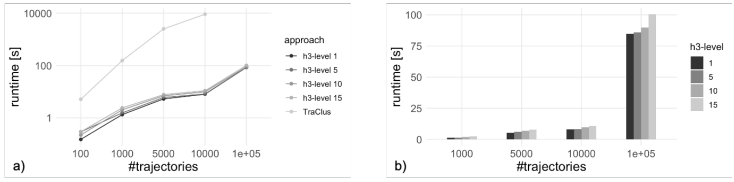

Figure 4: Left: Runtimes for the transition network approach for different resolutions (h3-levels) and for the TraClus implementation (lightest grey) over different numbers of trajectories. Right: Runtimes of the transition network approach for different numbers of trajectories and resolutions (h3-levels).

Island are only identified with our approach and remain hidden in TraClus (Figure 3 bottom).

4.2 Efficiency

To assess the efficiency of the transition network approach, we measured the runtime for building the transition probability networks for different numbers of trajectories (100, 1000, 5000, 10 000, 100 000) and compared these runtimes to the runtimes of TraClus. To further assess the scalability in terms of resolution we measured the runtime of the transition network approach for different levels of h3-resolution (1, 5, 10, 15). A h3-resolution of 1 corresponds to very coarse cells of $\sim 600\,000\ km^2$. A h3-resolution of 15 refers to the finest possible resolution of $\sim 1\ m^2$ per cell.

The left of Figure 4 compares the runtimes of the transition network approach and TraClus on different numbers of trajectories. The transition network approach is about an order of magnitude faster than the TraClus algorithm, even on a small number of trajectories. Additionally it scales much better with growing number of trajectories.

On the right of Figure 4, we evaluate the scalability of the transition network approach for varying resolutions. We can see that the resolution of the cells (h3-level) does only have minor impact on the runtime of the approach.

5 CONCLUSION AND OUTLOOK

In this paper we introduced our transition network based main pathway identification approach and compared it against the state-of-the-art trajectory clustering approach TraClus. In our experiments, we could show that the main pathways identified by our approach yield results of higher quality and cover more information relevant to Marine Scientists compared to the results of TraClus. At the same time, we showed that our approach is 1-2 orders of magnitude faster than TraClus. We showed that our approach scales well in terms of the number of trajectories at different resolution levels (Figure 4).

In the future, we will further investigate our approach on much larger data sets, higher spatial grid resolutions and other observational fields provided by the Marine Science community. Another interesting direction we want to investigate is to apply our methods on parallel systems to achieve higher scalability for very huge sets of trajectories. We will also investigate the impact of our approach in the Marine Science domain by using our approach for new insights about where, and how many fish larvae travel in the western Mediterranean Sea. Marine Scientists claim that our approach will help to estimate most likely target regions and distribution routes of marine life to enhance protection and fishing management.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Pedram Adibi, Fabio Pranovi, Alessandra Raffaetà, Elisabetta Russo, Claudio Silvestri, Marta Simeoni, Amilcar Soares, and Stan Matwin. 2020. Predicting Fishing Effort and Catch Using Semantic Trajectories and Machine Learning. In *Lecture Notes in Computer Science*. Springer International Publishing, 83–99. https://doi.org/10.1007/978-3-030-38081-6_7
[2] Isaac Brodsky. 2018. H3: Uber's hexagonal hierarchical spatial index.
[3] Zaiben Chen, Heng Tao Shen, and Xiaofang Zhou. 2011. Discovering popular routes from trajectories. In *2011 IEEE 27th International Conference on Data Engineering*. IEEE, 900–911. https://doi.org/10.1109/ICDE.2011.5767890
[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.
[5] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *KDD 2007*, Pavel Berkhin (Ed.). ACM Press, New York, New York, USA, 330. https://doi.org/10.1145/1281192.1281230
[6] Iraklis Varlamis, K. Tserpes, M. Etemad, Amilcar Soares Júnior, and S. Matwin. 2019. A Network Abstraction of Multi-vessel Trajectory Data for Detecting Anomalies. *undefined* (2019).
[7] Jiawei Han Jae-gil Lee. 2007. *Trajectory Clustering: A Partition-and-Group Framework*.
[8] Xiaoming Liu, Luxi Dong, Chunlin Shang, and Xiangda Wei. 2020. An Improved High-Density Sub Trajectory Clustering Algorithm. *IEEE Access* 8 (2020), 46041–46054. https://doi.org/10.1109/ACCESS.2020.2974059
[9] Alex Polcyn. 2016. traclus_impl. https://github.com/apolcyn/traclus_impl.
[10] Willi Rath, Christina Schmidt, and Siren Rühs. 2021. *Mediterranean Sea Trajectory Data Examples*. https://doi.org/10.5281/zenodo.4650317
[11] Vojtěch Uher, Petr Gajdoš, Václav Snášel, Yu-Chi Lai, and Michal Radecký. 2019. Hierarchical Hexagonal Clustering and Indexing. https://doi.org/10.3390/sym11060731
[12] Erik van Sebille, Stephen M Griffies, Ryan Abernathey, Thomas P Adams, Pavel Berloff, Arne Biastoch, Bruno Blanke, Eric P Chassignet, Yu Cheng, Colin J Cotter, et al. 2017. Lagrangian ocean analysis: fundamentals and practices. *Ocean Modelling* (2017).
[13] Xuantong Wang, Jing Li, and Tong Zhang, [n.d.]. A Machine-Learning Model for Zonal Ship Flow Prediction Using AIS Data: A Case Study in the South Atlantic States Region. https://doi.org/10.3390/jmse7120463
[14] Yu Zheng. 2015. Trajectory Data Mining. *ACM Transactions on Intelligent Systems and Technology* 6, 3 (2015), 1–41. https://doi.org/10.1145/2743025

Paper C

Data Fusion for Connectivity Analysis between Ocean Regions

Carola Trahms,
Yannick Wölker, Patricia Handmann, Martin Visbeck,
Matthias Renz

The paper has been published in
IEEE 18th International Conference on e-Science (e-Science), 2022.

Trahms, C., Wölker, Y., Handmann, P., Visbeck, M. and Renz, M. (2022), "Data Fusion for Connectivity Analysis between Ocean Regions," 2022 IEEE 18th International Conference on e-Science (e-Science), Salt Lake City, UT, USA, 2022, pp. 326-335, DOI: https://doi.org/10.1109/eScience55777.2022.00046.

C. Data Fusion for Connectivity Analysis between Ocean Regions

© 2022 IEEE. Reprinted, with permission, from Carola Trahms, Yannick Wölker, Patricia Handmann, Martin Visbeck, and Matthias Renz, "Data Fusion for Connectivity Analysis between Ocean Regions", IEEE 18th International Conference on e-Science (e-Science), 10/2022.[1]

[1] In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Kiel University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

Data Fusion for Connectivity Analysis between Ocean Regions

Carola Trahms
Data Science
Kiel University
Kiel, Germany
ctrahms@geomar.de

Yannick Wölker
Data Science
Kiel University
Kiel, Germany
ywoelker@geomar.de

Patricia Handmann
Ocean Dynamics
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Germany
phandmann@geomar.de

Martin Visbeck
Physical Oceanography
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Germany
mvisbeck@geomar.de

Matthias Renz
Data Science
Kiel University
Kiel, Germany
mr@informatik.uni-kiel.de

***Abstract*—The ocean is an important part of the global system. Tracking the connectivity between bodies of water is crucial for understanding local, regional and global changes in the ocean dynamics that mediate the spreading of nutrients and influence the marine ecosystem and ocean productivity. We developed a Data Fusion approach that enhances and automates the existing methods for the analysis of this connectivity. This approach combines and condenses two different data sources in two stages, Data Enhancement followed by Data Reduction. The Data Enhancement stage fuses equidistantly gridded data containing physical measurements and trajectories representing movement data. The Data Reduction stage aggregates the fused data into a Markov Model representation of the transition probabilities between ocean regions. We applied this framework to an exemplary analysis for the connectivity between two oceanic areas using real ocean data stemming from marine research. We show that this method directly tackles the limitations of existing marine data analysis methods and furthermore introduces new means to answer questions that had no quantitative answers up to now.**

***Index Terms*—Marine Data Science, Marine Data Fusion, Trajectory Embedding, Markov Models, Cluster Analysis**

I. Introduction

Connectivity between ocean regions is an important indicator on how heat, fresh water, nutrients and chemical tracers such as oxygen and carbondioxide but also organisms are distributed through the global ocean [1]. Existing methods to analyze the connectivity between ocean regions rely on expert knowledge that is weaved in manually [2]. This renders the whole process very slow and also subjective to the researchers' view on the research question at hand.

In this paper we describe a new *Data Fusion* framework (Figure 1) that combines two data sources using *Data Enhancement* and *Data Reduction* based on a stage based cross domain Data Fusion approach [3]. In the first Data Fusion stage *Data Enhancement*, the physical properties of the seawater (salinity and temperature) are clustered into categories. Then, the water flow trajectory data are fused with these clusters and embedded into sequences of water type clusters that are subsequently used to aggregate the data and to derive the final water cluster transition model in the *Data Reduction* stage.

As a test case, we focus on the subpolar North Atlantic part of the Atlantic Meridional Overturning Circulation (AMOC), consisting of the northward flowing Gulf stream system at the surface and deep and cold Deep Water return flows. We show in section IV how the connectivity between two large areas in the subpolar North Atlantic, the Labrador Sea and the Irminger Sea (Figure 2), can be analyzed fast and intuitively.

This paper is structured as follows. First, we give an insight into the problem we solve with our approach from an application, i.e., marine science, point of view. Then, we briefly describe the idea for the Data Fusion solution that applies very distinct methods from many different fields of machine learning and data science. In section II, we give an overview over methods for Data Fusion and different approaches that tackle application specific challenges similar to (parts of) the ones presented in this paper. In section III, we introduce the framework in an abstract way and explain how it mitigates the application specific challenges (section III-E). Section IV introduces a practical example from the marine sciences as a case study and proof of concept. It explains the actual implementation, choice of methods (e.g., for the clustering), and reasoning behind it for all steps in the framework. Section IV-D finishes with a detailed analysis of the contributions of our framework in tackling the challenges and limitations previous analytical methods stated for the application domain.

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

C. Data Fusion for Connectivity Analysis between Ocean Regions

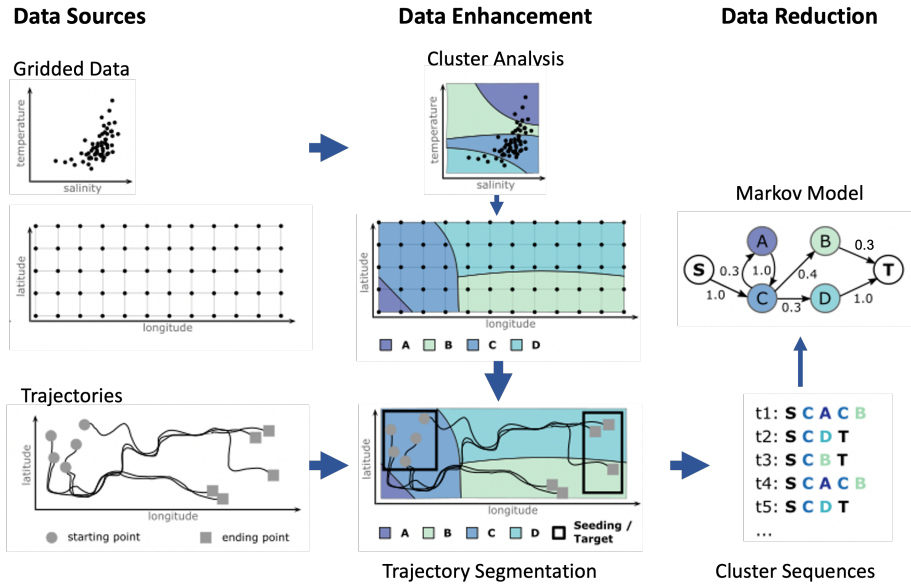

Fig. 1. This schema shows the complete Data Fusion process in this work. Two Data Sources are combined in the Data Enhancement step. Data Reduction then reduces the total amount of data by focusing on the information relevant to the domain research aspects. The *gridded data* are temperature and salinity measures (top) that are equidistantly sampled in the coordinate space (bottom). The *trajectories* are seemingly chaotically sampled positions following the underlying velocity fields. Data Enhancement combines *water clusters* obtained in the temperature-salinity space of the *gridded data* with the particle *trajectories* into water cluster *segmented trajectories*. Data Reduction condenses the trajectories into *cluster sequence* strings by removing repeating water cluster labels to focus only on the transitions. Then it aggregates the transition probabilities between the water clusters in a *Markov Model*.

A. Application-based Problem Description

Advances in ocean circulation understanding stem from direct observations in the ocean and the analysis of numerical ocean model simulations.

Direct in-situ velocity observations in the real ocean are expensive and rare. Thus, the connectivity is often derived from analyzing the distributions of chemical tracers such as salt [4], [5], oxygen, CFCs [6], [7] or tritium-helium [8]. These tracers are introduced into the ocean usually at the surface and they then spread at greater depths. These data are collected in elaborate ship surveys which are very costly and time consuming.

The observational data, however, have been used to develop Ocean General Circulation Models (OGCMs). These models solve the fundamental equations governing the ocean dynamics and calculate the oceanic velocities based on the hydrographic state variables temperature, salinity and pressure. Those models are initialized from data and forced with atmospheric properties at the surface. Particle release simulations in these models are used to generate simulated observational data for larger areas of the ocean or extended periods of time.

To examine the connectivity between two ocean regions, Seeding S and Target T, large numbers of particles are released in the seeding area S and their trajectory is recorded over time. To have statistically valid evidence, several thousands up to millions of particle trajectories are simulated and recorded. When analyzing the connectivity between S and T, generally four aspects are of interest:

1) How many of the seeded particles reach the target region?
2) Where do major changes in the hydrographic properties occur along the trajectories?
3) Where do the particles come from when they enter the target region?
4) What are major pathways (through underway landmarks) from seeding to target?

From these aspects, several problems arise concerning the efficiency and quantifiability of the analysis results.

Counting the particles reaching the target is usually done by manually defining a border for the target region and then backtracking the particles' full trajectories until they hit the border. This can be done forwards and backwards in time.

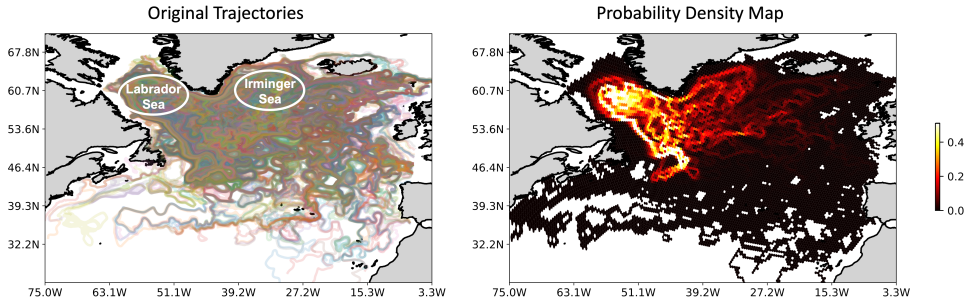

Fig. 2. Positional plot and Probability Density Map of trajectory data in the subpolar North Atlantic Ocean (SPNA).

The transition times of particles are quite long. Thus, they are usually tracked backwards to speed up the process. However, each positional measurement in the trajectories still needs to be checked and noted whether it is before or after the border. This is computationally very costly and thus not feasible for large amounts of data.

Due to the gradual change of the water characteristics along the trajectories *major changing points in these hydrographic properties* are hard to detect. Without having physically defined regions of similar water properties, which could be used as underway landmarks, the *pathway and sources of particles entering the target area even from different directions* is only qualitatively describable, e.g., in a probability density map (Figure 2 right). If literature based assumptions about interesting underway water characteristic landmarks are available, these regions need to be manually incorporated in addition to the target region. If there is no indication in literature or subjective reasoning, major changing points in the hydrographic properties have to be defined through analysis by an expert. This is again very time consuming and additionally very subjective in nature.

B. Application-based Basic Idea of our Approach

To mitigate the limitations mentioned above, we derived a method to fuse trajectory data with the gridded data from the OGCM (Figure 1). The equidistantly gridded data can be used to cluster the complete area based on water properties, i.e., salinity and temperature. These water type clusters can then be used to enhance the trajectory data by segmenting the trajectories according to the clusters. This makes changes in major water properties instantaneously visible, i.e., the switch from one cluster to another. To further speed up the processing of the trajectories, entries assigned to the same cluster can be removed except for the first consecutive occurrence. Thus, the trajectories are described and compressed by the sequence of clusters they visit. For a statistical analysis of the transition probabilities of particles moving between the clusters, a Markov Model can be calculated. This leads to a fast, intuitive, and quantitative visualization of the major movement, i.e., the connectivity, between the water type clusters.

II. Related Work

A. Data Fusion

There are several ideas on how cross domain Data Fusion can be done [9], [10]. There are three broad categories of Data Fusion approaches: Stage-based Data Fusion, Feature-level-based Data Fusion and Semantic Meaning-based Data Fusion [3]. While Feature-level-based Data Fusion simply concatenates all available measurements into one huge feature set, semantic meaning-based and stage-based Data Fusion incorporate different levels of abstraction into the Data Fusion. The semantic meaning-based Data Fusion focuses on combining high-level concepts from various data sources. Stage-based Data Fusion combines insights, e.g. from pattern recognition algorithms such as clustering, and raw data from different data sources into a stage based Data Fusion approach. As an example, [11] segments a city map into areas along the road network and fuses these areas with taxi cab trajectories. This approach uses solely spatial information for the segmentation and the fusion. However, it does not fuse data across different feature spaces.

B. Pattern Recognition and Clustering in Marine Data

There exists an abundance of methods for pattern recognition. These are, however, not always a good fit for marine data as they are usually very sparse. The ocean is a huge entity to observe with many different properties and regimes and even for ocean simulation models, it is not known enough about the inner and subgrid relations to provide dense sampling of the whole ocean [12]. This renders a completely new perspective on evaluating and selecting pattern recognition algorithms, e.g. for clustering or pathway detection, in these data as they are dependent on the process scale and data cover.

Pattern recognition, such as clustering, in hydrographic seawater properties (e.g., salinity, temperature) can be challenging, when the range and variance of the data is too large [13]. The variance of hydrographic data in great depths is much

lower than in the first 500 m near the surface [14]. Different techniques evolved to overcome this variance problem. One approach is to divide the water column into density intervals [13]. This division is needed for clustering the ocean because, e.g., temperature and salinity that define the density through a non-linear relation have very different orders of magnitude in their variance. They vary very differently throughout the water column, with strong variability in the near surface layers and a mostly homogeneous distribution in the abyssal ocean. Usually each density interval is rather concentrated around certain temperature and salinity values and a common cluster algorithm can produce results that contain important features [13].

A different approach is to introduce new dimension for each depth layer. This way, a standard scaling can fit to different variances of each depth layer [15]. The subsequent clustering yields a two dimensional cluster since the depth is encoded in the dimension and does respect the vertical movement of the density field.

C. Trajectory clustering and main pathway detection

Trajectories sampled in the real ocean or in simulations of the ocean move with the large scale currents and meso to small scale dynamics. Contrary to car trajectories, this movement appears seemingly chaotically and does not follow the concept of static roads on land. It follows ocean currents that can change their direction, position and strength over time. Additionally, some of the trajectories take very unlikely ("offroad") routes. The challenge of clustering such trajectories has been solved for animal tracks [16] although this approach does not scale well for the huge amounts of data needed to answer the marine research questions [17]. Other approaches compress the trajectories by transforming them into a sequence of automatically detected *points of interest* [18]. This has been successfully done for car trajectories [19], ship-routing [20] and animal tracks [21]. Such trajectories of point-of-interest-sequences can be statistically aggregated into Transition Probability Networks modeled in Markov Models [11] which can then be efficiently analyzed according to starting and destination areas [22]. In [23] an Absorbing Markov Chain was used to model the traffic flow to predict the most popular routes. The answered aspects are partly similar to the aspects we want to solve in this work.

III. Method

We will first give a quick overview of the components of our framework, followed by detailed descriptions for each of these components.

We established a stage-based cross-domain fusion framework [3] consisting of a Data Enhancement step and a subsequent Data Reduction step. Figure 1 shows the complete process. We first enhance the data available by combining gridded and trajectory data. Then we reduce the amount of data and distill the information of interest. Through the combination of the Data Enhancement and Data Reduction steps, it is possible to exclude manual definition and investigation from the analysis.

The *Data Enhancement* combines gridded and trajectory data (Figure 1 left). We do a *cluster analysis* on the gridded data (Section III-A) to then *segment* the trajectory data by assigning these clusters to them (Section III-B).
In contrast to [11], who segment a city map into areas along the road network and fuse this with taxi cab trajectories, our Data Fusion approach uses information from different feature spaces. The clustering is done in the space of the physical properties of seawater, *temperature* and *salinity*. The segmentation of the trajectories takes place in the longitude-latitude-depth-space, i.e. *coordinate space*.

After having enhanced the available data, they are reduced by focusing on the analysis aspects at hand (Figure 1 right). First, to investigate where the trajectories change the respective water cluster, the corresponding positions of the trajectories are extracted (Section III-C). Then, to get a view on the transition probabilities between the water clusters, the cluster changes of all trajectories are statistically summarized into a Markov Model (Section III-D). It is possible to create this Markov Model for each complete trajectory, or to select only (sub)trajectories that start at the seeding area S and end in the target area T.

A. Water Property Clustering on Gridded Data

The velocity (u, v) and hydrographic (salinity, temperature) data are equidistantly gridded in the coordinate space following the underlying grid architecture of the OGCM (x, y, depth, time). For each of the positions, latitude, longitude, depth, salinity, temperature as well as the three along grid velocities are given, which are physical properties of the water and as such not equidistant in the temperature-salinity-space (see upper-left diagram in Figure 1). We are using only the physical properties salinity and temperature to cluster the grid points. Then the clusters are mapped back to the native coordinate space. This is straight forward as the physical properties build up slowly changing volumes throughout the ocean and thus, the clusters found in the salinity-temperature space are already fully connected. Based on the cluster labels, the grid space is partitioned into spatially coherent regions of grid points belonging to the same hydrographic cluster. As a result we get a physical property-based segmentation of the grid space, where in each region, the grid points (water bodies) share the same physical properties (Figure 1).

Generally, any partition-based clustering algorithm can be applied here. To identify the clustering algorithm that yields the most plausible results, we applied fundamentally different types of cluster algorithms with a variety of parameters (see Section IV-B) and let domain experts interpret and evaluate the resulting clusters with the marine literature.

B. Segmenting Trajectories by Clusters

The water clusters from the previous step represent areas in the ocean that have similar physical properties with respect to salinity and temperature. They can be viewed as potentially

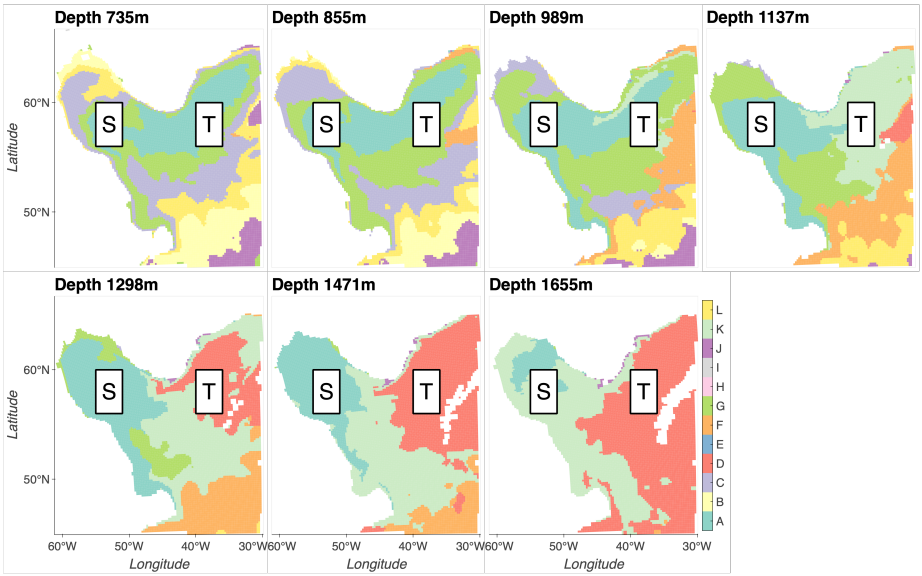

Fig. 3. Water Clusters obtained with Gaussian Mixture Models with twelve clusters and Seeding and Target areas, *S* and *T*. Seeding: lon $[-55, -51]$, lat $[56, 60]$, Target: lon $[-40, -36]$, lat $[56, 60]$

interesting water-landmarks. For the following we assume a good representation of these landmarks through the water clusters. Extra areas can be assigned for the seeding and target area of a particle release experiment (Figure 3, *S* and *T*). These artificial areas indicate the special interest for the current oceanographic research question, e.g., the connectivity between these regions. We are not interested in the timing of the trajectory movement at this point. Thus, the timestamps of the positional data points can be discarded. We use the gridded cluster information to label the time-independent positional trajectory data points by mapping them into the coordinate space of the clusters found in the previous step.

C. Trajectory Compression by Water Cluster Changing Positions

Having enhanced the data in the previous steps, now the data are going to be heavily condensed to focus on answering the domain research aspects.

The segmented trajectories can be described as a string of the clusters they cross. By removing consecutive same characters (e.g., AAABB $\rightarrow$ AB), the trajectories are massively compressed (Figure 1, bottom right). This is done either by using the last position of the to-be-left area or the first position of the area that will be entered. With the removal of characters information regarding the temporal evolution, i.e., the velocity, of the trajectories is lost. However, this information is irrelevant for the domain research aspects that this method focuses on.

D. Markov Model as Water Cluster Transition Network

The compressed string-based trajectories are used to calculate a Markov Model either for all the identified water clusters or specifically filtered for the designated seeding and target areas. We assume that the Markov Property holds here, as the water moves so slowly that the regions it visited before the current region can be regarded as irrelevant for the current transition. For the Markov Model filtered by seeding and target area, the compressed trajectories' substrings starting with the seeding area and ending on the target area are extracted. One trajectory can contain multiple substrings that are of interest for the Markov Model, but each of these substrings need to start with the seeding area and end with the target area. Based on the trajectories' substrings, the Markov Model's transition probabilities are calculated by first counting all leaving edges for each node and then normalizing the edges by the total number of leaving edges. Figure 1 shows a "complete" Markov Model without filtering for a seeding or target region. Note that the transition probabilities from node B to the target T is 0.3. This is because the Markov Model is not filtered according to the complete seeding to target area path and thus, the trajectory segments that end on B naturally reduce the overall transition probability into T. Figure 7 shows a filtered Markov

C. Data Fusion for Connectivity Analysis between Ocean Regions

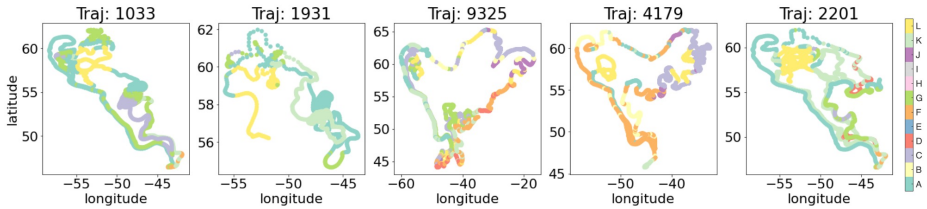

Fig. 4. A sample of segmented trajectories labelled by the respective clusters (Figure 3). The id of the trajectories is indicated in the title. The colors correspond to the clusters through which the trajectories move.

Model that indicates the transition probabilities from any of the clusters when filtering for subtrajectories between seeding and target region. Especially interesting are the seeding and target nodes (S and T). It is possible to assess fast and intuitively how the water distributes when leaving (or entering) the node. The transition probabilities for leaving S are directly visible on the edges in the Markov Model (Figure 1). For T, the incoming particles need to be counted and the probability distribution needs to be calculated from them (Table III shows an example).

E. Mitigation of current methodological limitations

Data Fusion with clustering and trajectory segmentation, objectifies the connectivity analysis of ocean regions. Up until now, marine researchers defined possible seeding and target regions for water by hand. These definitions of regions were defined by experience and previous scientific work. Due to the variability in connectivity between two oceanic regions connected to the complex interplay between the atmosphere and ocean dynamics, the exact positions of these regions are moving sometimes significantly and a spatially constant manual definition is prone to this variability. As discussed earlier the water clusters can vary also with the time from seasonal to decadal scales and differ between observations and single ocean simulation runs. If researchers were not aware of a possibly interesting water type region beforehand, this region would also not be regarded in the analysis.

The clustering in the temperature-salinity space finds general areas of similar hydrographic properties. Hence, these clusters of similar properties can be used to label the trajectories accordingly along their way. The clusters are objectively found and do not rely on manual definition before the analysis. It is possible to identify all transit areas between seeding and target region automatically, as the clusters cover the whole coordinate space of the trajectories.

The compression of the trajectory segments can be done to either mark the first entry into a new cluster or the last position before leaving the old one. These positions can then be used to plot a heatmap of the *hot* areas of water cluster transitions (see Figure 6 for an example).

The Markov Model automatically counts all the particles leaving a node to calculate the transition probabilities. Thus, by summing up the incoming edges to the target node, the total number of particles reaching the target region can be found fast and intuitively. There is no further need to backtrack all trajectories to answer this question. Related to this, the predecessors of the target node are the water regions where the particles come from before reaching the target region. These regions do not need to be known beforehand and manually predefined as the clustering does this automatically. Using the complete Markov Model, the major pathways of particles can be extracted from the number of transitions along the edges (visualized, e.g., in Figure 7). This approach makes it possible to assess these pathways quantitatively.

IV. Experimental Evaluation

To show the applicability of this method, we introduce a case study as an evaluation and proof of concept. The study focuses on the connectivity between the Labrador Sea and the Irminger Sea in the subpolar North Atlantic (SPNA). The Labrador Sea is in the basin to the west of Greenland (longitudes $> 42°$ West). The Irminger Sea lies to the East of Greenland and to the West of Iceland and the Reykjanes Ridge (Figure 2). The connectivity between these two areas plays an important role in how the Atlantic Meridional Overturning Circulation (AMOC) works.

A. Data Sources

For both data sources, the gridded and trajectory data, we chose the output of the VIKING20X-JRA-OMIP ocean simulation model [12]. It has an high spatial resolution of 1/20° in the region of interest and was run from 1958 to 2019. For our experiments we restricted the horizontal area to the North Atlantic ocean east to 30°W and north to 45°N.

For the gridded data, we decided to select the monthly ocean model output between 700m and 1800m depth to cover more than half of the trajectory data. In this way, we use the smallest range of depth layers and limit the variance in the physical measurements simultaneously to overcome the challenge of strong variabilities in temperature and salinity explained in section III-A. We averaged all the features over the year 2019.

For the trajectory data, 9809 particles were seeded in the Labrador Sea in the daily output of the VIKING20X-JRA-OMIP ocean model in the depth of 989m in the winter months and tracked daily for 10 years using the Parcels tool doing

offline trajectory calculation [24], [25]. Resulting in a total of 35148810 data points for this study.

B. Data Enhancement

Identified Water Clusters: The two dimensional data set (salinity and temperature) was scaled to zero mean and unit variance. We applied different types of partitioning clustering algorithms: distance-based (k-means), density-based (DBSCAN [26]), and model-based (Gaussian Mixture Models [27] trained with Expectation Maximization) clustering. The optimal algorithm and number of clusters was obtained by showing these clustering realizations with different number of clusters to domain experts. They declared the clusters found for twelve clusters with Gaussian Mixture Models to be in alignment with their intuition and understanding of this specific region in a way that well known features can be recognized. For the implementation we used the widely known and well established scikit-learn library to cluster our gridded data [28]. Figure 3 shows the identified clusters.

Trajectory Segmentation: To map the positional data points of the trajectories to the identified clusters, we used a k-nearest-neighbors classifier (k = 5) with randomly selected 60% of the gridded data. The selected features are longitude, latitude and depth. The features were scaled to reach from 0 to 1. We used 40% of the gridded (and labeled) data as test data to ensure that the classifier works well on unseen data. The mean accuracy on the test set is 0.92. Reaching this fairly good score boosts the confidence for using the model to segment the (unseen) positional data of the trajectories. Exemplary segmented trajectories are shown in Figure 4.

C. Data Reduction

Trajectory Compression: By removing the double entries of characters, the trajectories are hugely compressed from 3650 data points per trajectory down to at the most 1000. Figure 5 shows a histogram of the compressed trajectory lengths. The mean compressed trajectory length is 117 with a standard deviation of 59.6. The median length is 108. The segmented and compressed trajectories are used to plot a transition heatmap of where the transitions between the water clusters happen over all depths (Figure 6).

Markov Model: The compressed trajectories were filtered for complete seeding to target subtrajectories ($S \rightarrow T$) using regular expressions. Figure 7 shows a Markov Model that is calculated from these filtered cluster sequence strings. The transition probabilities are indicated by the width of the edges. For enhanced readability, only transition probabilities starting at 0.1 are shown in this Figure. Table I contains the complete adjacency matrix with all transition probabilities.

The Markov Model shows that in total 4365 particles travel from the seeding to the target area. Table II shows the distribution and transition probabilities from the seeding area to the next water cluster. Table III shows the distribution and transition probabilities from the previous water cluster into the target area.

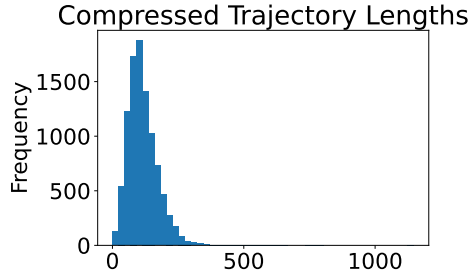

Fig. 5. Histogram of the lengths of the compressed trajectories. The original trajectory length is 3650 data points. The mean compressed trajectory length is 117 with a standard deviation of 59.6. The median length is 108.

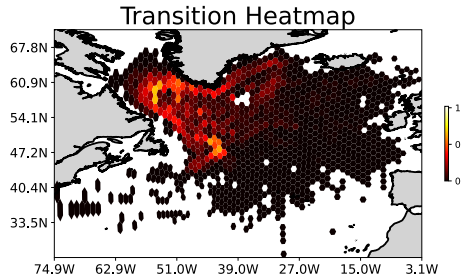

Fig. 6. Heatmap of where particles transition between water clusters in all depths.

D. Analysis by Aspects of Connectivity Analysis

In this section we assess the capabilities of our method to answer the aspects of connectivity analysis identified in section I-A.

1) How many particles reach the target region?: Looking at Figure 7, the seeding and target regions, *S* and *T*, are marked in white. This Markov Model, as explained in section IV-C, was build from all subtrajectories that start in the seeding area, *S*, and reach the target area *T* anytime during their spreading. Table II and III show that out of 9809 seeded particles a total of 4365 connect the two areas of interest. Some of them recirculated back to the seeding area and reached the target even more than once. This underlines a strong connectivity between the two regions, the Labrador and Irminger Sea, within the subpolar North Atlantic. As the traditional sections are replaced by the objective clusters the regions are now more related to their physical similarities than stiff and sharp geographical information. This can infer a major shift in the understanding of the circulation in the area, as the connectivity is clearly related to property shifts and geographic motion.

2) Where do major changes in the hydrographic properties occur along the trajectories?: Figure 6 shows a heatmap

C. Data Fusion for Connectivity Analysis between Ocean Regions

TABLE I
COMPLETE ADJACENCY MATRIX SHOWING THE TRANSITION PROBABILITIES OF THE CASE STUDY MARKOV MODEL (SECTION IV-C).

	S	A	G	F	C	L	K	D	B	H	J	**T**
S	0	0.75441	0.04170	0	0.00206	0	0.20183	0	0	0	0	0
A	0	0	0.54300	0.01903	0.00432	0	0.35751	0	0.00068	0	0.00003	0.07543
G	0	0.40516	0	0.04141	0.44031	0.00922	0.09559	0	0.00018	0.00039	0	0.00773
F	0	0.07230	0.44627	0	0.09175	0.12391	0.25530	0.01047	0	0	0	0
C	0	0.00348	0.80242	0.02026	0	0.16290	0.00009	0.00023	0.00324	0.00696	0.00042	0
L	0	0.00037	0.02778	0.09161	0.67899	0	0	0	0.17807	0.00975	0.01343	0
K	0	0.41129	0.12487	0.03596	0.00016	0.00004	0	0.40245	0	0	0.00048	0.02475
D	0	0	0.00010	0.00407	0.00010	0	0.87312	0	0	0	0.00010	0.12250
B	0	0.01173	0.00251	0	0.06114	0.84422	0	0	0	0.03769	0.04271	0
H	0	0	0.01916	0	0.52874	0.21073	0	0	0.23372	0	0.00766	0
J	0	0.02685	0.02685	0	0.02013	0.37584	0.08054	0.00671	0.46309	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0

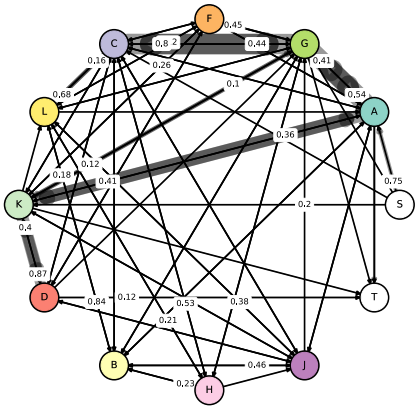

Fig. 7. Markov Model from analyzing 9809 trajectories. The width of the edges indicates the relative number of single transitions between the nodes. The corresponding *transition probabilities* are indicated near the tails of the arrows. Note that for this visualization, only transition probabilities starting at 0.1 are depicted. See Table I for the complete adjacency matrix. Table II shows the exact numbers of particles leaving the seeding area *S*. Table III shows the exact numbers of particles moving into the target area *T*. Note that this cannot be read out directly from the adjacency matrix (Table I).

TABLE II
SUCCESSORS OF THE SEEDING AREA *S*

Successor	**#Particles**	**transition probability**
A	3293	0.754
C	9	0.002
G	182	0.041
K	881	0.201

TABLE III
PREDECESSORS OF THE TARGET AREA *T*. THESE TRANSITION PROBABILITIES ARE THE PROBABILITIES FOR TRANSITION *into* THE TARGET AREA AND THUS, IS NOT DIRECTLY AVAILABLE IN THE ADJACENCY MATRIX IN TABLE I.

Predecessor	**#Particles**	**transition probability**
A	2220	0.509
D	1233	0.282
G	296	0.068
K	616	0.141

of the positions where the trajectories' labels change into a new cluster at any of the used depth layers. This is very relevant information when we want to know where the physical properties change substantially, and when we are looking to understand possible involved mechanisms. Due to the cluster analysis in the temperature-salinity space, the boundaries for substantial change in physical water properties are found in a data driven fashion and no manual thresholds have to be defined beforehand. The change in the accordingly labeled trajectory position data can thus be reliable reproduced and has a rational and methodological justification. The highlighted areas in the heatmap (Figure 6) are at positions where they are suspected by marine researchers, but up until now the analysis for proving this has been very complicated, slow and inconvenient. The areas highlighted are regions with enhanced mixing due to e.g the formation of Irminger rings west of Greenland or the interaction with the North Atlantic current at the Northwest Corner region (most southern highlight).

3) Where do the particles come from when they enter the target region?: Table III shows an overview over the transition probabilities for the predeceasing areas of the target area T. It indicates the number of transitions into the target area in all subtrajectories starting at seeding and reaching the target area. This spread of areas in combination with Figure 3 shows the main directions where the water changes into the target area. Most of the trajectories enter from areas A and D, 50.9% and 28.2%, which are also visually the biggest areas surrounding the target area. Interestingly, way smaller amounts stem from area K and G, 14.1% and 6.8%. These areas seem still very big in Figure 3, but the dynamics represented by the trajectories show that the particle moving along these dynamical fields do not visit these areas quite as often as areas A and D.

4) What are major pathways (through underway landmarks) from seeding to target?: Table II and III show an overview over the transition probabilities for the succeeding areas of the seeding area and for the predeceasing areas of the target area. It also indicates the number of transitions in all subtrajectories containing both seeding and target area. For the target area (Table III), we find that most of the transition from area A, i.e., 50.9%. As we can see in Figure 3, area A is directly connected to the seeding area *S*, e.g. in depth 1137m, and to the target area *T*, e.g. in depth 989m. This could mean that most of the transitions happen directly from *S* to *T*. It is also supported by Table II which shows that also most of the transitions from the seeding area go into area A, 75.4%. There seems to be a strong connection via area A between seeding and target area. These findings are also backed by previous domain research [e.g., [29], [30]]. Additional predeceasing areas are Area D, 28.2%, and area K, 14.1%. Area D is not a direct successor of the seeding area *S* and thus, there seems to exist another slightly more complicated path - in terms of substantial change in physical water properties - between the seeding area and the target area.

V. Outlook

This approach is tailored to a very specific set of marine research questions. For the clustering, we found an optimal number of twelve clusters in our specific case study through discussion with marine experts. In the future a more objective method to identify the optimal number of clusters while taking the physical constraints of water into account should be established. Furthermore, the temperature and salinity data varies greatly when going from the sea surface (high variance) to the bottom of the sea (very low variance). Existing clustering algorithms struggle with this kind of data yielding one big cluster for the low variance volume and concentrating on the high variance regions. For the application in water type clustering clustering algorithms are needed that have an adaptive sensitivity to the variance in the data.

The clustering on the gridded data as well as the labeling of the trajectories disregarded the temporal aspect of the moving and changing ocean. Here, a temporal average was used for the clustering, while the timestamps in the trajectories were discarded all together. Another way of approaching the trajectory labeling could be to classify the trajectory data in the temperature-salinity space instead of the coordinate space.

As promising as the results of this approach are, it would be very relevant to apply clustering algorithms for changing environments and map the identified clusters to the trajectories for subsequent consistent labeling. This solution would then have the potential to answer marine research questions regarding temporal aspects as well. Questions such as: How long do the particles travel from Seeding to Target? Not only where, but when do they change properties along the way? Are there different spreading regimes over time, highlighting different connections in the Markov Model?

VI. Conclusion

Our newly developed method mitigates many of the limitations of state-of-the-art marine research methods for particle trajectory based analysis of ocean dynamics and connectivity. It does so by combining well established Data Enhancement and Data Reduction methods in a Data Fusion approach. The results are solely data driven and objective, yet they compare very well with previous research results from more expert system approaches used in the marine sciences. This is a very promising first step into transforming previously manual and subjective analysis methods into a data driven and easily reproducible approach. It will allow to more efficiently and objectively address a whole series of questions in the marine science domain concerned with circulation and exchange between ocean regions in the context of climate change, ocean pollution and ecosystem functions.

References

[1] M. Srokosz, M. Baringer, H. Bryden, S. Cunningham, T. Delworth, S. Lozier, J. Marotzke, and R. Sutton, "Past, present, and future changes in the atlantic meridional overturning circulation," *Bull. Amer. Meteor. Soc.*, vol. 93, no. 11, pp. 1663–1676, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1175/BAMS-D-11-00151.1

[2] E. van Sebille, S. M. Griffies, R. Abernathey, T. P. Adams, P. Berloff, A. Biastoch, B. Blanke, E. P. Chassignet, Y. Cheng, C. J. Cotter *et al.*, "Lagrangian ocean analysis: fundamentals and practices," *Ocean Modelling*, 2017.

[3] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE transactions on big data*, vol. 1, no. 1, pp. 16–34, 2015.

[4] S. A. Cunningham and T. W. N. Haine, "Labrador sea water in the eastern north atlantic. part ii: Mixing dynamics and the advective-diffusive balance," *J. Phys. Oceanogr.*, vol. 25, no. 4, pp. 666–678, Apr. 1995. [Online]. Available: https://doi.org/10.1175/1520-0485(1995)025¡0666:LSWITE¿2.0.CO;2

[5] J. F. Read and W. J. Gould, "Cooling and freshening of the subpolar north atlantic ocean since the 1960s," *Nature*, vol. 360, p. 55, Nov. 1992. [Online]. Available: https://doi.org/10.1038/360055a0

[6] W. M. Smethie, R. A. Fine, A. Putzka, and E. P. Jones, "Tracing the flow of north atlantic deep water using chlorofluorocarbons," *J. Geophys. Res.*, vol. 105, no. C6, pp. 14 297–14 323, Jun. 2000. [Online]. Available: https://doi.org/10.1029/1999JC900274

[7] M. Rhein, J. Fischer, W. M. Smethie, D. Smythe-Wright, R. F. Weiss, C. Mertens, D.-H. Min, U. Fleischmann, and A. Putzka, "Labrador sea water: Pathways, cfc inventory, and formation rates," *Journal of Physical Oceanography*, vol. 32, no. 2, pp. 648–665, 2002.

[8] Z. Top, W. B. Clarke, and W. J. Jenkins, "Tritium and primordial 3he in the north atlantic: a study in the region of charlie-gibbs fracture zone," *Deep Sea Research Part A. Oceanographic Research Papers*, vol. 34, no. 2, pp. 287–298, Feb. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0198014987900872

[9] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519303902

[10] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang, "Urban big data fusion based on deep learning: An overview," *Inf. Fusion*, vol. 53, no. C, p. 123–133, jan 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.06.016

[11] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 89–98. [Online]. Available: https://doi.org/10.1145/2030112.2030126

[12] A. Biastoch, F. U. Schwarzkopf, K. Getzlaff, S. Rühs, T. Martin, M. Scheinert, T. Schulzki, P. Handmann, R. Hummels, and C. W. Böning, "Regional imprints of changes in the atlantic meridional overturning circulation in the eddy-rich ocean model viking20x," *Ocean Science Discussions*, pp. 1–52, 2021.

[13] M. Liu and T. Tanhua, "Water masses in the atlantic ocean: Characteristics and distributions," *Ocean Science*, vol. 17, pp. 463–486, 03 2021.
[14] L. D. Talley, *Descriptive physical oceanography: an introduction*. Academic press, 2011.
[15] I. Rosso, M. R. Mazloff, L. D. Talley, S. G. Purkey, N. M. Freeman, and G. Maze, "Water mass and biogeochemical variability in the kerguelen sector of the southern ocean: A machine learning approach for a mixing hot spot," *Journal of Geophysical Research: Oceans*, vol. 125, no. 3, p. e2019JC015877, 2020, e2019JC015877 10.1029/2019JC015877. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015877
[16] J. H. Jae-gil Lee, *Trajectory Clustering: A Partition-and-Group Framework*, 2007.
[17] C. Trahms, P. Handmann, W. Rath, M. Visbeck, and M. Renz, *Where Have All the Larvae Gone? Towards Fast Main Pathway Identification from Geospatial Trajectories*. New York, NY, USA: Association for Computing Machinery, 2021, p. 126–129. [Online]. Available: https://doi.org/10.1145/3469830.3470896
[18] Y. Zheng, "Trajectory data mining," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 3, pp. 1–41, 2015.
[19] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2390–2403, 2013.
[20] Iraklis Varlamis, K. Tserpes, M. Etemad, Amílcar Soares Júnior, and S. Matwin, "A network abstraction of multi-vessel trajectory data for detecting anomalies," *undefined*, 2019.
[21] J. A. Downs and M. W. Horner, "Analysing infrequently sampled animal tracking data by incorporating generalized movement trajectories with kernel density estimation," *Computers, Environment and Urban Systems*, vol. 36, no. 4, pp. 302–310, 2012.
[22] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Zufle, "Querying uncertain spatio-temporal data," in *2012 IEEE 28th International Conference on Data Engineering*, 2012, pp. 354–365.
[23] Z. Chen, H. T. Shen, and X. Zhou, "Discovering popular routes from trajectories," in *2011 IEEE 27th International Conference on Data Engineering*, 2011, pp. 900–911.
[24] M. Lange and E. van Sebille, "Parcels v0. 9: prototyping a lagrangian ocean analysis framework for the petascale age," *arXiv preprint arXiv:1707.05163*, 2017.
[25] P. Delandmeter and E. v. Sebille, "The parcels v2. 0 lagrangian framework: new field interpolation schemes," *Geoscientific Model Development*, vol. 12, no. 8, pp. 3571–3584, 2019.
[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.
[27] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[29] J. Fischer, J. Karstensen, M. Oltmanns, and S. Schmidtko, "Mean circulation and eke distribution in the labrador sea water level of the subpolar north atlantic," *Ocean Science Discussions*, pp. 1–27, 2018.
[30] R. S. Pickart, M. A. Spall, M. H. Ribergaard, G. W. K. Moore, and R. F. Milliff, "Deep convection in the irminger sea forced by the greenland tip jet," *Nature*, vol. 424, p. 152, Jul. 2003. [Online]. Available: https://doi.org/10.1038/nature01729

Paper D

Manuscript: Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions

Carola Trahms,
Patricia Handmann, Martin Visbeck, Matthias Renz

To be submitted to DASFAA 2023

D. Manuscript: Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions

Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions

Carola Trahms, Patricia Handmann, Martin Visbeck, and Matthias Renz

Abstract—Connectivity analysis of ocean regions has a pivotal role in understanding the dynamics of the ocean. A static Data Fusion framework for automating and objectifying connectivity analysis has been introduced [1]. It consists of a Data Enhancement and a Data Reduction stage. The framework neglects temporal information for the analysis throughout both stages. We propose an extension that integrates temporal aspects to this Data Fusion framework by changing time slices for averaging the gridded input data for the clustering step during the Data Enhancement stage. Data for this study were generated using a Ocean General Circulation Model that returned gridded hydrographic time series data as well as trajectories of (virtually) released particles for the subpolar North Atlantic. This study has shown that by changing the input data for the Data Fusion framework, it is possible to render it much more sensitive towards temporal variations. Furthermore, the experiments showed a sensible way for estimating the optimal number of clusters. The segmentation of trajectories gets more stable. This is another step towards fully automating the connectivity analysis between ocean regions.

Keywords—Marine Data Science, Marine Data Fusion, Temporal Cluster Analysis

I. Introduction

EACH winter in the subpolar North Atlantic, the superficial water cools down and starts to sink into deeper layers of the ocean. This process is called *deep convection*. Deep convection is an important process in the Atlantic Meriodonal Overturning Circulation (AMOC), part of which is, for example, the Gulf stream. The processes related to deep convection can be stronger or weaker in irregular intervals. Part of the ongoing research is to understand how water formed by deep convection spreads throughout the Atlantic Ocean. Physical Oceanographers found several main pathways of this spreading, one of which is the spreading from the Labrador Sea to the Irminger Sea (see Figure 1 left). To find out more about these pathways, particles are seeded in the starting region of interest (Labrador Sea) and tracked over time in a *Lagrangian Particle Release Experiment*. Although this is possible with physical tracers (e.g. salt [2], CFCs [3] or tritium-helium [4]), it is very expensive and time consuming. Using virtual particles in a Ocean General Circulation Model (OGCM) is thus very common and yields huge amounts of particle trajectory data [5]. Figure 1 shows an example of trajectories computed a simulated Lagrangian Particle Release Experiment. The probability density map on the right side is calculated from a heatmap of all particle positions normalized by the number of particles.

A. Problem Description

To aid and objectify the analysis of Lagrangian Particle Experiments, a Data Fusion approach has been introduced [1]. The Data Fusion consists of two stages, *Data Enhancement* and *Data Reduction* .

The first stage, *Data Enhancement*, hydrographic properties, i.e. temperature and salinity of the water, from different time periods are clustered to identify volumes with similar hydrographic properties. These clusters are then fused with particle trajectory data by segmenting the trajectories according to the clusters using geo-coordinates as features. *Data Reduction* builds transition probability networks from these segmented trajectories.

In the *Data Enhancement* step, the gridded data are averaged over ten years. This returns stable clusters that are prevalent most of the time. It neglects, however, seasonal and interannual differences. Hydrographic data from the subpolar North Atlantic suggest different hydrographic structures throughout the basin during strong deep convection periods versus weak ones [6]. Averaging over the complete observed time period gives only a general overview on average situation in terms of hydrographic properties. It fails completely to capture the possibly significantly different processes during strong versus during weak deep convection. Figure 2 shows this problem clearly at a depth of 855 m. In the weak deep convection time period (bottom row), cluster B is not present in the Labrador Sea (the left "ear" of the subpolar North Atlantic, see Figure 1). This cluster seems to correspond to the hydrographic properties of water that is formed during strong deep convection. This was not happening during this time. In the the other two time periods, cluster B is present in the Labrador Sea. The clustering for the changing intensity of deep convection (top row) mixes the strong and weak deep convection clusters, which is to be expected.

Another caveat of the clustering in [1] is the selection of the optimal number of clusters for clustering. While striving to automate the ocean region connectivity analysis process, the optimal number of clusters was determined by asking an expert instead of making a data-driven decision. This should be done in a data driven way. The problem here is that the clusters of hydrographic properties that this algorithms look for are not well defined. They do not have distinct borders. The clusters are regions in the hydrographic space that border onto one another. They do not have space between them. The clustering should help drawing the borders between them in an

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

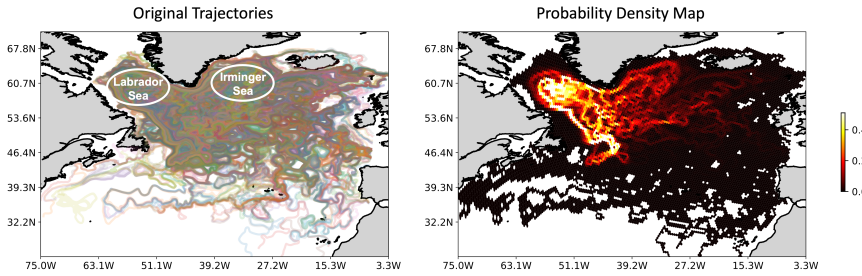

Fig. 1. Positional plot and Probability Density Map of trajectory data in the subpolar North Atlantic Ocean (SPNA).

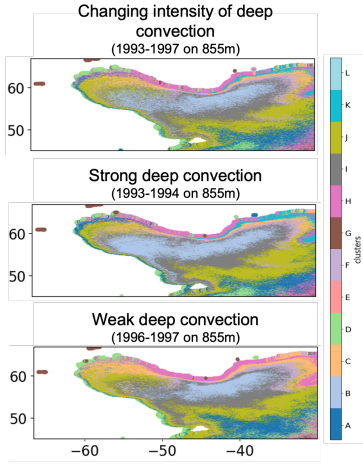

Fig. 2. Clustering results on 855 m depth for averaging over the complete time period for all three time periods examined in this study.

objective way. An example of defining these regions manually can be seen in [7].

In this work we extend the static spatial *Data Fusion* framework for connectivity analysis of ocean regions [1] by integrating time-sensitive clustering. We propose a way of preparing the data for clustering such that this seasonal variance in the clusters can be used to segment the trajectories of the particles. Integrating time into the clustering, not only regards temporal aspects, but also leads to a nice way of deciding on the appropriate number of clusters for the area and time period of interest.

II. Related Work

Analyzing and clustering trajectories stemming from atmospheric Lagrangian Particle Experiments, such as HYSPLIT [8] is an area of ongoing research, e.g., [9], [10], [11]. These approaches, however, do not use gridded data, but trajectory data.

There are several classes of clustering problems depending on how the data is available [12]. In the Data Fusion framework for ocean region connectivity analysis, trajectory data are fused with clustered geo-referenced time series. In this work we concentrate on clustering geo-referenced time series.

Clustering geo-referenced time series needs to regard not only the measurement values at the geo-coordinate and time, but also the spatial and temporal information [13]. One of the biggest challenges with this data is regarding how to integrate time and space information into the clustering algorithm [14]: The time and space could be used as features, but it is also possible to pre-cluster the data and thus aggregate according to similar time spans or regions. This can be done data driven [14] or by using previous knowledge on the data, such as the physical spatio-temporal constraints on hydrographic properties, that we will introduce in section III-B.

To assess how well a particular clustering represents the underlying structure, there exist many well established measures that are generally used for assessing the goodness of clustering results when the ground truth is unknown. Three of them are convieniently implemented in the scikit-learn machine learing toolkit[1]: the Silhouette Coefficient [15], the Davies-Bouldin Score [16] and the Calinski-Harabasz Index [17]. These measures try to evaluate the compactness of the clusters (Silhouette Coefficient and Calinski-Harabasz Index) or similarity between clusters (Davies-Bouldin Score) to assume whether the clustering separates the feature space well.

For comparing clustering results across different datasets there exist several matching methods of different efficiency [18]. A simple, yet convincing approach is to track the similarity of clusters across data sets by counting the shared data points inside two clusters [19].

[1]https://scikit-learn.org/

D. Manuscript: Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions

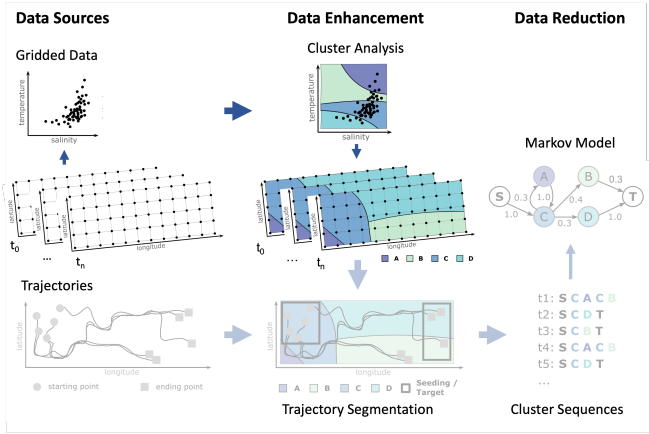

Fig. 3. This shows how the schema taken from [1] can be extended and modified to accommodate temporal information in the clustering step.

III. Method

We extend the Data Fusion framework to incorporate time information by changing the gridded input data for the clustering (Figure 3). Instead averaging over the complete time period of interest, following the ideas in [14], we suggest averaging over single time slices leading to a grid for each time span, $t_0 - t_n$. The key idea behind this averaging-approach is to have a finite set of hydrographic measures that are representative for their respective time slices. Clustering on these data yields a time-sensitive clustering. By segmenting trajectories according to these time-sensitive clusters we achieve a stable trajectory segmentation over the total length of the trajectories. Static clustering, i.e., averaging over the complete time period, leads to artefacts in trajectory segmentation. Trajectories change seemingly frequently back and forth between clusters when moving along the borders of the static clusters.

In this section, we explain what data we use for evaluating this method, what ways exist to average sensible time spans with regard to the marine domain research and how these can be used to identify an optimal number of clusters for the observed region.

A. Data Selection

Figure 4 shows how the Lagrangian Particle Release Experiment for observing water spreading ways during deep convection in the subpolar North Atlantic has been set up. Particles are set out daily. They are observed and recorded daily over 10 years leading to 3650 observations per trajectory. The intensity of deep convection changes over the years. In the years from 1987-1994 there has been exceptionally strong deep convection followed by a period of time with weak(er) deep convection until 2000 [6].

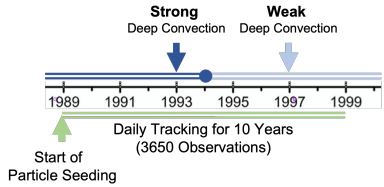

Fig. 4. Experimental setup for the simulated Lagrangian Particle Release Experiment in the Labrador Sea

For this analysis, we focus on the time period from 1993 to 1997. In this period, the deep convection changed from its strongest to a very weak intensity [2]. It is especially crucial and beneficial to apply time-sensitive hydrographic clustering in this time period (see Figure 2 for a comparison of clustering results when averaging over the complete time period 1993-1997 against clustering results when averaging over 1993-1994 and 1996-1997 separately).

To ensure comparable clustering results, we randomly subsample the averaged gridded data to 1 000 000 data points.

We selected trajectories for segmentation starting in January of each of the covered years. This leads to trajectories of 5 different lengths ranging from five years (trajectories starting in 1993) to only one year (trajectories starting in 1997).

[2]Record strong deep convection 1987-1994, and weak or no deep convection 1995-1999 [6]

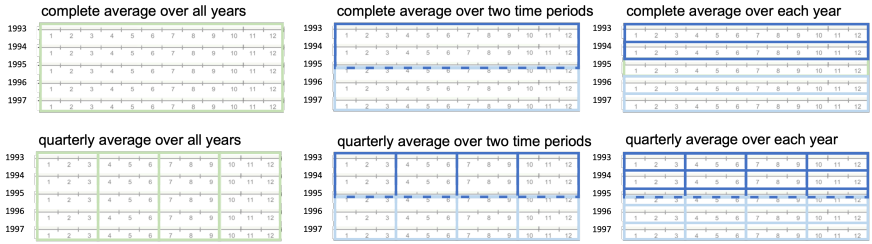

Fig. 5. Possible and marine scientific sensible time slices for averaging time in the gridded data.

B. Time slices for averaging the gridded input data

The gridded data are monthly measurements of hydrographic properties in a fixed geo-coordinate grid. Figure 5 shows time averaging slices that are available for these data and that make sense for the marine domain research. Each of the plots in Figure 5 shows the months on the x-axis and the years on the y-axis. The green and blue boxes indicate the possible spans for averaging the time. Green boxes indicating averaging over the complete time period and blue averaging over periods with strong and weak deep convection separately. Excluding year 1995 separates these two ranges. The top row shows yearly averages and the bottom row shows quarter-wise averages. Quarter-wise averging means averaging over January to March, April-June, July-September, and October-December. This approximately captures the seasons. A third row could be added that shows the same principles for month-wise averaging, but we will not use this in this analysis.

Averaging over all quarters for two time periods (Figure 5 middle, bottom) leads to a good overview on general seasonal processes in these time periods. For the full quarterly resolution in the complete time period, the data is averaged over every quarter in every year (Figure 5 right, bottom).

We aim to assess whether this way of averaging indeed extends the Data Fusion framework for ocean region connectivity. We focus here on quarterly averaging to get insights into whether this mitigates the problems shown in Figure 2.

To compare the impact of the averaging time slices, we calculate nine gridded data sets for clustering. We analyze three time periods: 1993-1994 (strong deep convection), 1996-1997 (weak or normal deep convection) and the complete time period including 1995, 1993-1997 (changing intensity of deep convection). For each of these (sub) data sets, we average over the complete time period (static), over all quarters for all years (quarterly) and over each quarter in each year (quarter-yearly).

C. Spatio-Temporal Clustering on Hydrographic Ocean Data

The gridded hydrographic data are recorded for geo-coordinates, but for clustering, these coordinates are irrelevant. The hydrographic properties in the subpolar North Atlantic do not appear in two unlinked areas at the same time [20]. So in our special case, we can abandon the time and space information while clustering only the time-averaged gridded data. The information on space and time only becomes relevant when plotting the clusters onto a map. We selected two clustering algorithms for comparison, k-means and Gaussian Mixture models. These two methods assume fundamentally different starting points. Kmeans is purely data driven, while Gaussian Mixture models build statistical models to describe the patterns in the data.

When clustering hydrographic properties, there are no nicely separated clusters to be expected. As sea water is a fluid in motion, it constantly mixes and changes its properties at stationary geo-coordinates. When aiming at determining a sensible or good number of clusters to separate volumes with similar hydrographic properties, we do not expect any of the cluster separation measures available to find the one perfect number of clusters. We hope much more to find a data-driven decision helper as to where to cut the ocean into clusters. This is relevant for marine scientists as current methods to find a good segmentation of the ocean rely on expert-defined constraints [7].

We use two measures for cluster separation to determine a good number of clusters, the Davies-Bouldin Score[16] and the Calinski-Harabasz Index[17]. We re-run clustering with k-means and Gaussian Mixture models 20 times and calculate the separation scores for each run. We use all nine time period and averaging (sub) data sets for clustering and determine these scores for each data set separately.

When performing cluster analysis on hydrographic data by testing different number of clusters, the optimal number of clusters is either obtained by finding the global optimum of the clustering separation measures or - if this is not possible - by detecting the ”elbow” or ”knee” in them [21].

For mapping the clustering results across the different data sets, we use pairwise cluster comparison [19] implemented in [18].

D. Trajectory Segmentation in the Hydrographic Space

Other than [1], we decided to segment the trajectories in the *hydrographic space* instead of the geo-coordinate and time-space. This has two advantages: First, the hydrographic space only contains two features, salinity and temperature, that are also used for water mass identification in the Marine Sciences. Second, as the classes of the trajectory segments are stable

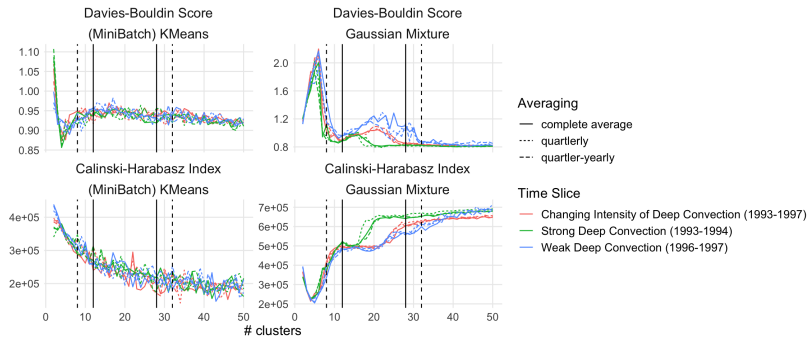

Fig. 6. Comparison of two measures for cluster separation for two clustering algorithms on data subsets of different size and different temporal aggregation. The black line indicates a good pick for the number of clusters for Gaussian Mixture models according to these plots. It is the same number of clusters that was established by showing the spatial clustering results with different numbers of clusters to domain experts [1].

over time, this allows to neglect the time dimension for the segmentation without loss of information. The time dimension is implicit in the length of the (daily sampled) trajectory data points.

IV. Evaluation

A. Deciding on a good number of clusters for the subpolar North Atlantic

We use nine different configurations of time slices and averaging methods of the gridded hydrographic data. This leads to nine different clustering results for the same region. Figure 6 shows the Davies-Boulding Score (top row) and the Calinski-Harabasz Index (bottom row) for the different clustering algorithms, k-means (left) and Gaussian Mixture models (right). The colors indicate the time slice that was regarded and the line type indicates the averaging that was used.

The optimal Davies-Boulding Score is a minimum while the optimal Calinski-Harabasz Index is a maximum. Gaussian Mixture models outperform k-means almost always. For very low number of clusters below 10, k-means works better, but overall, Gaussian Mixture models score better, i.e., lower for the Davies-Boulding Score and higher for the Calinski-Harabasz Index. We can also see that there is no visible difference between the performance measures for k-means across the time slices and averaging methods. In the remainder of this analysis we will focus on Gaussian Mixture models, as they clearly are able to catch the underlying patterns in the subpolar North Atlantic much better than k-means.

The averaging methods do not yield very different results per time slice for Gaussian Mixture models. Each time slice has its own color, so it is easy to see that the type of values per time slice follow approximately the same lines. For both performance measures, however, we can see that there are distinctly different behaviours between the time slices. While the clustering behaves similarly below eight clusters (dottet black line), there is some divergence between eight and 32 clusters depending on the time slice selected. On one hand, during strong deep convection (green) both measures saturate much earlier than during other time slices. On the other hand, during weak deep convection (blue) the saturation happens at the highest number of clusters. When looking at the complete time period (changing intensity of deep convection, red), the saturation happens between the two extrema. This is a very interesting finding as it suggests that the underlying hydrographic structure in the subpolar North Atlantic is indeed quite different during time periods of strong and weak deep convection. It is not only, as we hypothized, that there exist different types of hydrographic property clusters, but also the number of clusters and thus different hydrographic types is different.

Finding the optimal cluster numbers can be done by identifying the ”elbow” or ”knee” in the performance measure over number of clusters [21]. In this case, there exist two possible optimal number of clusters. One in the local optimum (black line at 12 clusters) just before the clustering for the time slices start to diverge and one after (black line at 28 clusters). Although the number of clusters in the saturation area might be the global optimum, we decided to continue with the 12 clusters being a local optimum and confirming - in a sense - the previous findings that 12 clusters are a good number of clusters [1].

B. Comparison of static and time-sensitive clustering results

Based on the results shown in Figure 6, we decided to pick 12 clusters for further analysis. Figure 7 shows the clustering results in 855 m depth. Appendix A contains plots of the other depths. The first two rows in Figure 7 show the quarterly averages for the time periods with strong and weak deep convection respectively. They show how the clusters

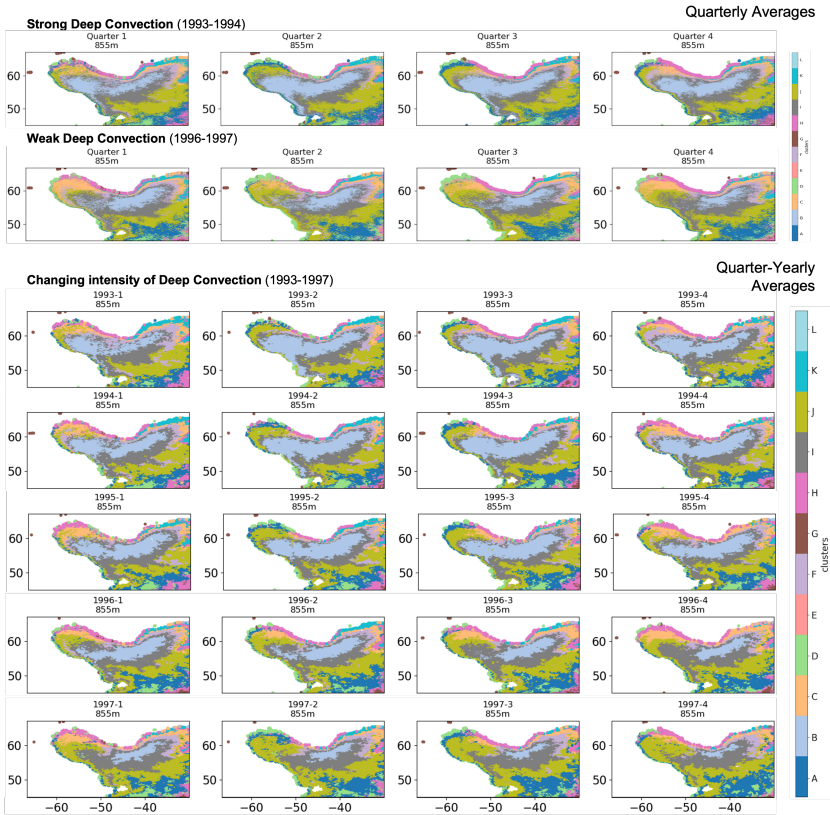

Fig. 7. Comparison of the clustering results for quarterly averages on the strong and weak deep convection periods, 1993-1994 and 1996-1997 against the clustering results for quarter-yearly averages on the complete time span from 1993-1997.

move accross the quarters for the identified processes. Note that similarly to Figure 2, during strong deep convection (1993-1994), cluster B is present in the Labrador Sea and spreads towards Irminger Sea in Quarter 2 and 3. It starts separating between Labrador Sea and Irminger Sea again in Quarter 4 which then can be connected back to Quarter 1 for the next round. During weak deep convection, cluster B is rarely present in the Labrador Sea, but very present towards Irminger Sea throughout all Quarters. There is not much cluster movement during weak deep convection in this depth.

The remainder of Figure 7 shows the quarter-yearly clusters found in the time period with changing intensity of deep convection containing the years 1993-1997. When comparing the cluster movement, e.g., for cluster B, it is clearly visible that the clustering on the quarter-yearly averaged data managed to catch the prevalent cluster movement in the strong deep convection phase (1993-1994) as well as the cluster movement during the weak deep convection phase.

This means that we managed to extend the static Data Fusion framework for Ocean Region connectivity analysis with time-sensitive clustering by changing the way of averaging the input data.

D. Manuscript: Spatio-Temporal Data Fusion for Connectivity Analysis between Ocean Regions

TABLE I
STATISTICS ON CLUSTER CHANGES IN THE TRAJECTORIES WHEN SEGMENTING WITH DIFFERENT CLUSTERINGS.

	static segmentation	time-sensitive segmentation (each quarter for all years)	time-sensitive segmentation (each quarter for each year)
mean	26.14	23.02	26.24
std	36.61	36.52	36.07
min	1.00	1.00	1.00
25^{th} percentile	2.00	1.00	2.00
median	11.00	5.00	10.00
75^{th} percentile	37.00	30.50	35.50
max	220.00	233.00	201.00

C. Comparison of trajectory segmentations for time-sensitive and time-averaged segmentation

Table I shows the statistic on how many cluster changes occur in the trajectories when segmenting with different time averaging for the time period of changing intensity of deep convection, 1993-1997. The mean number of changes between clusters is lowest for the quarterly averaged data and highest for the quarter-yearly averaged data. The standard deviation does not vary much between the averaging spans. The median shows a very distinct difference between the quarterly averaged and both the complete and quarter-yearly averaged data. It is roughly half as many changes between clusters. This is especially peculiar as there is only a very small difference between the medians of the static and quarter-yearly averaged data. On the other side, when looking at the maximum changes, the quarter-yearly averaged data based segmentation is lowest. This definitely needs deeper investigation with more trajectories that are seeded not only each January, but in each month of each year. It is already clear that different ways of averaging the data before performing the cluster analysis influences the segmentation outcome of the trajectories.

Figure 8 shows the resulting Markov Models that were computed from time-sensitive and static clustering for the complete time span 1993-1997. Further analysis is needed as to how to sensibly interpret these transition probability models. Already now we can clearly see that, e.g., the transition probabilities between cluster I (gray, bottom left) and C (light violet, left) are widely over estimated in the quarterly averaging because cluster L does not appear in this clustering method. In the quarter-yearly and curiously also in the static segmentation, we can see that most of the traffic goes not directly between cluster I and C, but via cluster L.

V. CONCLUSION AND OUTLOOK

We propose a way of flexible analyzing connectivity between ocean regions on different time scales. Using different temporal averaging approaches on different time slices of the gridded hydrographic data and then performing cluster analysis on these data sets seems to be a promising way to estimate the underlying number of clusters and thus, get informed about the differences in the underlying hydrographic structure.

The special case of marine hydrographic data significantly simplifies the tasks of spatio-temporal clustering and segmentation. Due to the spatial relative unambiguousness of hydrographic properties in the subpolar North Atlantic and the temporal averaging according to relevant slices of time, the spatio-temporal information does not need to be included in the clustering and segmentation process. So both clustering and segmentation can be done only on the hydrographic properties without loosing any information on the evolution of the clusters over time and their position in space.

The marine scientific analysis of the results of this framework, however, needs still to be done. The findings that can be derived from time-sensitive markov models need to be compared to findings that already exist.

This contribution is a step towards a fully automated framework for analyzing spatial and temporal connectivity in the ocean. Integrating temporal sensitivity into the Data Fusion framework for ocean region connectivity analysis [1] is a much need and necessary step. This only regards the cluster analysis in the framework. It is important to do a thorough analysis on the trajectory segmentation as well. The transition probability markov models could be extended to contain information on the retention time of particle within the clusters. This could be either done by introducing more edges to the graph for each retention time interval or by storing the trajectories traveling time information in the network themselves by adding another layer.

APPENDIX A

Figures showing all depths of the quarter yearly averages of the complete time period (Figure 9), the quarterly clustering of the strong deep convection period (Figure 10), and weak deep convection period (Figure 11).

ACKNOWLEDGMENT

The authors would like to thank Nelson Tavares de Sousa for providing the cluster matching code.

REFERENCES

[1] C. Trahms, Y. Wölker, P. Handmann, M. Visbeck, and M. Renz, "Data fusion for connectivity analysis between ocean regions," in *in Print: 2022 IEEE 18th International Conference on eScience (eScience)*, 2022.

[2] S. A. Cunningham and T. W. N. Haine, "Labrador sea water in the eastern north atlantic. part ii: Mixing dynamics and the advective-diffusive balance," *J. Phys. Oceanogr.*, vol. 25, no. 4, pp. 666–678, Apr. 1995. [Online]. Available: https://doi.org/10.1175/1520-0485(1995)025⟨0666:LSWITE⟩2.0.CO;2

[3] M. Rhein, J. Fischer, W. M. Smethie, D. Smythe-Wright, R. F. Weiss, C. Mertens, D.-H. Min, U. Fleischmann, and A. Putzka, "Labrador sea water: Pathways, cfc inventory, and formation rates," *Journal of Physical Oceanography*, vol. 32, no. 2, pp. 648–665, 2002.

[4] Z. Top, W. B. Clarke, and W. J. Jenkins, "Tritium and primordial 3he in the north atlantic: a study in the region of charlie-gibbs fracture zone," *Deep Sea Research Part A. Oceanographic Research Papers*, vol. 34, no. 2, pp. 287–298, Feb. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0198014987900872

[5] P. Handmann, M. Visbeck, and T. Kanzow, "Ocean current changes," in *Climate Change*. Elsevier, 2021, pp. 219–249.

[6] I. Yashayaev and J. W. Loder, "Recurrent replenishment of l abrador s ea w ater and associated decadal-scale variability," *Journal of Geophysical Research: Oceans*, vol. 121, no. 11, pp. 8095–8114, 2016.

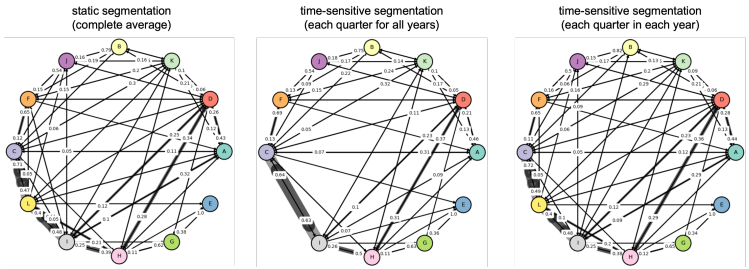

Fig. 8. Different Markov Models for the complete time span 1993-1997 resulting from time-sensitive and static clustering during Data Fusion. Increasing clustering time-sensitivity from left to right. First column: static, no time-sensitivity (1 time span), Second column: averaged per quarter over all years (4 time spans), Third column: averaged per quarter for each year separately (20 time spans)

[7] M. Liu and T. Tanhua, "Water masses in the atlantic ocean: Characteristics and distributions," *Ocean Science*, vol. 17, pp. 463–486, 03 2021. [Online]. Available: https://os.copernicus.org/articles/17/463/2021/

[8] A. F. Stein, R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. D. Cohen, and F. Ngan, "Noaa's hysplit atmospheric transport and dispersion modeling system," *Bulletin of the American Meteorological Society*, vol. 96, no. 12, pp. 2059 – 2077, 2015. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/96/12/bams-d-14-00110.1.xml

[9] L. Cui, X. Song, and G. Zhong, "Comparative analysis of three methods for hysplit atmospheric trajectories clustering," *Atmosphere*, vol. 12, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2073-4433/12/6/698

[10] S. Cheng, F. Wang, J. Li, D. Chen, M. Li, Y. Zhou, and Z. Ren, "Application of trajectory clustering and source apportionment methods for investigating trans-boundary atmospheric PM10 pollution," *Aerosol and Air Quality Research*, vol. 13, no. 1, pp. 333–342, 2013. [Online]. Available: https://doi.org/10.4209%2Faaqr.2012.06.0154

[11] A. Valenzuela, F. J. Olmo, H. Lyamani, M. Antón, A. Quirantes, and L. Alados-Arboledas, "Classification of aerosol radiative properties during african desert dust intrusions over southeastern spain by sector origins and cluster analysis," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D6, 2012. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD016885

[12] S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo, *Spatio-temporal clustering*. Boston, MA: Springer US, 2010, pp. 855–874. [Online]. Available: https://doi.org/10.1007/978-0-387-09823-4_44

[13] M. Y. Ansari, A. Ahmad, S. S. Khan, G. Bhushan *et al.*, "Spatiotemporal clustering: a review," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2381–2423, 2020.

[14] X. Wu, C. Cheng, R. Zurita-Milla, and C. Song, "An overview of clustering methods for geo-referenced time series: from one-way clustering to co- and tri-clustering," *International Journal of Geographical Information Science*, vol. 34, no. 9, pp. 1822–1848, 2020. [Online]. Available: https://doi.org/10.1080/13658816.2020.1726922

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[16] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[17] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[18] N. T. de Sousa, C. Trahms, P. Kröger, M. Renz, R. Schubert, and A. Biastoch, "Tracking the evolution of water flow patterns based on spatio-temporal particle flow clusters," in *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 246–253. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/MDM55031.2022.00054

[19] P. Kalnis, N. Mamoulis, and S. Bakiras, "On discovering moving clusters in spatio-temporal data," in *Advances in Spatial and Temporal Databases*. Springer Berlin Heidelberg, 2005, pp. 364–381.

[20] L. D. Talley, *Descriptive physical oceanography: an introduction*. Academic press, 2011.

[21] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

Fig. 9. Quarter yearly averages of the complete time period 1993-1997.

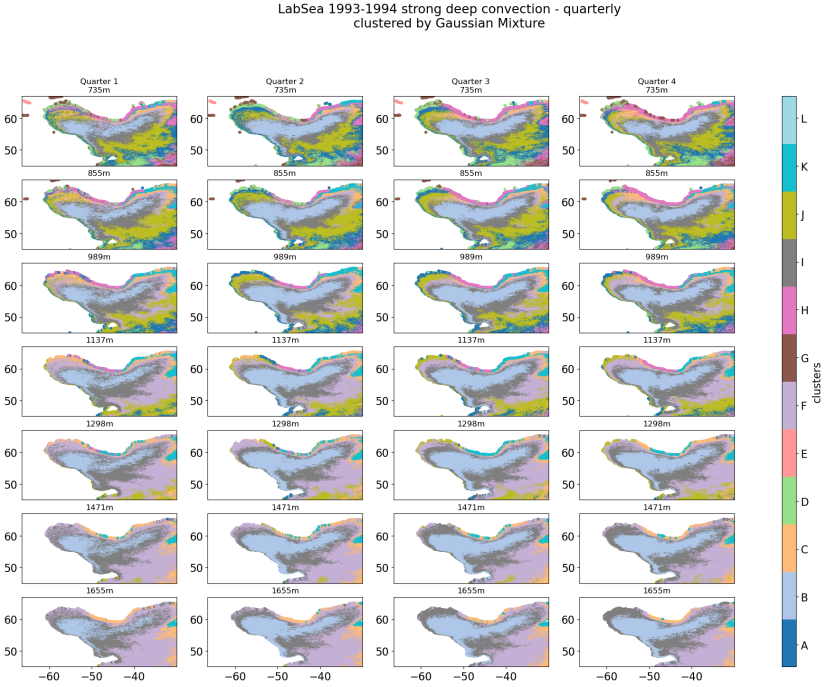

Fig. 10. Quarterly clustering of the strong deep convection period 1993-1994.

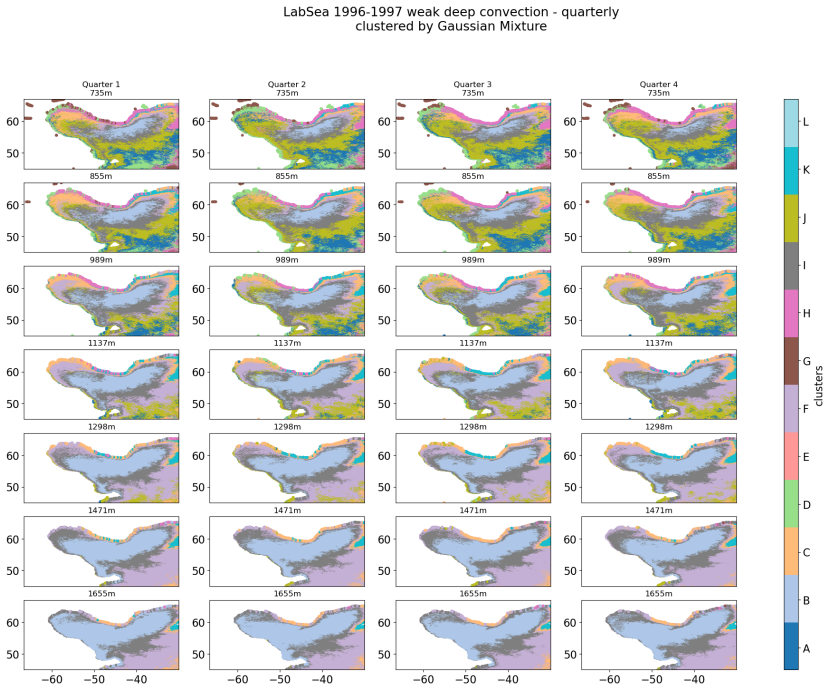

Fig. 11. Quarterly clustering of the weak deep convection period 1996-1997.

Paper E

Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters

Nelson Tavares de Sousa,
Carola Trahms, Peer Kröger, Matthias Renz, René Schubert,
Arne Biastoch

The paper has been published in
23rd IEEE International Conference on Mobile Data Management (MDM),
2022.

Tavares de Sousa, N., Trahms, C., Kröger, P., Renz, M., Schubert, R. and Biastoch, A. (2022), "Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters," 2022 23rd IEEE International Conference on Mobile Data Management (MDM), Paphos, Cyprus, 2022, pp. 246-253, DOI: https://doi.org/10.1109/MDM55031.2022.00054.

E. Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters

© 2022 IEEE. Reprinted, with permission, from Nelson Tavares de Sousa, Carola Trahms, Peer Kröger, Matthias Renz, René Schubert, and Arne Biastoch, "Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters", 23rd IEEE International Conference on Mobile Data Management (MDM), 06/2022.[1]

[1] In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Kiel University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters

Nelson Tavares de Sousa*, Carola Trahms*†, Peer Kröger*, Matthias Renz*, René Schubert†§, Arne Biastoch†‡
*Department of Computer Science, Christian-Albrechts-University, Kiel, Germany
{ntd,cat,pkr,mr}@informatik.uni-kiel.de
†GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany
‡Christian-Albrechts-University, Kiel, Germany
§Univ Brest, CNRS, IRD, Ifremer, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, Brest, France
{rschubert,abiastoch}@geomar.de

***Abstract*—Marine scientists investigate the movement of oceanic water particles with floating measurement devices released in the real ocean, as well as with virtual particles released in numerical model simulations. The detection, visualization, and evolution of clustered particles is key for gaining a comprehensive understanding of the underlying processes in the oceans. Thereby, vast amounts of mobility data (3D coordinates of these particles over time) need to be analyzed using mobility data science methods. In this paper, we describe the application of data science techniques to detect particle clusters and, more importantly, to track the evolution of these clusters over time in order to support the analysis of oceanic flows. In particular, we apply a well-known concept for tracking the cluster evolution from the data mining community that relies on pair-counting and, thus, is rather inefficient. In order to be applicable to large amounts of particles, we further elaborate two heuristic solutions to compute the cluster transitions based on spatial approximations. Experiments on real world data show a considerable speed-up while sacrificing marginal accuracy drops. Our prototype is used by domain experts for the analysis of the large-scale ocean by virtual particle release experiments in ocean simulations.**

***Index Terms*—data science, clustering, cluster evolution, spatiotemporal data**

I. Introduction

Ocean research is interested in the movement of water. This movement can be big lateral ocean currents such as the Gulf Stream, but also vortices - so called *eddies* - that move the water swirlingly into a broad direction. Both types of movement can be investigated by means of examining particles drifting in the water. This *Lagrangian* perspective follows respective particles along a *trajectory* from point A to point B (in contrast to the fixed *Eulerian* perspective) [1]. The analysis of their associated volume, temperature and salinity characteristics can explain changes in dynamics and hydrography that often are of climatic relevance. An example is the spreading of upper water masses from the southern hemisphere into the North Atlantic [2] or the deep return flow [3], both constituting the Atlantic Meridional Overturning Circulation (AMOC) that is important for the climate variability and climate change for the North Atlantic rim countries. Lagrangian techniques are also involved to describe the (nearly) passive spreading of organisms, such as marine animal larvae [4] or juveniles [5], or other drifting objects such as plastics [6] or larger debris (even up to parts from missing airplane MH370 [7]). The required input (velocities and hydrography) for Lagrangian calculations usually comes from ocean models [8], gridded satellite observations, or directly from surface drifter observations [9].

Lagrangian strategies require large numbers to provide robust statistics. $\mathcal{O}(10^5 - 10^6)$ particle trajectories are often computed based on ocean simulations. Moreover, the spatial and temporal resolution of the computed particle trajectories need to be sufficiently large to resolve the turbulent oceanic flows, the *eddies*, at the mesoscale (horizontal extends of $\mathcal{O}(10 - 100$ km)) or even the submesoscale ($\mathcal{O}(1$ km and below)) [10]. Mesoscale eddies can be detected on the basis of e.g their associated sea-surface height elevation. With respective detection methods and regridding of the particle densities onto regular maps, it could be shown that specific particles tend to siphon in eddies and form dense clusters (e.g. [9], [10]). Figure 1 shows us a visualization of a particle flow simulation around the southern tip of Africa. Water originating from the Agulhas Current east of Africa is carried toward the Atlantic Ocean, mainly by joining large eddies that are identifiable by their closed sea-surface height contours in this figure and are associated with higher particle transports (darker colors).

Fast visualization of clustered particles is needed to analyze and understand the underlying processes in the oceans. Additional information can be extracted from the evolution of the particle flows and the resulting eddies over time, requiring further methods to provide such functionality. However, the large amount of particles provide a significant challenge to handle such data efficiently enough to provide a fast and interactive tool to inspect the data. Clustering techniques, such as DBSCAN [11], allow to detect eddies in the water flow. However, they do not enable tracking over time instances because the identities (*ids*) of the clusters may vary. Figure 2 gives an example on how clusters may be identified over different time steps without consistent cluster ids. The order

This work has been partially conducted in the project FON 2020-24 funded by the KMS Kiel Marine Science - Centre for Interdisciplinary Marine Science at Kiel University. We acknowledge support by the Helmholtz School for Marine Data Science (MarDATA) funded by the Helmholtz Association (Grant HIDSS-0005). R.S. acknowledges funding from the French National Agency for Research (ANR) through the project DEEPER (ANR-19-CE01-0002-01).

E. Tracking the Evolution of Water Flow Patterns Based on Spatio-Temporal Particle Flow Clusters

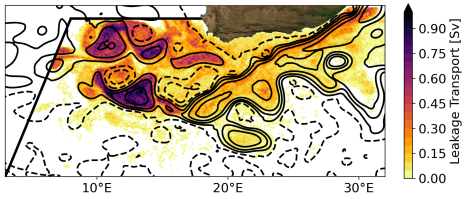

Fig. 1. A snapshot of the regridded water transport of particles released in the Agulhas Current east of Africa cut from the animation https://vimeo.com/609782202. Sea-surface height contours (black contours) mark the presence of the Agulhas Current, as well as of oceanic eddies (closed contours) that are associated with dense clusters of particles.

of identification (identity is depicted as color in this example) is not stable and therefore usually varies over time. Providing a stable identification of clusters, as required in this use case, is a separate task not yet covered by clustering algorithms. With focus on fast and efficient visualizations, the tracking of clusters must not introduce significant overhead in addition to the clustering itself.

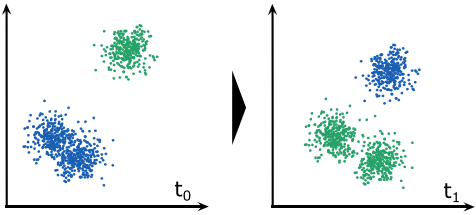

Fig. 2. DBSCAN over time. Note the different cluster assignments in each timestep.

In this paper we present techniques to track clusters which enable a stable visualization of such throughout a temporal domain. These techniques build on clustering algorithms and track their resulting clusters. We use DBSCAN to showcase and evaluate the presented techniques, but do not restrict the approaches to this specific algorithm.

In Section II we motivate and describe the basic data science techniques that we apply to the problem: density-based clustering and monitoring cluster evolution. In Section III we detail how these basic data science techniques are extended in order to support a large scale marine science application. In Section IV we evaluate our solutions on real-world marine data.

II. Preliminaries

There are two main requirements for the tracking algorithm. One, the cluster need to be detected in the first place. We showcase an example using DBSCAN, but our tracking approach is not limited to this clustering algorithm. The second requirement is to be able to identify the same clusters in successive time steps. We propose using MONIC [12] as a starting point. We will propose ways to mitigate the limitations of this approach in section III.

A. Cluster Detection Using DBSCAN

The first step in the analysis of the evolution of water flow patterns over time is to cluster the particles at each time step into groups that form eddies. This clustering is done according to the particles' coordinates that are represented as 3D points with longitude, latitude and depth. As it can be seen from Figure 1, the particle clusters representing eddies are dense point clouds of arbitrary size and shape. Not all points will be participating in eddies throughout all time slots, i.e., we expect some noise points. In addition, the patterns (clusters and noise) are highly dynamic, i.e., it is unlikely that we will have the same number of eddies throughout all time slots.

Density-based clustering algorithms, in particular DBSCAN [11], are capable of solving these challenges by detecting clusters of arbitrary size and shape, being robust against noise, and not requiring the number of clusters as input parameter. DBSCAN finds clusters as regions in the data space, where the data points are *dense*, separated by regions where the data points are sparse. The input parameters, a real number ε and a natural number $minPts$, specify the notion of density: a point that finds at least $minPts$ neighbors within radius ε (assuming a distance function among data points; we use the Haversine distance here) is considered to be within the core of a dense area and is called a "core point". Each of the neighbors within radius ε of a core point are in the same dense area (cluster), too, and each of them may be core point themselves. As long as we follow a transitive chain of core points that find themselves within radius ε, we never leave the dense area. DBSCAN computes these transitive chains starting at an arbitrary core point to compute each cluster separately. Picking suitable parameters ε and $minPts$ is not trivial but beyond the scope of this application paper. For our study, we manually "optimized" the parameters by running different settings and let domain experts selecting the scenario where the resulting clusters were "most meaningful" for them.

B. Monitoring Cluster Evolution Using MONIC

Having computed a set of clusters C_i for each time slot i, we need a method to keep track of their evolution, i.e., to track changes from the previous set C_{i-1} to the current C_i. We used MONIC [12] since it is model agnostic, i.e., it is independent of the clustering algorithm that produced the set C_j for any time slot j. It also offers a comprehensive and systematic definition of possible cluster transitions. In addition, even though it is based on comparing point ids when matching clusters, it is generalizable to a scenario where we do not have point ids.

MONIC maps a cluster $X \in C_i$ to one of the other clusters $Y \in C_{i-1}$, denoted by $f(X)$, based on the overlap of X and Y relative to X:

$$f(X) = \text{argmax}_{Y \in C_{i-1}} \frac{|X \cap Y|}{|X|},$$

i.e., $X \in C_i$ is mapped to the cluster $Y \in C_{i-1}$ that has the highest overlap amongst all clusters in C_{i-1} based on pairwise point matching. This requires that points have fixed ids over the entire time span. In case of ties, random mapping is performed such that the mapping is always unique. Based on this mapping, MONIC systematically defines several possible transitions for a given cluster X, including internal transitions related to the cluster itself (e.g. growth, shrinkage, etc.) as well as external transitions related to several clusters (e.g. merge, split, etc.).

III. Approximating Cluster Transitions

To track the evolution of a given cluster, we can use MONIC as described above to identify the same cluster in subsequent time steps. Since MONIC is based on pairwise point matching, this only works if we can uniquely identify data points over the entire life span. However, in particle flow analysis (as in many other spatiotemporal data applications) data points are not uniquely identifiable over all time slots, e.g. because the points are just snapshot observations and there is not clear matching of one observation at time slot $i-1$ to an observation at time slot i. As a consequence, w propose to extend MONIC such that the cluster matching between different time slots is approximated comparing clusters based on their spatial proximity. This allows us to identify the most similar cluster and to create a link between these clusters throughout time frames.

The following subsections introduce three different approaches we have investigated to implement such an extension.

A. Distance-based Pairwise Cluster Matching Approximations

Even though we cannot reliably match data points from different time slots in our data set using point ids, we can estimate that matching, which in the following will be our baseline approach.

With this information, we are able to determine the most similar cluster for any other given cluster in a similar way as in MONIC (see II-B). Equation 1 shows how to calculate this most similar cluster for a given cluster X.

$$f(X,\epsilon) = \text{argmax}_{Y \in C_{t-1}} \frac{|d(X,Y,\epsilon)|}{|X|} \tag{1}$$

C_i is a super set of clusters for a given time frame i. Furthermore, we assume $X \in C_t$. The function $d(X,Y,\epsilon)$ returns a subset of the data points of X, for which there is a neighboring data point in Y with a distance below ϵ. A ratio is determined by comparing the size of this resulting subset with the size of X. Eventually, $f(X,\epsilon)$ returns the cluster Y which shows the highest ratio of similarity to the cluster X. A filtering step based on a lower bound for the ratio allows to only consider results with a certain ratio or above.

The specific implementation of the function d may vary. For instance, different metrics for the distance calculation can be used. Furthermore, a pairwise distance calculation is relatively expensive with an execution time of $\mathcal{O}(n \cdot n^2)$. Optimized data indexing structures, such as K-D Trees, allow to further reduce the execution time to $\mathcal{O}(n \cdot log\ n)$ [13].

As this approach compares all data points within the clusters, it should in theory yield in the most precise results. However, some disadvantages emerge from this small scaled comparison. Depending on the value set for ϵ, noise will increase due to mismatched data points, especially at the border region of the cluster. If data points of two different clusters are within the distance of ϵ, these data points could be mismatched. This could become crucial for smaller clusters, as small amount of mismatched data points result in a higher variance in the resulting ratio.

B. Grid-based Clustering Matching Approximations

To circumvent the potentially high calculation times of the pairwise comparison, we also introduce a grid-based approach which allows to approximate the similarity between two clusters. For this, as depicted in Figure 3, we divide the space by applying a grid upon it. For each cluster, we determine the set of cells within the grid, which are at least partially occupied by data points of the cluster. This will be also applied to a different time frame, which yields in two different cell sets, which we can use for comparison. Figure 3 shows an example of this approach.

We see two different time frames with a moving cluster within. To the left side we see the first time frame. The relevant cells of the grid are marked red. To the right side we see the next time frame depicting the same cluster on a slightly different location due to its movement. By determining relevant cells, we can see that not all the same cells are covered. Some cells are lost due to the movement, recognisable due to the lesser amount of red cells. However, new cells are covered, as the movement of the cluster enters the region of not yet covered cells (green in Figure 3). To compare both clusters, we need to count all cells which are covered by both and set it into relation to the size of the cluster we want to track, as depicted by Equation 2.

$$f(X) = \text{argmax}_{Y \in C_{t-1}} \frac{|e(X) \cap e(Y)|}{|e(X)|} \tag{2}$$

The result of function $e(X)$ represents the subset of grid cells, which are covered by a cluster X. To continue with our example, we assume cluster X to be the cluster in the right time frame in Figure 3 and Y to be the cluster in the previous time frame to the left side. Therefore, $e(X)$ returns all cells colored in red and green on the right side. The red cells on the right time frame represent the intersection of cells covered by each cluster. As a result we get a coverage ratio of 0.5. For this approach, we also solely consider the cluster with the highest ratio and define a cut-off value afterwards to set a lower bound from which on we deem clusters to be matches.

The accuracy of this approach relies on the grid resolution. Cell sizes can be set arbitrarily and may not be uniform throughout all dimensions. If the resolution is chosen to be at the lower range of possible values, noise at the edges of clusters can be detected. For instance, let a grid cell be big

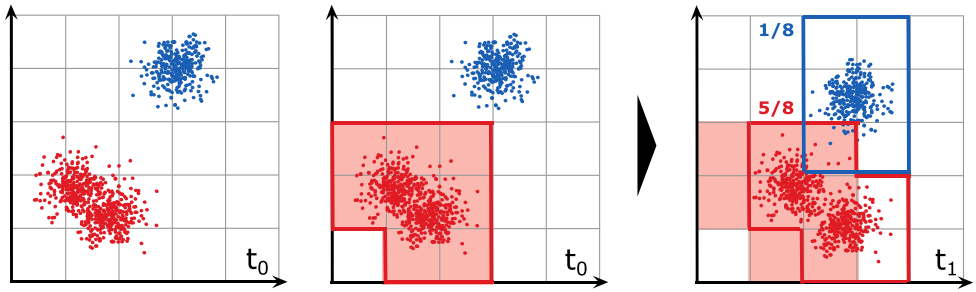

Fig. 3. Grid-based cluster matching.

enough to subsume a whole cluster. The number of covered cells by this cluster in this case will be 1. Any other cluster which covers this grid cell in some way will therefore yield in a ratio of 1, deeming this second cluster to be an incorrect match. However, higher resolutions result in higher memory consumption, though this can be estimated by $\mathcal{O}(n)$. All required operations can be run in $\mathcal{O}(n)$ or lower, leading to an overall estimate of $\mathcal{O}(n)$.

C. Centroid-based Clustering Matching Approximations

In our third approach we circumvent the difficulties of noise nearby cluster edges by analyzing the centroids of clusters. By only considering the centroid, we move the focus of this approach away from the edges of the cluster, reducing the interference between adjacent clusters.

As shown in Figure 4, we determine the centroid for each cluster and compare them between time frames based on their distance. The distance a cluster moves will be comparably smaller than the distances between two unlinked clusters. This can be represented by a threshold which gives an upper limit for the distance moved. Any movement beyond this threshold will be considered to be a different cluster. To eventually detect the correct cluster, we perform this calculation upon all clusters and search for the minimal distance traveled (see Equation 3).

$$f(X) = \text{argmin}_{Y \in C_{t-1}} \, d(c(X), c(Y)) \tag{3}$$

$c(X)$ returns the coordinates of the centroid of cluster X. $d(x, y)$ calculates the distance for two points x and y. Eventually, $f(X)$ searches for the cluster in the previous time frame which minimizes the distance traveled.

We further perform a filtering step including the aforementioned threshold to inhibit cluster mismatches.

We argue this approach shows similar behavior regarding memory consumption and execution time compared to the grid-based approach. Execution times will grow linearly with the input size, as the operations required can be performed in linear time. The same is valid for memory consumption, as in worst case, each data point forms an individual clusters, which again returns an individual centroid.

Although this approach reduces interference at cluster edges, mismatches cannot be phased out for all cases. For instance, if one cluster is surrounded by a second cluster, both will share similarly located centroids which could lead to mismatches.

IV. Evaluation

We conduct experiments to evaluate and compare the performance and accuracy of each approach. As experimental setup, we run each approach independently of each other for a total of five times each. We record the execution times for each run and calculate the mean value of all five runs. As a reference data set we extract one time frame from the particle transport simulation (see Figure 1) and track the clusters for the same time instance. We get a testing data set with 334,398 data points, out of which 54,805 data points compose 36 clusters. For the implementation, we rely on Python 3.9.7 and the libraries Scikit-Learn 1.0.1 and Numpy 1.20.3. Scikit-Learn provides some machine learning utilities, such as the calculation of nearest neighbors. As implementation of an indexing structure we use Scikit-Learn's KD Tree implementation.[1] The experiments are executed on an Apple M1 CPU with a total of 16 GB of memory.

A. Efficiency

One main goal is to avoid overhead in the execution time for the cluster tracking methods. Therefore, we implemented the various methods to compare these and eventually provide statements on the estimated execution times of each. The results of the measured execution times are shown in Figure 5. The upper plot shows a comparison between the four different algorithm types. Note the pseudo-logarithmic scale of the upper plot.

We see great differences in the execution times between the different algorithm types. The implementations of the grid-based and centroid-based algorithms outperform our baseline, the pairwise-distance-based algorithm, by far. Whereas

[1] https://scikit-learn.org/stable/modules/neighbors.html#k-d-tree

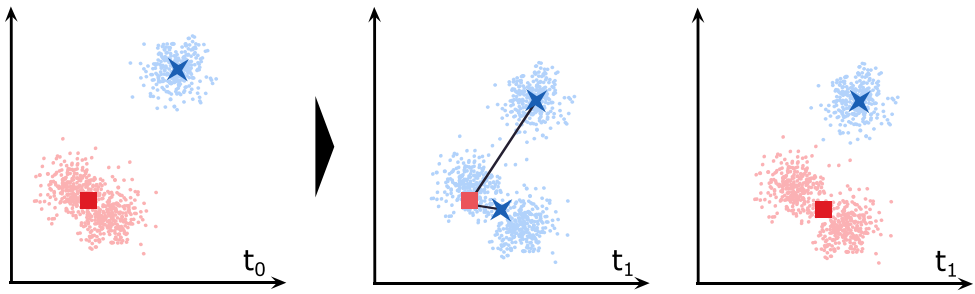

Fig. 4. Centroid-based cluster matching.

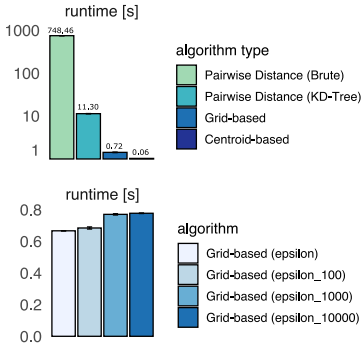

Fig. 5. Runtime for the different algorithms. The error bars indicate the standard *error*. Upper plot: runtime comparison between the groups of algorithms. The y axis is in pseudolog. Lower plot: runtime comparison for different sizes of ϵ.

B. Accuracy

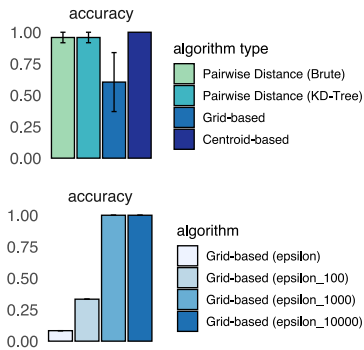

Fig. 6. Accuracies for the different algorithms. The error bars indicate the standard *error*. Upper plot: accuracy comparison between the groups of algorithms. Lower plot: accuracy comparison for different sizes of ϵ.

the pairwise-distance-based algorithm requires at least 11.3 seconds to calculate the cluster matches, the approximating algorithms require 0.72 and 0.06 seconds. This performance difference can be explained with the reduced search space used by both approaches. Comparing operations are only applied on cluster- or cell-level, while the pairwise-distance-based algorithm calculates distances between all data points. This is also reflected in the execution times of both pairwise-distance implementations. The *brute* implementation employs a naive pairwise comparison between all data points, resulting in a runtime estimated to be $\mathcal{O}(n^2)$. This estimate can be reduced by using *K-D Trees*, which effectively reduce the search space and this resulting in runtimes depicted as $\mathcal{O}(n \cdot log\ n)$. To investigate the impact of different grid resolutions of the grid-based algorithm, we also compared different parameters for this approach. The lower plot in Figure 5 shows the runtime comparison between them. We see differences of up to 15%, which in absolute terms make up a difference of 0.12s.

Our experiments were also conducted to show the accuracy of the different algorithms. To calculate the accuracy, we use a single time slice twice as our ground truth. The algorithms should correctly identify all clusters, as the clusters are effectively not moved through the simulated progress in time. The ratio between correctly identified clusters and the total amount of clusters represents our accuracy score. The results of this experiment are shown in Figure 6, again the upper plot compares the different algorithm types and the lower plot different grid sizes of the grid-based algorithm. We see high scores throughout all algorithms, but with an increased variance upon the grid-based algorithm. Depending on the upper bound for distance traveled, the pairwise-distance-based algorithm returns results with an accuracy between 91.7% and 100%. The centroid-based algorithm identifies clusters with a 100% accuracy. The variance of accuracy with the grid-based algorithm is a result of the different resolution parameters we used. Increasing the edge length, and therefore reducing the

resolution, leads to higher amount of mismatches, as more clusters reside may reside in the same grid cell. In such cases, the algorithm is not able to differentiate between the clusters and to identify them correctly.

C. Marine Science Application

As stated in section I, our real-world application is represented by a simulation of particle flows in the Agulhas current. Eddies are formed in this current which are detected as clusters by DBSCAN. We then apply our approaches for cluster tracking to comprehend the evolution of these clusters throughout time. To evaluate this application, we selected a time range of the simulation which contains a higher amount of clusters of varying sizes.

We choose a time range of four days, which allows to demonstrate the results in this paper. On average, 377366.5 particles are present each day, composing 378.5 clusters. Figure 7 shows the data and the corresponding clusters with different representation types. Each column represents a single day, following the chronological order from left to right. The first row shows the raw data which is a plot of each data point at its corresponding location. This is followed by a visualization of the distinct clusters which are detected by DBSCAN. The identity of a cluster is represented by its color. We can therefore see that clusters are not identified correctly throughout the time range, as their colors differ from one day to another. The same color may be used for different clusters as the color palette is limited to 20 colors.

Following the raw DBSCAN clusters visualizations, we see the results of each of our three algorithms. For the pairwise-distance-based algorithm, we expect a ratio of 0.9 or higher to consider clusters as a match. We see how clusters of bigger sizes are reliably identified throughout all days. However, due to the movement of the underlying particles, one cluster of bigger size cannot be matched correctly on day 3.

Similar behavior can be observed with the centroid-based algorithm, whereas different clusters are affected. For this algorithm, we choose as an upper bound for the distance between the centroids a value 400 times the ε value used by DBSCAN. Any value below this returns worse results further showing the great amount of movement between the days.

The grid-based approach correctly matches the biggest clusters, but fails to correctly match mid-sized clusters. However, in direct comparison with the two aforementioned approaches, the grid-based algorithm shows the best performance for our real-world example.

Erroneous matches upon small clusters occur regardless of the approach. This is especially the case for smaller clusters at the edges of cluster of bigger sizes. Figure 8 shows a cropped image of a region including smaller clusters. The distances travelled by particles are big enough such that either clusters appear or disappear, or simply cannot be reliably identified, as the distance traveled does not allow to infer this information. We argue that the data provided by this use-case is not suitable to track the evolution of small clusters. However, we clearly see the benefits of our approaches with clusters of greater sizes, which comparably move shorter distances between two time steps.

V. Related Work

Kalnis et al. [14] introduced a concept to determine the similarity between clusters. Counting the amount of identifiable data points between to clusters allows to measure the similarity. Enhanced approaches are also presented to speed-up the cluster tracking over time. However, the shown approaches rely on data points to be uniquely identifiable, which is not the case for our real-world example. Therefore, we adapt the presented approach to allow for a defined margin of error, which takes movement into consideration, to detect a data point in different points in time. In addition, we introduced two different approximation methods, which allow for faster calculation of cluster similarities.

The Moving Micro-Clustering (MMC) algorithm allows to track microclusters by approximating them with bounding boxes [15]. Boxes provide an efficient way of calculating approximations and can be used as guide in the next time slice. However, this approach is only applied on microcluster level. We focus on the application on cluster level, to track directly the evolution of clusters. As a faster tracking option, we introduced a grid-based approach, which also exploits the faster hashing operations coming from rectangles.

ChronoClust employs the concept of microclusters and introduces the centroid-based tracking of clusters to the micro-cluster level [16]. Moving clusters are tracked by analysing the involved microcluster in each time step. ChronoClust applies two techniques to track moving clusters, either by lineage or historical proximity. The lineage approach determines a moving cluster by comparing microclusters. Clusters in different time steps will be linked, if those two clusters share a certain amount of microclusters. This requires an already existing microclustering. Historical proximity describes the location-based linking, by searching for for the nearest constituent microcluster of the previous time step. The cluster including this microcluster will then be considered as a match. We employ similar techniques with our centroid-based approach, but perform such on cluster level, as we argue microclustering to not be necessary for our real-world use case.

In the oceanographic community, there are alternative methods to identify and track coherent eddies based on e.g. Lagrangian rotational coherence [17], [18]. It would be interesting to compare such approaches to the clustering presented here.

VI. Conclusion and Future Work

In this paper, we presented different approaches to track clusters throughout a temporal domain and applied these approaches on a real-world application for marine sciences. This allows marine researchers to gain more insights into the water currents and to perform further analyses on particle movements throughout time. We introduced three approaches, one providing a sophisticated comparison method and two using approximation techniques to match and track clusters

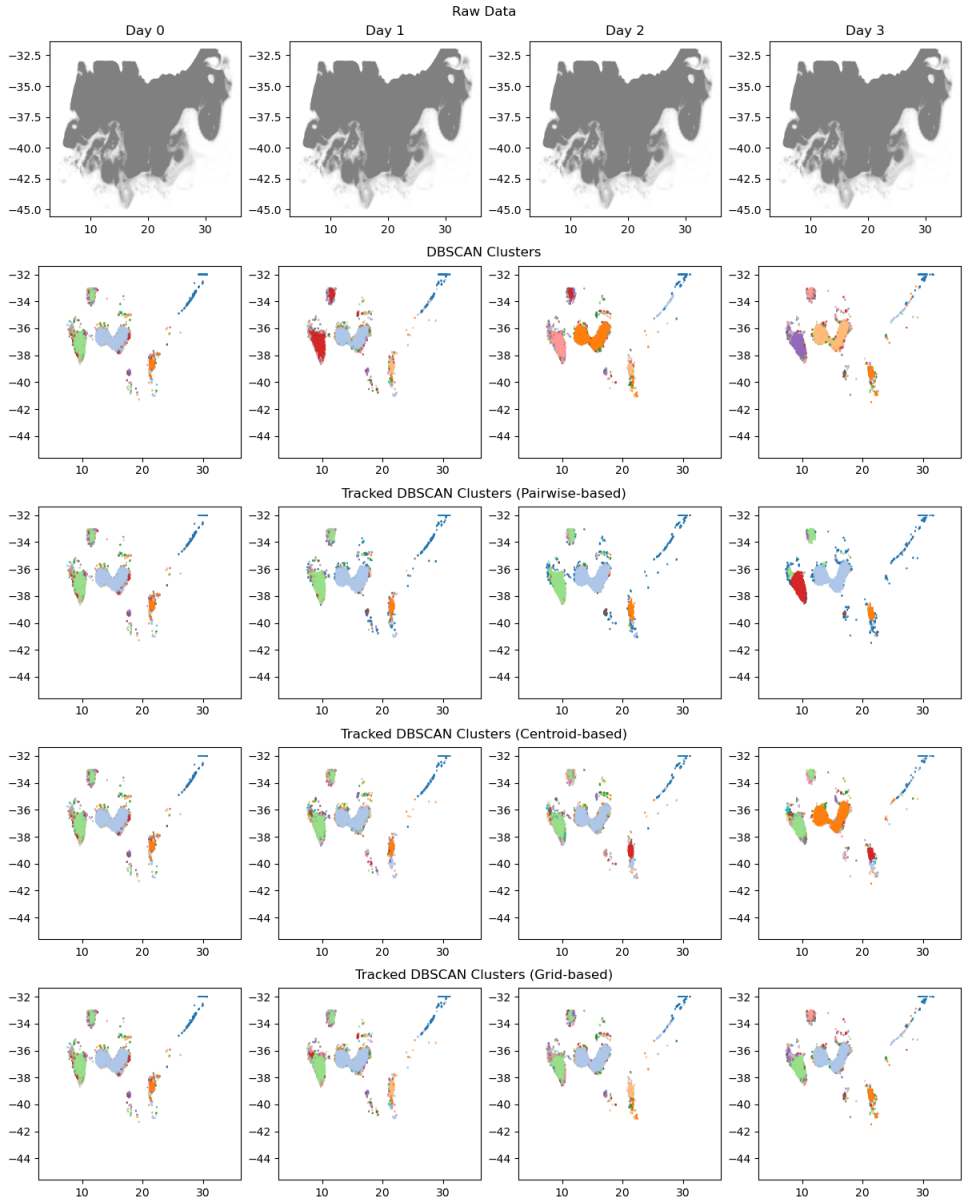

Fig. 7. Visualization of the application data with different representation types including raw data, clustered data, and tracked clustered data from our approaches.

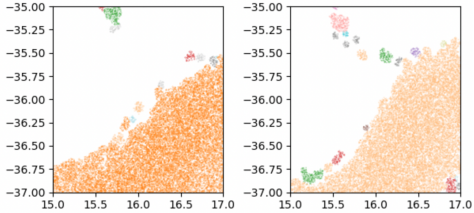

Fig. 8. Cropped excerpt of the real world example. The excerpt is taken from days 2 and 3 of the DBSCAN clusters data set.

through time. An evaluation of the approaches regarding execution times and accuracy however, shows that the latter can be at least as accurate as the first approach, while requiring less than 1% of the original execution of the first approach. An application of the approaches on a real-world example shows that these can be used as a reliable tracking method for clusters of greater sizes, as the movement of particles in the data prohibits the reliable use of our approaches for smaller clusters. However, we argue that applications showing data with smaller deltas between two time steps will be able to make use of our approach.

Future work will include more information for matching purposes, such as direction and speed of clusters. Including this information reduces the search area, resulting in more accurate algorithms and reducing the margin of error. Furthermore, enhanced methods to track cluster evolution, such as in ChronoClust or MMC [15], [16], can also be applied to our approaches, further enhancing the insights to each individual cluster. Adapting such approaches allows to detect splitting or joining clusters and to trace data points entering and leaving each cluster. As the introduced algorithms are agnostic regarding the calculation method of the clusters, a comparison of different clustering algorithms (e.g. OPTICS, KNN) may reveal benefits by not relying solely on DBSCAN.

The approaches introduced in this work are suitable for other applications as well. The methods shown here are not limited to two-dimensional data, but can also be applied to data of arbitrary number of dimensions.

VII. Data Availability

A python implementation of the presented algorithms (including the test data used in the evaluation) is available as open source at https://github.com/cau-kiel-ai/py-cluster-tracking.

References

[1] E. van Sebille, S. M. Griffies, R. Abernathey, T. P. Adams, P. Berloff, A. Biastoch, B. Blanke, E. P. Chassignet, Y. Cheng, C. J. Cotter, E. Deleersnijder, K. Döös, H. F. Drake, S. Drijfhout, S. F. Gary, A. W. Heemink, J. Kjellsson, I. M. Koszalka, M. Lange, C. Lique, G. A. MacGilchrist, R. Marsh, C. G. Mayorga Adame, R. McAdam, F. Nencioli, C. B. Paris, M. D. Piggott, J. A. Polton, S. Rühs, S. H. Shah, M. D. Thomas, J. Wang, P. J. Wolfram, L. Zanna, and J. D. Zika, "Lagrangian Ocean Analysis: Fundamentals and Practices," pp. 49–75, 2018.

[2] S. Rühs, F. U. Schwarzkopf, S. Speich, and A. Biastoch, "Cold vs. warm water route-sources for the upper limb of the Atlantic Meridional Overturning Circulation revisited in a high-resolution ocean model," *Ocean Sci.*, vol. 15, no. 3, pp. 489–512, 2019.

[3] T. Schulzki, K. Getzlaff, and A. Biastoch, "On the Variability of the DWBC Transport Between 26.5°N and 16°N in an Eddy-Rich Ocean Model," *J. Geophys. Res. Ocean.*, vol. 126, no. 6, p. e2021JC017372, jun 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1029/2021JC017372

[4] S. F. Gary, A. D. Fox, A. Biastoch, J. M. Roberts, and S. A. Cunningham, "Larval behaviour, dispersal and population connectivity in the deep sea," *Sci. Rep.*, vol. 10, no. 1, p. 10675, dec 2020. [Online]. Available: http://www.nature.com/articles/s41598-020-67503-7

[5] M. Baltazar-Soares, A. Biastoch, C. Harrod, R. Hanel, L. Marohn, E. Prigge, D. Evans, K. Bodles, E. Behrens, C. W. Böning, and C. Eizaguirre, "Recruitment collapse and population structure of the european eel shaped by local ocean current dynamics," *Curr. Biol.*, vol. 24, no. 1, pp. 104–108, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982213014504

[6] E. Van Sebille, S. Aliani, K. L. Law, N. Maximenko, J. M. Alsina, A. Bagaev, M. Bergmann, B. Chapron, I. Chubarenko, A. Cózar, P. Delandmeter, M. Egger, B. Fox-Kemper, S. P. Garaba, L. Goddijn-Murphy, B. D. Hardesty, M. J. Hoffman, A. Isobe, C. E. Jongedijk, M. L. Kaandorp, L. Khatmullina, A. A. Koelmans, T. Kukulka, C. Laufkötter, L. Lebreton, D. Lobelle, C. Maes, V. Martinez-Vicente, M. A. Morales Maqueda, M. Poulain-Zarcos, E. Rodríguez, P. G. Ryan, A. L. Shanks, W. J. Shim, G. Suaria, M. Thiel, T. S. Van Den Bremer, and D. Wichmann, "The physical oceanography of the transport of floating marine debris," p. 023003, feb 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1748-9326/ab6d7d

[7] "Strategies for simulating the drift of marine debris," *J. Oper. Oceanogr.*, vol. 14, no. 1, pp. 1–12, apr 2021. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1755876X.2019.1602102

[8] A. Biastoch, F. U. Schwarzkopf, K. Getzlaff, S. Rühs, T. Martin, M. Scheinert, T. Schulzki, P. Handmann, R. Hummels, and C. W. Böning, "Regional imprints of changes in the Atlantic Meridional Overturning Circulation in the eddy-rich ocean model VIKING20X," *Ocean Sci.*, vol. 17, pp. 1177–1211, sep 2021. [Online]. Available: https://os.copernicus.org/articles/17/1177/2021/

[9] C. Vic, S. Hascoët, J. Gula, T. Huck, and C. Maes, "Oceanic mesoscale cyclones cluster surface lagrangian material," *Geophysical Research Letters*, p. e2021GL097488.

[10] R. Schubert, J. Gula, and A. Biastoch, "Submesoscale flows impact Agulhas leakage in ocean simulations," *Commun. Earth Environ. 2021 21*, vol. 2, no. 1, pp. 1–8, sep 2021. [Online]. Available: https://www.nature.com/articles/s43247-021-00271-y

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[12] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "MONIC," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*. ACM Press, 2006.

[13] P. Ram and K. Sinha, "Revisiting kd-tree for nearest neighbor search," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2019.

[14] P. Kalnis, N. Mamoulis, and S. Bakiras, "On discovering moving clusters in spatio-temporal data," in *Advances in Spatial and Temporal Databases*. Springer Berlin Heidelberg, 2005, pp. 364–381.

[15] Y. Li, J. Han, and J. Yang, "Clustering moving objects," in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. ACM Press, 2004.

[16] G. H. Putri, M. N. Read, I. Koprinska, D. Singh, U. Röhm, T. M. Ashhurst, and N. J. King, "ChronoClust: Density-based clustering and cluster tracking in high-dimensional time-series data," *Knowledge-Based Systems*, vol. 174, pp. 9–26, jun 2019.

[17] G. Haller, "Dynamic rotation and stretch tensors from a dynamic polar decomposition," *Journal of the Mechanics and Physics of Solids*, vol. 86, pp. 70–93, 2016.

[18] N. Tarshish, R. Abernathey, C. Zhang, C. O. Dufour, I. Frenger, and S. M. Griffies, "Identifying lagrangian coherent vortices in a mesoscale ocean model," *Ocean Modelling*, vol. 130, pp. 15–28, 2018.