The background of the slide is a complex, abstract pattern of thin, overlapping contour lines. These lines are colored in a variety of shades including light blue, teal, green, yellow, and purple, creating a marbled or topographic effect. The lines are most dense in the lower half of the image, where they form larger, more irregular shapes, and become sparser towards the top.

Large-scale structure extraction in Lagrangian ocean trajectories

David Wichmann

Large-scale structure extraction in Lagrangian ocean trajectories

David Wichmann

Institute for Marine and Atmospheric research Utrecht (IMAU)
Department of Physics, Utrecht University
Princetonplein 5, 3584 CC Utrecht, The Netherlands

ISBN: 978-94-6332-736-7

Printed by GVO drukkers & vormgevers B.V.

Large-scale structure extraction in Lagrangian ocean trajectories

**Grootschalige Structuurdetectie in Lagrangiaanse
ocean trajecten**
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 16 juni 2021 des ochtends te 10.15 uur

door

David Wichmann

geboren op 25 mei 1990
te Balingen, Duitsland

Promotor:

Prof. dr. H.A. Dijkstra

Copromotor:

Dr. E. van Sebille

This work was conducted within the TOPIOS (Tracking Of Plastic In Our Seas) project and supported through funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement No 715386).

Contents

Summary	vii
Samenvatting	ix
1 Introduction	1
1.1 Motivation.	1
1.2 Lagrangian ocean transport	2
1.3 Challenges for Lagrangian data analysis	4
1.4 Macroscopic features of the ocean flow at different scales . . .	6
1.4.1 Basin-scale mixing	6
1.4.2 Transport barriers	8
1.4.3 Mesoscale eddies	8
1.5 Goals and research questions of this thesis	9
2 Influence of Near-Surface Currents on the Global Dispersal of Marine Microplastic	11
2.1 Introduction.	12
2.2 Methods	15
2.3 Results	17
2.3.1 Particle distributions	17
2.3.2 Regional transport	18
2.4 Conclusion.	22
2.5 Supplementary material to chapter 2	25
3 Mixing of passive tracers at the ocean surface and its implications for plastic transport modelling	35
3.1 Introduction.	36
3.2 Methods	38
3.2.1 Entropy of mixing	38
3.2.2 Markov chain mixing	39
3.2.3 Division of the ocean surface into basins.	42
3.3 Results	43
3.3.1 Entropy of mixing	43
3.3.2 Markov chain mixing	44
3.4 Conclusion.	48
3.5 Supplementary material to chapter 3	51
4 Detecting flow features in scarce trajectory data using networks derived from symbolic itineraries: an application to surface drifters in the North Atlantic	73
4.1 Introduction.	74

4.2	Drifter data set	76
4.3	Methods	76
4.3.1	Preliminaries	76
4.3.2	Definition of the network	77
4.3.3	Normalized cut and spectral relaxation	79
4.3.4	Comparison to existing methods	82
4.4	Results	84
4.4.1	Periodically driven double-gyre flow	84
4.4.2	Surface drifters in the North Atlantic	87
4.5	Conclusion.	88
4.6	Appendices	90
4.6.A	Relation of Q to almost-invariant sets with the transfer operator	90
4.6.B	Supplementary figures double-gyre flow	91
4.6.C	Supplementary figures ocean drifters	95
5	Ordering of trajectories reveals hierarchical finite-time coherent sets in Lagrangian particle data: detecting Agulhas rings in the South Atlantic Ocean	97
5.1	Introduction.	98
5.2	Trajectory datasets.	100
5.1	Quasi-periodically perturbed Bickley jet	100
5.2	Agulhas rings in the South Atlantic	101
5.3	Methods	101
5.1	Detecting coherent structures in Lagrangian trajectory data	101
5.2	Trajectory embedding	103
5.3	Clustering with OPTICS	105
5.4	Comparison to related methods	107
5.4	Results	109
5.1	Bickley jet flow	109
5.2	Agulhas rings	112
5.5	Conclusion.	114
5.6	Appendices	116
5.6.A	Additional figures for the Bickley jet flow	116
5.6.B	Additional figures for the Agulhas flow.	118
5.6.C	Detecting Agulhas rings with a particle-based network	119
6	Conclusion and outlook	121
6.1	Research questions and answers.	121
6.2	Future perspectives	122
	Bibliography	125
	Acknowledgement	135
	Curriculum Vitæ	137

Summary

The ocean currents constantly transport and redistribute large amounts of tracers such as heat, salt or plastic on the global scale. Understanding this large-scale dispersal of tracers in the ocean is critical for many topics in oceanography. In this thesis, we study the meso- and large-scale features of the ocean flow in a Lagrangian framework, i.e. with trajectories that represent the motion of individual tracer parcels. Due to the chaotic behaviour of trajectories in a nonlinear system such as the ocean flow, single trajectories have only very limited meaning, as they strongly depend on the precise initial conditions. At the same time, on the macroscopic scale, the ocean flow often possesses a certain form of order that is less sensitive to the initial conditions of particles, and is reflected by the collective and probabilistic behaviour of trajectories for intermediate and long times. The different chapters in this thesis present, in different contexts, methods to extract such macroscopic properties from sets of Lagrangian ocean trajectories, covering both measured and numerically simulated ocean trajectories.

This thesis was conducted as part of the TOPIOS project (Tracking Of Plastic In Our Seas). Chapters 2 and 3 thus aim to contribute to a better understanding of plastic transport and modelling in the ocean. In chapter 2, we show that the well-known accumulation pattern of buoyant plastic in the centres of the five major ocean basins, the so-called Subtropical Gyres, changes significantly when particles are transported not at but close to the surface. Although it is known that buoyant plastic particles can be submerged to deeper layers by different mechanisms, the impact on the global accumulation pattern had not been studied so far. We show that near-surface transport can lead to an increased export of plastic from subtropical to polar regions compared to surface transport, providing a potential explanation for the discovery of microplastic particles in polar seas. The results of chapter 2 show that a surface-only view of microplastic transport in the ocean might be misleading. It is therefore important to develop more realistic transport models that accurately reflect the actual dynamics of plastic in the ocean, including the vertical direction.

Chapter 3 focuses on basin-scale mixing of tracers at the ocean surface, which is a direct consequence of the converging Ekman currents in the centres of the Subtropical Gyres. Using concepts from Markov chain theory, our results show that the correlation between initial and final particle distributions in each subtropical basin decays on time-scales that are relevant for plastic transport modeling. Our results imply that the backtracking of particles at the surface ocean is not meaningful for times beyond a certain mixing time scale any more, which is typically around 10 years for most of the ocean. The results of chapter 3 demonstrate that a straight-forward interpretation of long modelled particle trajectories is not possible at the surface ocean, and, more generally, that fu-

ture particle-based models of plastic transport in the ocean should recognize the limited value of long modeled trajectories in a highly chaotic system.

Chapters 4 and 5 focus more generally on data-driven techniques to extract large-scale structure from sets of measured and modelled particle trajectories. Important concepts in both chapters are the reduction of the dimensionality of Lagrangian particle data, and data clustering. In chapter 4, we define a network based on measured drifter trajectories to detect quasi-stationary features in the ocean flow. The network is built upon concepts from symbolic dynamics, and takes into account the quasi-stationarity of the features to be detected. We successfully detect the major structures in the North Atlantic Ocean, such as the Western Boundary Current region, and the transport barrier between the Subpolar and Subtropical Gyres. The method presented in chapter 4 makes use of the correspondence between an individual particle trajectory defined on a continuous space, and its symbolic itinerary that is defined on a discrete space. In particular, the method allows one to define a clustering problem on the discrete, low-dimensional space, while the result of the clustering problem can be mapped to the initial continuous space. This allows the extraction of features at scales below the coarse graining scale.

Finally, in chapter 5, we show how density-based clustering can be used to overcome problems of partition-based clustering for the detection of coherent sets of particles, such as eddies and jets, in networks that are derived from Lagrangian trajectories. In the chapter, we successfully detect Agulhas rings in the South Atlantic Ocean within an ocean model, and show how a purely data-driven application of an established clustering technique can outperform many recent algorithms. A main conclusion of chapter 5 is that future research should focus more on exploring established data analysis techniques to identify large-scale structure in Lagrangian ocean trajectories.

Samenvatting

De oceaanstromingen transporteren en herverdelen voortdurend grote hoeveelheden tracers zoals warmte, zout of plastic op wereldschaal. Inzicht in deze grootschalige verspreiding van tracers in de oceaan is van cruciaal belang voor veel onderwerpen in de oceanografie. In dit proefschrift bestuderen we de meso- en grootschalige kenmerken van de oceaanstroming in een Lagrangiaans kader, d.w.z. met trajecten die de beweging van individuele tracers weergeven. Vanwege het chaotische gedrag van trajecten in een niet-lineair systeem zoals de oceaanstroming, hebben afzonderlijke trajecten slechts een zeer beperkte betekenis, omdat ze sterk afhankelijk zijn van de precieze beginvoorwaarden. Tegelijkertijd bezit de oceaanstroming op macroscopische schaal vaak een bepaalde vorm van orde die minder gevoelig is voor de beginvoorwaarden van deeltjes, en die tot uiting komt in het gezamenlijke en probabilistische gedrag van trajectoriën op de middellange en lange termijn. De verschillende hoofdstukken in dit proefschrift presenteren, in verschillende contexten, methoden om dergelijke macroscopische eigenschappen te extraheren uit verzamelingen Lagrangiaanse oceaantrajecten, zowel van gemeten als numeriek gesimuleerde oceaantrajecten.

Dit proefschrift is geschreven in het kader van het TOPIOS project (Tracking Of Plastic In Our Seas). De hoofdstukken 2 en 3 hebben tot doel bij te dragen aan een beter begrip van plastic transport en modellering in de oceaan. In hoofdstuk 2 laten we zien dat het bekende accumulatiepatroon van drijvend plastic in de centra van de vijf grote oceaanbekkens, de zogenaamde Subtropische Gyres, aanzienlijk verandert wanneer deeltjes dieper onder het oppervlak worden getransporteerd. Hoewel bekend is dat drijvende plasticdeeltjes door verschillende mechanismen naar diepere lagen kunnen worden getransporteerd, was het effect op het globale accumulatiepatroon tot nu toe niet bestudeerd. Wij laten zien dat transport op enkele tientallen meters onder het oppervlak kan leiden tot een verhoogde export van plastic van subtropische naar polaire gebieden in vergelijking met transport aan het oppervlak, wat een potentiële verklaring biedt voor de observatie van microplastic deeltjes in polaire zeeën. De resultaten van hoofdstuk 2 laten zien dat het alleen in beschouwing nemen van het transport van microplastic aan het oceaanoppervlak misleidend kan zijn. Het is daarom belangrijk om meer realistische transportmodellen te ontwikkelen die de werkelijke dynamiek van plastic in de oceaan nauwkeurig weergeven, inclusief de verticale richting.

Hoofdstuk 3 richt zich op menging van tracers aan het oceaanoppervlak, een direct gevolg van de convergerende Ekman stromingen in de centra van de Subtropische Gyres. Met behulp van concepten uit de Markov chain theorie tonen onze resultaten aan dat de correlatie tussen initiële en uiteindelijke deeltjesdistributies in elk subtropisch bekken afneemt op tijdschalen die relevant zijn voor

modellering van plastic transport. Onze resultaten impliceren dat het backtracen van deeltjes aan het oppervlak van de oceaan niet meer zinvol is voor tijden voorbij een bepaalde mengtijschaal, die typisch rond de 10 jaar ligt voor het grootste deel van de oceaan. De resultaten van hoofdstuk 3 tonen aan dat een simpele interpretatie van lange gemodelleerde deeltjestrajecten niet mogelijk is aan het oppervlak van de oceaan, en, meer generiek, dat toekomstige deeltjes-gebaseerde modellen van plastic transport in de oceaan de beperkte waarde van lange gemodelleerde trajecten in een zeer chaotisch systeem moeten erkennen.

De hoofdstukken 4 en 5 richten zich meer in het algemeen op data-gedreven technieken om grootschalige structuur te extraheren uit sets van gemeten en gemodelleerde deeltjestrajecten. Belangrijke concepten in beide hoofdstukken zijn de reductie van de dimensionaliteit van Lagrangiaanse deeltjes data, en data clustering. In hoofdstuk 4 definiëren we een netwerk gebaseerd op gemeten drifter trajecten om quasi-stationaire kenmerken in de oceaanstroming te detecteren. Het netwerk is gebaseerd op concepten uit de symbolische dynamica, en houdt rekening met de quasi-stationariteit van de te detecteren kenmerken. We detecteren met succes de belangrijkste structuren in de Noord-Atlantische Oceaan, zoals de Westelijke Grensstroomregio en de transportbarrière tussen de Subpolaire en Subtropische Gyres. De methode die in hoofdstuk 4 wordt gepresenteerd, maakt gebruik van de overeenkomst tussen een individueel deeltjestraject gedefinieerd op een continue ruimte, en de symbolische route ervan die is gedefinieerd op een discrete ruimte. De methode maakt het mogelijk om een clusterprobleem te definiëren en op te lossen op de discrete, laag-dimensionale ruimte, terwijl het resultaat van het clusterprobleem in kaart kan worden gebracht op de initiële continue ruimte. Dit maakt het mogelijk om kenmerken te extraheren op kleinere schalen dan de coarse graining schaal.

Tenslotte laten we in hoofdstuk 5 zien hoe een op dichtheid gebaseerde clustering kan worden gebruikt om problemen op te lossen die partitie-gebaseerde clustering hebben bij de detectie van samenhangende sets van deeltjes, zoals wervelingen en stromingen, in netwerken die zijn afgeleid van Lagrangiaanse trajecten. In dit hoofdstuk detecteren we met succes Agulhasringen in de Zuid-Atlantische Oceaan binnen een oceaanmodel, en laten we zien hoe een puur data-gestuurde toepassing van een gevestigde clusteringstechniek veel recente algoritmen kan overtreffen. Een belangrijke conclusie van hoofdstuk 5 is dat toekomstig onderzoek zich meer zou moeten richten op het verkennen van gevestigde data-analysetechnieken om grootschalige structuur in Lagrangiaanse oceaantrajecten te identificeren.

1

Introduction

1.1. Motivation

The ocean currents constantly transport and redistribute large amounts of tracers such as heat, salt, nutrients and plastic¹. Tracers in the ocean can be passively transported by the currents, or influenced by other mechanisms such as active swimming of fish [1] or biofouling related sinking of plastic [2]. Tracer specific transport mechanisms lead to different temporal and spatial scales on which transport takes place, or on which transport is of interest. For example, the time fish larvae are transported by the ocean currents is restricted by the larvae settling at a certain location, which typically takes place after less than a month [3]. In contrast, for the accumulation of plastic in the centres of the Subtropical Gyres², time scales of several years to a decade are important [4]. On the microscopic level, a small tracer parcel exposed to the ocean currents is transported by a time-dependent and often chaotic flow. This makes it difficult to predict the pathway of such a parcel, in particular for long times. On the larger scale, however, the ocean flow can possess a certain form of order that is less sensitive to the underlying microscopic chaos. Such order is reflected by the collective and probabilistic behaviour of sets of tracer parcels for intermediate and long times. Chapters 2-5 present, in different contexts, methods to extract such macroscopic properties from sets of Lagrangian ocean trajectories.

Understanding the distribution of tracers in the ocean is important for many topics in oceanography. An important example of ocean tracers are heat and salt, which are influenced by freshwater and heat fluxes at the ocean surface, and which almost entirely determine the density of sea water. Resulting density gradients in turn affect the large-scale ocean circulation, creating the so-called

¹In the following, we use the word ‘tracer’ to include both classic tracers like heat and salt, as well as particulates like plastic or fish larvae.

²The Subtropical Gyres are basin-scale circular flows in the five major ocean basins, i.e. in the North Pacific, South Pacific, North Atlantic, South Atlantic and the Indian Ocean.

thermohaline part of this circulation [5]. Changes in precipitation or enhanced melting of polar ice under climate change have an impact on the distribution of heat and salt in the ocean. Such changes can have an effect on the large-scale ocean circulation, which in turn affects the distribution of heat and salt. This interaction between heat and salt and the ocean currents leads to a range of mathematically interesting phenomena such as positive feedback loops and bistability, which are active subjects of research in physical oceanography (see e.g. [6]). Understanding the distribution and transport of these tracers in the ocean is crucial to understand potential consequences of climate change.

Marine plastic is another, yet human-made, tracer that has received significant public and scientific interest in the past years. Entering the ocean from different sources such as cities, river mouths or ships, plastic can be transported by the ocean currents, while being subject to a multitude of other biophysical and chemical mechanisms such as biofouling, degradation and weathering [7]. Despite recent efforts, the transport pathways and the fate of plastic in the ocean are still largely not understood. Gaining a better understanding of the dispersal of plastic in the ocean is important to assess the risk plastic poses for marine ecosystems, but also to identify potential pathways of plastic into the human food system.

1.2. Lagrangian ocean transport

The distribution of tracers in the ocean as a function of time can be studied within two frameworks. The Eulerian perspective on tracer transport describes the evolution of a tracer density $\rho(x, t)$ as a function of time t , defined on the ocean domain Ω , with $x \in \Omega$. In the case of no diffusion, the time evolution of the tracer density $\rho(x, t)$ is described by the continuity equation:

$$\frac{\partial \rho(x, t)}{\partial t} = -\nabla \cdot (v(x, t)\rho(x, t)), \quad (1.1)$$

where $v(x, t)$ is the fluid velocity, which is governed by the Navier-Stokes equations. The Eulerian perspective comes naturally within fluid dynamical simulations on discretized spatial domains, where discretized fields such as density and surface wind stress force the ocean circulation. The Lagrangian perspective describes how individual tracer *parcels*, i.e. small concentrated amounts of a tracer, spread across the domain Ω , i.e. the object of study are the position vectors of such parcels, $x(t) \in \Omega$, as a function of time. The Lagrangian approach is particularly useful in situations where the actual tracer pathway is of interest, for example to study the connectivity of different regions in the ocean for plankton [8], or the crossing of oceanographic fronts by surface tracers [9]. The Lagrangian viewpoint has also been useful in understanding the pathways and time scales of the global transport of water masses, and thus heat and salt [10], which in turn determine parts of the large-scale ocean circulation.

There are essentially two approaches to study Lagrangian ocean transport: by

analysing drifter data, or with numerically integrated virtual particles. The NOAA Global Drifter Program³ is an ongoing project of surface drifters sampling the ocean velocity in the upper water column [11]. Drifter data has been used to compute the mean flow and other climatological quantities of the surface ocean [12], see LaCasce [13] for an account of the statistical analysis of drifter data. Drifters have also been used to estimate a time-averaged form of the transfer operator (cf. section 1.4.1), the transition matrix, in the context of surface plastic transport modeling [4, 14]. In chapter 4, we analyse ocean surface trajectories derived from drifter experiments to detect the large-scale flow structures of the North Atlantic Ocean.

Despite the growing number of drifters released in the ocean - in total more than 20,000 drifters have been released in the last decades - the coverage of sampling points in space and time is still too low for many applications. For example, it is hard to study mesoscale time-dependent phenomena such as eddies with ocean drifters, simply because eddies are almost never sufficiently sampled by the drifters. The spatial and temporal inhomogeneity of the data in addition causes problems in deriving unbiased statistical properties of the ocean flow, such as large-scale transport rates. Numerically integrating virtual Lagrangian particles in time based on a given velocity field allows one to study the flow in regions that are poorly or not sampled by drifter experiments, such as the deep ocean. It also allows the application of statistical methods to analyse the behaviour of particle ensembles, as particles can be distributed arbitrarily in space and in time to generate sufficiently large amounts of data. Virtual Lagrangian trajectories have been used in different contexts, such as the dispersion of plastic in the ocean [4, 15–18], or to study the outcropping of isopycnals into the surface mixed layer [19].

The equations of motion for a deterministic tracer (without diffusion) are

$$\frac{dx(t)}{dt} = v(x, t), \quad (1.2)$$

where $x = x(t)$ is a particle's position at time t , and $v(x, t)$ the velocity. The velocity field $v(x, t)$ is context dependent. For passive and infinitesimally small particles in the ocean, $v(x, t)$ corresponds to the velocity of the ocean currents. In other situations, $v(x, t)$ might contain additional terms, e.g. a sinking term for dinoflagellate cysts [20], or an inertial velocity component resulting from the finite size of the fluid parcel [21]. In any case, for applications in the ocean, $v(x, t)$ contains the ocean currents in some form, giving rise to chaotic behaviour of the trajectories resulting from the integration of eq. (1.2), with the typical associated phenomena such as unpredictability of individual trajectories and mixing, see e.g. Lasota and Mackey [22] for an introduction to these concepts. The integration of eq. (1.2) from an arbitrary initial condition x_0 at time t_0 yields the flow $\phi_{t_0}^t$:

³<https://www.aoml.noaa.gov/phod/gdp/>

$$x(t) = \phi_{t_0}^t(x_0) = x_0 + \int_{t_0}^t v(x(t'), t') dt'. \quad (1.3)$$

In chapters 2, 3 and 5 of this thesis, we use *parcels*⁴ (Probably A Really Computationally Efficient Lagrangian Simulator) for the numerical integration of eq. (1.2) for large amounts of initial conditions. *Parcels* comes with standard integration routines such as the 4th-order Runge-Kutta method and an easy way to define custom particle behaviour. *Parcels* also has in-built methods to interpolate the velocity field $v(x, t)$ in space and time to the particle position and time for different grid types. The details of the simulations of each study in this thesis can be found in the corresponding chapters.

1.3. Challenges for Lagrangian data analysis

There are many methods to analyse Lagrangian trajectory data, going beyond the application to the ocean. *van Sebille et al.* [23] provide an overview over some of the most common methods that have been applied in the ocean context. As described in section 1.1, individual trajectories have only very limited meaning due to the chaotic nature of the ocean flow. To illustrate the dependence on initial conditions, fig. 1.1 shows the result of 100 simulated trajectories starting in the western North Atlantic Ocean. The underlying ocean velocities are taken from a NEMO ORCA-N006 model simulation [24] with a horizontal resolution of $1/12^\circ$, which is also used in chapters 2, 3 and 5. The virtual trajectories are integrated with the 4th-order Runge-Kutta routine and a time step of five minutes, for a total time of 500 days. The initial conditions are chosen at random within a square cell of $1/12^\circ$, i.e. below model resolution. The large spread of the trajectories in fig. 1.1 over the course of time demonstrates the influence of small uncertainties in the initial drifter location, even if these are below the ocean model grid resolution. Due to the chaotic property of the ocean flow, for a sufficiently long time interval, a single initial condition is not representative for the action of the flow on the points in its surrounding any more.

In some situations, where integration times are short, a few individual trajectories might be good enough to study a certain phenomenon. An example for this case are fish larvae, which are transported by the ocean currents for timescales of less than a month [3]. In the case of surface plastic, particle trajectories are often integrated for several years to a decade [4, 16, 25]. It is clear from fig. 1.1 that individual trajectories can have only very limited meaning after such long integration times, cf. chapter 3.

A further, more practical, challenge arises in the analysis of measured drifter data: drifter trajectories are inhomogeneously distributed in space, and are of different length depending on the drifter lifetime. Figure 1.2 contains some statistics of the drifters of the Global Drifter Program [11]. The data⁵ contains

⁴<https://oceanparcels.org/>

⁵downloaded in March 2020

Figure 1.1: Illustration of 100 trajectories with random initial locations in a $1/12^\circ \times 1/12^\circ$ degree cell, simulated for 500 days. The triangle marks the initial conditions of the particles. The large spread of the virtual trajectories over time is clearly visible, even if the initial distance between them is below the ocean model resolution.

more than 24,000 drifter trajectories in the global ocean. Figure 1.2a shows the large variation of drifter lifetimes, which ranges from a few weeks to more than five years. The number of active drifters present in the global ocean as a function of time is shown in fig. 1.2b, with a strong increase in the 1990s and 2005. The number of drifters that are present in the global ocean at the same time is close to 1,500 in 2020, a number that is too low to study time-dependent phenomena on small scales, e.g. on scales of ocean eddies. Further, drifters are often released at places with a strongly repelling flow, and typically accumulate in the centers of the Subtropical Gyres over the course of time (cf. section 1.4.1 and chapter 4), such that spatial coverage of drifters is highly inhomogeneous. This makes the analysis of drifter trajectories more challenging than for simulated trajectories.

Figure 1.2: Statistics of the data of the Global Drifter Program [11]. a: histogram of the drifter lifetimes. b: number of active drifters as a function of time.

1.4. Macroscopic features of the ocean flow at different scales

This section briefly introduces some of the oceanographic phenomena studied in the chapters of this thesis. They all have in common that they result from the collective behaviour of groups of fluid parcels, and possess some form of insensitivity to exact choices of initial conditions. Ordinary differential equations as the one in eq. (1.2) and their macroscopic properties are described by the theory of finite-dimensional dynamical systems. There are many textbooks on the topic, and we refer to the open source ‘Chaosbook’ [26] for a general introduction with many examples and exercises. In this section, we assume that $v = v(x)$ in eq. (1.2) is independent of time for simplicity of the presentation. The generalization to time-dependent systems is straightforward.

1.4.1. Basin-scale mixing

The chaotic property of a dynamical system implies that two initially close particles can separate in finite time. This sets a microscopic limit on the predictability of individual particle trajectories. On the other hand, for many systems, the long-term distribution of an initial particle cloud is completely predictable and independent of the precise initial locations. For example, in classical chaotic Hamiltonian systems, the preservation of phase space volumes (incompressibility) as stated by Liouville’s theorem implies that the long-term distribution of any initial particle cloud is the uniform distribution in phase space. The attraction of an arbitrary distribution of initial conditions to such an *invariant density* defines a notion of *memory loss*, and the system is thus called a mixing dynamical system.

One can study the mixing properties of a flow with the *Perron-Frobenius operator* $\mathcal{P}$, also called *transfer operator* [22]. The transfer operator is the time evolution operator of particle densities, i.e. it arises from the formal integration of the continuity equation, eq. (1.1), for a time Δt :

$$\begin{aligned}\mathcal{P}_{\Delta t} &= e^{\mathcal{A}\Delta t} \\ \mathcal{A} &= -\nabla \cdot v(x) - v(x) \cdot \nabla.\end{aligned}\tag{1.4}$$

The transfer operator is a Markov operator, i.e. its action on a density $\rho(x, t)$ preserves the L^1 -norm, $\|\rho(x, t)\|_1 = \|\mathcal{P}_{\Delta t}\rho(x, t)\|_1 = \|\rho(x, t + \Delta t)\|_1$. It is a useful tool to study a fluid flow, as it replaces the non-linear evolution operator $\phi_{t_0}^t$ of eq. (1.2) by a linear evolution operator, yet on the infinite space of L^1 -integrable positive functions. As described e.g. by Froyland [27], a finite dimensional approximation $P_{\Delta t} \in \mathbb{R}_{\geq 0}^{M \times M}$ of $\mathcal{P}_{\Delta t}$ based on a partition (binning) of the underlying domain into M bins can be obtained with Lagrangian particles by seeding every bin with a certain number of N particles. The entries of the matrix $P_{\Delta t}$, $(P_{\Delta t})_{ij}$, are then defined as the ratio of particles starting in bin i at $t = 0$ and ending up in bin j at $t = \Delta t$, divided by N :

$$(P_{\Delta t})_{ij} = \frac{\# \text{ particles } (i, 0) \rightarrow (j, \Delta t)}{N}. \quad (1.5)$$

$P_{\Delta t}$ acts on discretised densities, which we denote by $\bar{\rho} \in \mathbb{R}_{\geq 0}^M$. With the definition of eq. (1.5), the action of $P_{\Delta t}$ on discretised densities $\bar{\rho}$ is defined by multiplication from the right:

$$\bar{\rho}(t + \Delta t) = \bar{\rho}(t)P_{\Delta t}. \quad (1.6)$$

The matrix $P_{\Delta t}$ is a finite-dimensional Markov operator, i.e. its rows add up to one, and time evolution by $P_{\Delta t}$ describes a finite Markov chain:

$$\bar{\rho}(t + n\Delta t) = \bar{\rho}(t)(P_{\Delta t})^n. \quad (1.7)$$

The transfer operator allows to define the *invariant measure* of the discretized dynamical system, π , by the density corresponding to eigenvalue 1 of $P_{\Delta t}$:

$$\pi = \pi P_{\Delta t}. \quad (1.8)$$

The fundamental theorem of Markov chains states that if a Markov chain is aperiodic and irreducible, there is a unique stationary density π with π_i , $i = 1, \dots, M$, strictly larger than zero [28]. Periodicity describes a situation where the Markov chain returns to its original state after iterating for an integer $n > 1$ in eq. (1.7), an unlikely situation if P results from the discretisation of chaotic dynamical system. Irreducibility means that every state can be reached from any other state. For the dynamical system of eq. (1.2) this implies that there are no boundaries to particle transport, cf. section 1.4.2. If $P_{\Delta t}$ has a unique stationary density, then we have for an arbitrary density field $\bar{\rho}_0$,

$$\lim_{n \rightarrow \infty} \bar{\rho}_0 P_{\Delta t}^n = \pi.$$

In practice, $\bar{\rho}_0 P_{\Delta t}^n$ will be indistinguishable from π already for a finite value of n , which motivates the concept of the Markov chain mixing time $t_{mix} = n\Delta t$ by the smallest value of n with $|\bar{\rho}_0 P_{\Delta t}^n - \pi| < \epsilon$, where $\epsilon > 0$ is some form of cutoff (typically $\epsilon = 1/4$, cf. chapter 3). As this is valid for any initial density $\bar{\rho}_0$, the information of the initial state is lost over time.

In the ocean, the horizontal wind stress at the surface leads to the formation of the Ekman current, with non-zero divergence of the 2-dimensional surface velocity, see e.g. chapter 8 of Cushman-Roisin and Beckers [29]. The consequence is a non-zero vertical velocity at the ocean surface. In the centres of the Subtropical Gyres, this velocity points downwards, i.e. the surface currents converge, leading to the attraction of particles transported at the ocean surface. This convergence, together with other components of the flow, leads to the large-scale subtropical accumulation zones of microplastic, also called ‘garbage patches’. In chapter 3, we study the mixing properties of the surface ocean flow, with important implications of the interpretability of individual particle trajectories.

1.4.2. Transport barriers

A strict transport barrier of a flow defined on an n -dimensional manifold Ω is simply a $(n-1)$ -dimensional hypersurface that cannot be crossed by any trajectory. Transport barriers divide a fluid into disjoint sub-manifolds $\{\Omega_i\}$, $\cup_i \Omega_i = \Omega$. If a fluid flow possesses a strict transport barrier, one can study the flow on each of the disjoint manifolds separately. In general, systems with noise do not possess strict transport barriers. For $t \rightarrow \infty$, any two regions of the fluid domain can be connected by the fluid flow. Even in such a situation, however, one can often find approximate transport barriers by looking for non-overlapping sets such that transport between the different sets is unlikely in a given time interval. Finding such approximate transport barriers can yield valuable insights and greatly simplify the understanding of a flow [30].

The surface ocean in fact consists of many approximate transport barriers. In the North Atlantic, the Subpolar Gyre is known to be approximately separated from the Subtropical Gyre [31]. Approximate transport barriers are also relevant for the long-term transport of plastic at the ocean surface, where transport between different subtropical basins is small [4]. We study approximate barriers to transport in chapters 2 and 4. In chapter 2, we define a partition of the surface ocean based on known transport barriers for passive surface particles, and study how these ‘barriers’ can be crossed by particles that are transported slightly below the surface. In chapter 4, we develop a network based on drifter trajectories and detect, through data clustering, well-known transport barriers in the North Atlantic Ocean.

1.4.3. Mesoscale eddies

The ocean is full of mesoscale eddies, i.e. circular flow structures on the scale of around 10-100 km. The formation of eddies can be explained within quasi-geostrophic theory [5]. Large horizontal shear of the flow or horizontal density gradients lead to instabilities of geostrophic flows, such that ageostrophic perturbations amplify. To first order, the resulting perturbation of the fluid velocity leads to the circular flow observed in an eddy. An important feature of eddies is their ability to trap water masses, while being transported by the background flow. Fluid parcels that are trapped in the eddy stay in the eddy until its decay, and have, on the larger scale, very similar pathways. Consequently, precise initial conditions (below the size of the eddy) only play a minor role for these particles. The identification of eddies in simulated ocean trajectories allows a comparison with observations, e.g. in terms of their size or frequency. Many methods to detect so-called Lagrangian coherent structures, or finite-time coherent sets, have been proposed in the context of ocean eddy detection [32–37]. Identifying eddies and other coherent sets of particles in an otherwise noisy background flow is the focus of chapter 5.

1.5. Goals and research questions of this thesis

This thesis was conducted within the TOPIOS project (Tracking Of Plastic In Our Seas)⁶. Chapters 2 and 3 specifically address questions arising in transport modeling of plastic in the ocean, the methods presented there are however not limited to plastic and can be applied in other situations as well. The transport of plastic in the ocean is a particularly interesting example, as plastic particles can be transported on the global scale, for years or even decades [4], a time scale after which only very limited meaning can be attributed to an individual trajectory (cf. chapter 3). Chapters 4 and 5 focus more generally on meso- and large-scale structure detection in Lagrangian trajectory data in the ocean.

The main goals of this thesis can be summarized as follows:

1. Further develop Lagrangian models to better understand the large-scale and long-term transport properties of plastic in the ocean
2. Investigate the limits regarding the interpretation of basin-scale long-term Lagrangian trajectory data at the ocean surface, which is particularly relevant for plastic transport modelling
3. Propose new, data-driven, methods to efficiently analyse large data sets of Lagrangian particles and extract macroscopic information from them, for both ocean drifters and numerically simulated trajectories

The research questions for each individual chapter of this thesis are summarized in the following paragraphs.

How does our understanding of global plastic transport in the ocean change if we take into account near-surface currents?

A major paradigm of buoyant plastic transport is its accumulation in the subtropical accumulation zones, called garbage patches. Simulations of virtual particles at the ocean surface have shown that this accumulation comes with a depletion of particles in polar and equatorial regions. However, buoyant plastic entering the ocean does not necessarily stay at the ocean surface for all times. It is known that buoyant plastic can lose its buoyancy due to biofouling, or be transported to deeper layers through wind-induced turbulent mixing. Further, very small particles, though buoyant, can follow the three-dimensional currents and not stay at the surface. Chapter 2 focuses on the long-term fate of microplastic and goes beyond studying transport in the ocean surface flow only. Particles are constrained to different depths close to the surface to mimic buoyancy loss through biofouling, advected with the three-dimensional velocity field or randomly displaced along the vertical to simulate wind-driven turbulent mixing in the upper water column.

What is the mixing time of the surface ocean flow, and how does that affect our current understanding of plastic transport?

⁶<https://topios.org/>

Mixing, as described in section 1.4.1, has implications for the interpretability of individual particle pathways: beyond the mixing time scale, initial conditions only play a minor role. This potential loss of correlation between initial conditions and final particle distribution has not been studied in the plastic transport modeling community, although the time scales that are relevant for the global accumulation of surface plastic are relatively long. In chapter 3, we study the mixing time of surface particles in the ocean, using the concepts of Markov chain mixing and mixing entropy.

Can we identify the major features of the large-scale flow in the North Atlantic Ocean based on drifter data with network theory?

There are several important large-scale features in the North Atlantic Ocean flow, such as the Western Boundary Current, the Caribbean Sea, and the transport barrier between the Subtropical and Subpolar gyres. The detection of these features from drifter data is however not trivial, in particular as the data is very inhomogeneous in space and time, and contains data gaps. In chapter 4, we develop a network based on drifter trajectories to detect such large-scale flow features from Lagrangian trajectories. Our method is designed for the application to scarce and incomplete drifter data.

Can we accommodate the concept of hierarchy and noise in detecting coherent structures in Lagrangian particle data?

There are many methods to detect coherent structures, such as eddies, in Lagrangian particle data, cf. section 1.4.3. A few recent methods explore the possibility to detect coherent structures using particle based networks together with data clustering techniques. Yet, most of these methods are based on graph partitioning, i.e. each particle trajectory by definition belongs to a cluster. This is problematic for applications in the ocean, where many eddies are present in large, noisy fluid domains. In chapter 5, we explore if eddies can be detected with the OPTICS clustering algorithm of Ankerst *et al.* [38], which comes with a natural notion of noisy trajectories and a hierarchy of coherent structures.

All four research questions can be seen as different facets of the same, central question: **how can we identify macroscopic structure in large sets of Lagrangian ocean trajectories?** Each chapter addresses this question from a different viewpoint and in a different context, and with different temporal and spatial scales. Yet, the overall goal of each of the methods presented in this thesis is to get rid of the superfluous information of individual trajectories, and to provide tools for the analysis of the large-scale, emergent features of ocean flow.

2

Influence of Near-Surface Currents on the Global Dispersal of Marine Microplastic

Buoyant microplastic in the ocean can be submerged to deeper layers through biofouling and the consequent loss of buoyancy or by wind induced turbulent mixing at the ocean surface. Yet the fact that particles in deeper layers are transported by currents that are different from those at the surface has not been explored so far. We compute 10-year trajectories of 1 million virtual particles with the Parcels framework for different particle advection scenarios to investigate the effect of near-surface currents on global particle dispersal. We simulate the global-scale transport of passive microplastic for (i) particles constrained to different depths from the surface to 120 m depth, (ii) particles that are randomly displaced in the vertical with uniform distribution, (iii) particles subject to surface mixing and (iv) for a 3D passive advection model. Our results show that the so called 'garbage patches' become more 'leaky' in deeper layers, and completely disappear at about 60 m depth. At the same time, subsurface currents can transport significant amounts of microplastic from subtropical and subpolar regions to polar regions, providing a possible mechanism to explain why plastic is found in these remote areas. Finally, we show that the final distribution in the surface turbulent mixing scenario with particle rise speed $w_r = 0.003$ m/s is very similar to the distribution of plastic at the surface. This demonstrates that it is not necessary to incorporate surface mixing for global long-term simulations, although this might change on more local scales and for particles with lower rise speeds.

This chapter has been published as:

Wichmann, D., Delandmeter, P., and van Sebille, E. (2019): Influence of near-surface currents on the global dispersal of marine microplastic. *Journal of Geophysical Research: Oceans*, 124, 6086–6096. <https://doi.org/10.1029/2019JC015328>

2.1. Introduction

Plastic pollution in the ocean has become one of the major environmental concerns in recent years. Plastic debris has been found in all five ocean basins [39], in the deep sea [40], and has been measured at the poles and in sea ice [41–43]. Debris can be transported thousands of kilometres away from its source by ocean currents, and is hence a global environmental problem. Owing to the complexity of the ocean currents, numerical modelling is an essential tool for understanding the transport pathways and accumulation patterns of marine plastic debris [18].

Transport modelling has proven to be successful in explaining the distribution of surface microplastic. The high concentrations in the subtropical gyres, often called ‘garbage patches’, is mainly a consequence of the converging Ekman component of the surface velocities [15, 44]. The fact that many numerical studies find more or less the same surface accumulation pattern irrespective of circulation models, temporal and spatial resolution and spatio-temporal plastic input scenarios [4, 14, 16, 17, 25] indicates a rather strong robustness of surface transport, sharing some similarity with an attractor of a dynamical system: small perturbations such as differences in the circulation models, sub-grid scale effects and uncertain initial conditions hardly change the long-term distribution of surface debris. Note that the individual garbage patches all play the role of (approximate) attractors [45], such that the kind of robustness described above is only expected within each basin separately. While a strong attractor is a convenient property in modelling transport rates and final distributions in a highly uncertain flow (such as in marine plastic modelling), one may be tempted to over-emphasize the role of the surface *flow* based on more or less accurately simulating the spatial *distribution* of surface plastic. A strong attractor at the ocean surface could in principle eliminate some of the history of a particle ensemble. Degradation processes or sinking would then not produce an easily observable signal in the surface distribution. Simulations of the 2-dimensional surface dynamics, though inaccurate on the individual particle level, could still produce accurate results for the distribution in the attractor region.

This is likely to be the case in the ocean due to two facts. First, it is difficult to reconcile the total surface plastic amount estimated with observations and modelling [17, 25, 39, 46] with the estimated amount of plastic available to enter the ocean [47]. Typically, in these studies the observed amounts at the surface are at the order of 1% of the estimated amount entering the ocean in 2010 only (according to Jambeck *et al.* [47]), whereas around half of the produced plastic is initially buoyant [48]. Second, microplastic has been found in regions far off the subtropical accumulation regions such as the Southern Ocean and Antarctica [43, 49, 50] and in the Arctic [41, 42, 51]. As discussed in these studies, it is often not clear whether microplastic in the polar regions originates from local sources or if there are significant transport pathways from lower latitudes into these regions, and the answer may very well depend on the specific sampling location. If transport pathways from lower, subtropical latitudes exist however, these can hardly be explained by the strongly converging surface currents, al-

though some of the polar transport can also be explained by wave-induced Stokes drift [9], and perhaps also by atmospheric transport [52]. These points show that near-surface transport, rather than surface transport, in the water column could play a crucial role in the global dispersion of marine microplastic. In fact, it has been shown that through biofouling, initially buoyant plastic might lose its buoyancy within a few weeks to months [53, 54], which is much less than the typical time scale of attraction of a garbage patch (several years, see fig. S2.1 in the supplementary materials). What happens with non-buoyant microplastic in the water column is however rather poorly understood. Kooi *et al.* [2] developed a model for explaining the size-selective behaviour of surface plastic removal and showed that particles can in principle stay in the water column, performing oscillations in the vertical caused by subsequent fouling and de-fouling, as was also described by Ye and Andrady [55]. For such kind of particles, it is obvious that subsurface currents are the ones that govern their motion and not the currents at the surface, and the former are expected to differ strongly from the latter due to the exponential decay of Ekman velocities as well as baroclinic pressure gradients. In particular, we expect less pronounced accumulation regions at depth due to the decay of the Ekman velocity.

In addition to biofouling and buoyancy loss, buoyant plastic particles at the ocean surface can be transported to deeper layers through wind-induced turbulent mixing, which leads to an exponentially decreasing depth profile of plastic concentrations depending on the particle rise velocity and local wind conditions [56–58]. This effect was shown to be important when comparing simulation results to measured surface concentrations, as surface trawl nets usually sample the uppermost part of the ocean only. The potential underestimation of the total amount of marine plastic was at first demonstrated by Kukulka *et al.* [56], and since then simulation results and measured concentrations have been corrected for such effects (e.g. [17, 25, 46]). In the simulations performed so far however, the effect of turbulent mixing was considered only after the simulation of trajectories, which does not take into account a potential effect of varying flow fields with depth. Clearly, a weakly buoyant particle will spend, through turbulent mixing, more time deep in the water column than at the surface because once submerged it needs more time to rise back to the surface.

Lastly, very small plastic particles are expected to behave very similarly to passive water parcels, with only little dependence on particle density. This is intuitively clear by considering Stokes' law for the rise velocity of a suspended particle, which depends linearly on the density difference, but quadratically on the particle diameter (e.g. [59]). Very small particles, even with a lower density than water, are therefore only very weakly buoyant and are likely transported by the full 3-dimensional currents, rather than the surface currents.

It should be emphasized that there is a range of other possible processes that can have an effect on plastic transport pathways, and which are not addressed in this paper. To name only a few, particle properties such as composition, size and shape can have an impact on global transport pathways because they affect the surface removal rate through biofouling [53, 60], on the rise velocity of

submerged particles [58, 61], and can lead to inertial effects of a particle in a fluid (e.g. [62]). Shape, size and buoyancy of a plastic particle can also change in time due to fragmentation through degradation [63]. The beaching of plastic particles, e.g. through breaking waves is far from being understood and could constitute a large sink of plastic. More realistic plastic source scenarios, such as the one of Lebreton *et al.* [64], can clearly have an effect on plastic distributions. Oceanographic phenomena such as downwelling and tides may also have a local effect on the transport pathways of plastic particles. All of these processes come with relevant temporal and spatial scales of importance, and it is currently not known to what extent the different processes are important for the global long-term fate and distribution of marine plastic. The large number of uncertainties makes it challenging to model the global dispersion of realistic marine debris. More experimental and numerical efforts are needed to constrain the parameter space that governs the dispersion of marine plastic debris, together with extended measurement campaigns of marine plastic that not only contain plastic counts, but also size, composition and other parameters such as the amount of biofouling.

Here, we do not intend to fully model the realistic dispersion of plastic in the ocean. Neither do we parameterise relevant processes such as degradation nor do we use realistic plastic input scenarios. We rather aim to assess the importance of sub-surface advection for the global accumulation and transport pathways of marine debris. Despite the evidence for the importance of vertical transport of microplastic presented above, to our knowledge no study so far has addressed the question how the global accumulation pattern and pathways of microplastic could be influenced by the transport through subsurface instead of surface currents. Yet, the importance of measuring plastic throughout the entire water column is starting to get recognised in the plastic community, and our understanding of the dispersal, accumulation and quantity of plastic might change substantially when deeper measurements become available [65]. With this paper, we wish to contribute to the understanding of the dynamics of submerged particles, which will also help with the interpretation of deep measurements.

Here, we study this effect through numerically simulating the trajectories of approximately 1 million particles in the global ocean for 10 years of advection for different simulation scenarios. We test the importance of biofouling-related subsurface transport through constraining particles to certain fixed depths and through randomly displacing them in the upper ocean. These models are highly idealised, but due to the lack of precise understanding of the vertical transport dynamics this can be seen as a first step in studying the consequences of subsurface transport on the global dispersion of marine microplastic. We also study the effect of turbulent surface mixing through explicitly including the wind-dependent vertical mixing process according to Kukulka *et al.* [56] in the particle trajectory integration. Finally, we run a simulation where particles are subject to the full 3-dimensional currents of our dataset.

2.2. Methods

We simulate the transport of microplastic through modelling the trajectories of Lagrangian particles. We use version 1.11 of the Parcels framework [66] (oceanparcels.org) for the 2D simulations, and version 2.0 for the 3D simulation [67]. We use current and wind stress data for the 10 years of 2000-2009 from a NEMO ORCA-N006 run, a standard run of NEMO [24], forced by the Drakkar forcing [68], available at <http://opendap4gws.jasmin.ac.uk/thredds/nemo/root/catalog.html>. The flow and wind stress fields are available on an ORCA grid [69] with a $1/12^\circ$ resolution, provided for every 5 days. The model has 75 vertical levels with an increasing spacing between them for larger depths, 26 of which are within the upper 120 m of the water column.

The fields are interpolated in space and time to the particle position and time with the C-grid interpolation scheme described in Delandmeter and van Sebille [67]. We use the 4th order Runge-Kutta method with a time step of 10 minutes for the trajectory integration. For each of the different simulation scenarios (see table 2.1), we study the long-term distribution of marine plastic and the transport pathways between different ocean regions (shown in fig. S2.2 in the supplements) with approximately 1 million particle trajectories starting from a uniform distribution on the vertices of a $0.2^\circ \times 0.2^\circ$ grid on January 5, 2000 and advecting them for 10 years.

We perform runs for four kinds of passive particles: (i) particles constrained to different depths, (ii) particles randomly displaced in the vertical according to a uniform distribution at each time step, (iii) particles that are, again at each time step, randomly displaced in the vertical according to the exponential distribution suggested by Kukulka *et al.* [56], and (iv) particles that follow the full 3-dimensional flow. (i) also includes a surface run ($z = 0$ m), which serves as a baseline to compare the other simulations to. Simulations with particle constrained to certain depths (i) represent the horizontal advection of particles in the water column. The assumption of a fixed depth is a strong simplification, but it allows us to gain first insights into the transport of submerged plastic particles with a reasonable computational cost and with no additional assumptions on real vertical transport mechanisms. Similarly, the simulation of the uniformly displaced particles serves as a simple model where particles can move to any other position in the upper water column, either due to biofouling or wind-induced mixing with very low buoyancy. Simulations (i)-(iii) neglect vertical flow velocities and are confined to the depth interval [0 m, 120 m]. 120 m is the depth where the concentration for very weakly buoyant particles with rise speed $w_r = 0.003$ m/s under strong wind conditions of $\tau = 0.8$ Pa decreases to roughly 20 % of the surface concentration within the parameterisation of Kukulka *et al.* [56] (see below), such that it captures the largest part of the turbulent mixing range in the surface layer. It also comprises the range of depths within which bio-fouled initially buoyant microplastic particles oscillate in the water column according to the model of Kooi *et al.* [2] (which ranges typically to around 60 m, see their fig. S1). The simulation of passive particles at the surface (at $z = 0$ m) is similar to the approach of most previous studies on surface

microplastic (e.g. [16, 17, 25]). For the 3D simulation (iv), we place the particles at an initial depth just below the surface at $z = 1.5$ m as we work with a non-moving ocean surface such that the vertical velocity is set to zero at $z = 0$ m by definition. This is similar to the methodology in Delandmeter and van Sebille [67].

According to Kukulka *et al.* [56], the depth-dependent plastic concentration $n(z)$ follows an exponential distribution (we take the z -axis pointing downwards)

$$n(z) \propto \exp(-\lambda z). \quad (2.1)$$

Here $\lambda = w_r/A_0$ where w_r is the rise velocity of the buoyant particle and $A_0 = 1.5u_{*w}\kappa H_s$ with u_{*w} the frictional velocity of water, $\kappa = 0.4$ the van Karman constant, $H_s = 0.96g^{-1}\sigma^{3/2}u_{*a}^2$ with gravitational constant g , wave age σ and frictional air velocity u_{*a} . As in [56] we take $\sigma = 35$. The frictional velocities of air and water are determined by the wind conditions, given by $u_{*w} = \sqrt{\frac{|\vec{\tau}|}{\rho_w}}$ for water with wind stress $\vec{\tau}$ and water density ρ_w . Finally, the frictional velocity of air is defined by $u_{*a} = \sqrt{\frac{|\vec{\tau}|}{\rho_a}}$. Taking $\rho_w = 1027$ kg/m³ and $\rho_a = 1.22$ kg/m³ leads to $A_0 = 0.31 \cdot |\vec{\tau}|^{3/2}$, where $|\vec{\tau}|$ is given in Pa and A_0 in m²s⁻¹. We model turbulent mixing through randomly displacing the virtual particles at each time step along the vertical with an exponential probability distribution $p(z) = \lambda e^{-\lambda z}$, where λ depends on the wind stress that is taken from the NEMO dataset. This approach assumes that the time scale of turbulent mixing is much faster than the one of particle advection, which is a reasonable assumption. We take a small value of the rise velocity of $w_r = 0.003$ m/s, which is at the lower end of the rise velocities reported by Reisser *et al.* [58] and hence simulates an extreme case of strong mixing effects. Table 2.1 shows the different simulation scenarios of this study (1 million particle trajectories each). The code for the simulations and data analysis is available on GitHub (https://github.com/OceanParcels/near_surface_microplastic).

Table 2.1: Simulation overview

Simulation	Parameters
(i) Fixed depth	$z = 0, 2, 5, 9, 16, 26, 41, 61, 87, 97, 120$ m ^a
(ii) Uniform mixing	$p(z) = U(0 \text{ m}, 120 \text{ m})$ ^b
(iii) Exponential mixing	$p(z) = \lambda e^{-\lambda z}$, $w_r = 0.003$ m/s, $z < 120$ m ^c
(iv) Full 3D advection	Initial depth = 1.5 m

^aThe values correspond to the depths within different layers of the hydrodynamic data.

^bRandom uniform distribution between top layer at $z = 0$ m and $z = 120$ m.

^cSee the text for the definition of λ .

2.3. Results

2.3.1. Particle distributions

Figure 2.1 shows the resulting long-term distribution of the surface ($z = 0$ m) particles at the end of the simulation. The accumulation regions in all five subtropical gyres are clearly visible and agree roughly with the expected locations from measured data (compare to e.g. fig. 1 of van Sebille *et al.* [25]). Note that many regions outside of the gyres are completely depleted of particles after the 10 years of simulation. This distribution did not change significantly in the last few simulated years, indicating that the particle distribution is in an almost-equilibrium state (see fig. S2.1 in the supplementary materials).

Figure 2.1: Modelled distribution for surface particles (fixed depth of $z = 0$ m), in logarithmic scale with square binning of 2° . As described in the text, the particles were released uniformly and advected for 10 years.

Figure 2.2 shows the final distribution for a selection of the different scenarios (see fig. S2.3 for the corresponding maps for the other scenarios), which we compare to the distribution of the surface simulation shown in fig. 2.1. Note that for figs. 2.2d-f, the distributions are shown as depth-integrated sums, i.e. distributions of the particles' longitudes and latitudes only. As expected, the strong converging property of the surface flow weakens at depth, see figs. 2.1 and 2.2a-c. The garbage patch structure is still clearly visible in the top part of the water column, with a transition towards more smeared out distributions at 41 m. Note that $z = 16$ m shown in fig. 2.2a is close to the depth sampled by drogued drifters [70]. At 61 m (fig. 2.2c), no garbage patches exist any more, and the distribution remains essentially unchanged up to the deepest layer considered at $z = 120$ m. Different from the surface flow, the polar and equatorial regions are not as strongly depleted of particles at larger depth any more. This is also the case for the random uniform simulation (fig. 2.2d), although the patches are still clearly visible. The exponential mixing simulation according to Kukulka *et al.* [56] does not show a significant visible difference

to the surface simulation on the global scale, as seen in fig. 2.2e. It should be noted however that this could be an effect of the choice of the rise speed (here $w_r = 0.003$ m/s), which depends on particle composition and size. Smaller particles with even smaller rise speeds would be mixed into deeper layers, and consequently the effects of turbulent surface mixing would increase. In addition, turbulent mixing could still be important on smaller spatial and temporal scales. The distribution of the 3D simulation (fig. 2.2f) does show only very little spatial variation, together with a removal of particles from the Southern Ocean region. Note here that this can also be an effect of our choice of initial condition for the 3D simulation, as we position the particles close to the surface at a depth of 1.5 m. More than 80% of the particles are still within the upper 350 m of the water column at the end of the simulation (see fig. S2.6).

Figure 2.2: Final distributions of a selection of simulations. The colour bars show the number of particles per bin on a square binning of 2° . a-c show a selection of cases for the fixed depth simulations (i), see table 2.1. d-f show the results for the other cases (ii)-(iv). In d-f, the distributions are depth-integrated sums. Similar figures for the other scenarios can be found in fig. S2.3.

2.3.2. Regional transport

We next analyse the inter-regional transport by dividing the ocean into seven regions: North Pacific, North Atlantic, South Pacific, South Atlantic, Indian Ocean, Arctic, and Southern Ocean. A map of the division is shown in fig. S2.2. The

Arctic region is chosen to be north of 65°N latitude, and the Southern Ocean to be south of 60°S . We analyse the regions of attraction of the different garbage patches and the connectivity between the polar and lower-latitude regions for the different simulations. Figures 2.3 and 2.4 show the final destination plotted in the respective colour at the initial location of the particles for those ending up in one of the five subtropical basins, and the two polar regions, respectively. In fig. 2.3a, we see rather well-defined boundaries between the different colours for the surface simulation, as expected. The area of each colour here corresponds to the region of attraction of the garbage patches. It is clearly visible that surface particles hardly cross the Equator in the Pacific and Atlantic basins. For larger depths (figs. 2.3b-c) however, these borders become smeared out, in particular the border between the Atlantic and Pacific basins, respectively: there are no clear-cut boundaries between the different basins any more, already at the relatively shallow depths we consider here. Similar plots for the other scenarios are shown in fig. S2.4. Figures 2.3a-c also show that the Indian Ocean plays a somewhat special role: the region of attraction becomes much less coherent at intermediate depths of $z = 41$ m, and more coherent again for deeper particles at $z = 120$ m. This is different from the other four subtropical basins, where the respective regions of attraction become less coherent monotonously with depth. The result for the uniform mixing scenario in fig. 2.3d resembles in its broad features the simulation at a depth of $z = 41$ m (fig. 2.3b). It is again notable that there is no significant visible difference between the surface (fig. 2.3a) and exponential mixing (fig. 2.3e) scenarios, although we would expect this to change for smaller rise speeds, as discussed above. Finally, we see in fig. 2.3f that in the 3D simulation the boundaries between the basins are also less clearly defined, in particular for the Atlantic basin. This is expected due to the transport of shallow water from the South to the North Atlantic where deep water is formed (see fig. S2.6) as part of the thermohaline circulation.

In figs. 2.4a-c, we can see that transport towards the poles increases with increasing depth. We also see that the polar North Atlantic is quite well connected to the subtropical North Atlantic already for the surface flow, as was also observed to be the case for surface drifters [31], although this connection increases for larger depths. For the Southern Ocean, at $z = 120$ m particles can originate from parts of the Indian Ocean and from the South Pacific and South Atlantic on the 10-year time scale considered here. This is very different from the surface flow and the 3D flow, where almost nothing enters the Southern Ocean from the other basins. The effect of the Antarctic Circumpolar Current to prevent transported matter from entering the Southern Ocean hence weakens significantly with depth. Figure 2.4d shows that the uniform mixing simulation leads to a stronger flow of particles from the northern North Pacific to the Arctic region, compared to the simulations at fixed depths. This is surprising, but it highlights that vertical dynamics can actually matter, and that the dynamics of the random uniform simulation is not simply an average of the fixed depth simulations. As can be seen in fig. 2.4e, the exponential mixing sim-

Figure 2.3: Scatter plot of the initial position of the particles that end up in one of the subtropical basins (indicated by the colours) for a selection of the performed simulations. a-c show a selection of cases for the fixed depth simulations (i), see table 2.1. d-f show the results for the other cases (ii)-(iv). Similar figures for the other scenarios can be found in fig. S2.4.

ulation does again not show significant differences to the surface simulation. Finally, we note that there is also significant inflow from the northern basins into the Arctic region in the 3D simulation (fig. 2.4f), as expected from the circulation in the global thermohaline circulation. Note that the fact that there are almost no particles that reach the Southern Ocean regions in the 3D simulation is again likely to be a consequence of our choice of initial depth of 1.5 m for the particles. Simulation results for the other scenarios are provided in the supplementary materials (fig. S2.5).

With figs. 2.3 and 2.4 one has to take into account that our simulation is of course limited by an unrealistically low number of particles released only once into the ocean with a uniform distribution. Continuous, spatially heterogeneous release of many more particles, as in the real world, would also lead to different transport into these regions.

Based on the regional division (fig. S2.2), we form a transport matrix F : if n_{ij} is the number of particles that starts in region i and ends up in region j after 10 years, we define this matrix as $F_{ij} = n_{ij} / \sum_j n_{ij}$, i.e. the row normalised form of n_{ij} . The matrix F_{ij} describes the share of those particles that are transported from region i to j compared to all particles that start in region i , and

Figure 2.4: Initial position of the particles that end up in the polar regions (indicated by the colours) for a selection of the performed simulations. a-c show a selection of cases for the fixed depth simulations (i), see table 2.1. d-f show the results for the other cases (ii)-(iv). Similar figures for the other scenarios can be found in fig. S2.5.

its entries therefore indicate the strength of the particle transport between the different regions. The transfer matrices for all simulations are shown in tables S2.1-S2.14. The main observation is that, apart from the Indian Ocean, the diagonal elements F_{ii} for i a subtropical basin decrease significantly for larger depths, by roughly 7-13% from the surface to $z = 120$ m, demonstrating that the garbage patches in these four basins become less stable in the deep or that their basins of attraction shrink. An exception is the Indian Ocean, whose diagonal element increases by about 10%. Note that these percentages are all sensitive to the choice of basins, which is especially difficult at the boundary between South Pacific and Indian Ocean. The diagonal elements for $i = \text{polar region}$ increase by 12% (Arctic) and 60% (Southern Ocean) between the surface and 120 m, showing that particles in these regions are more likely to stay there at larger depth than at the surface.

Figure 2.5 shows the significant non-diagonal (a and c) and diagonal (b and d) entries of F for transport towards the two polar regions. To assess the robustness under different choices of regional divisions, we computed the transport matrix for different definitions of the Southern Ocean boundary (56° , 60° , 62° S) and the Atlantic-Arctic boundary (60° , 65° , 70° N), illustrated by the light-coloured

areas (for different depths) and error bars (for the other scenarios) in the figure. For the Southern Ocean, we can see that there is no particle transport from the subtropical basins at the surface, but this increases to roughly 7% (South Atlantic), 4% (Indian Ocean) and 1% (South Pacific). In addition to the stronger import of particles into the Southern Ocean, there is also a weaker export of particles from the Southern Ocean, as can be seen by the strongly increasing diagonal element of F (fig. 2.5b). The situation is slightly less clear for the Arctic, where transport from the North Atlantic increases with depth, but transport from the North Pacific decreases (fig. 2.5c). As mentioned earlier, it is especially remarkable here that the random uniform simulation leads to significantly stronger transport from the North Pacific as any of the fixed-depth simulations. Finally, fig. 2.5d shows that over the whole range of depths the export of particles from the Arctic also seems to decrease (as the diagonal element increases), but this trend is not as significant as for the Southern Ocean. It is visible that the transport rates of the exponential mixing scenario are again very similar to the $z = 0$ m case. It should be noted that there is no monotonous trend between the transport rates and the choice of boundaries of the two polar regions, as is especially visible for the North Atlantic-Arctic transport shown in green colour in fig. 2.5c.

2.4. Conclusion

The long-term dynamics of surface microplastic is governed by strong basin-scale attracting regions, the garbage patches. Due to their robustness to different simulation scenarios and their ability to explain measured surface concentrations, surface simulations have dominated the recent efforts of global oceanic microplastic modelling. Nevertheless, if transport from subtropical to polar regions exists, this cannot be explained by considering the surface layer only, at least if no additional effects such as wave-induced Stokes drift [9] or atmospheric transport are taken into account. We provide another possible mechanism to explain such poleward transport: the advection by near-surface currents.

In this study, we tested the potential role of subsurface currents through simulating Lagrangian particle trajectories for 10 years. We perform runs for four kinds of passive particles: (i) particles constrained to different depths, (ii) particles randomly displaced in the vertical according to a uniform distribution at each time step, (iii) particles that are, again at each time step, vertically displaced according to the exponential distribution suggested by Kukulka *et al.* [56], and (iv) particles that follow the full 3-dimensional flow. The results show that the transport of submerged microplastic is determined by completely different dynamics than at the surface. The decay of the Ekman component with depth not only leads to less convergence and hence less pronounced accumulation spots of plastic, but there is also stronger transport from the subtropical and subpolar regions into the polar regions for particles at larger depth. This means that plastic particles in regions of lower latitudes that lose their buoyancy can in principle be transported to originally non-contaminated polar re-

Figure 2.5: Matrix elements for particle transport to the poles for the different scenarios: fixed depth (i), uniform mixing (ii), exponential mixing (iii) and the 3D simulations (iv). Note that the plots have different y-axes. The edges of the light coloured areas (for fixed depths) and error bars ((ii)-(iv)) in the figures display the result under different definitions of the Southern Ocean (a, b) boundary (56°, 60°, 62° S) and the Atlantic-Arctic (c, d) boundary (60°, 65°, 70° N). The solid markers correspond to the boundaries at 60° S and 65° N respectively. Note that changing the region boundaries does not affect the transport rates monotonously, such that the rates for different region definitions can intersect, as is particularly visible for the North Atlantic in (c).

gions. The exponential mixing scenario according to Kukulka *et al.* [56] did not show any visible significant difference to the surface simulation, although this would clearly change for particles with smaller rise speeds, which have a higher probability to stay away from the surface for longer times, and potentially also on smaller temporal and spatial scales. The 3D simulation also showed significant transport into the Arctic, which is a possible mechanism to explain the transport of neutrally buoyant (very small) plastic particles into that region.

We studied the effects of subsurface currents through constraining particles to certain depths, randomly displacing them or advecting them passively with the 3D flow, all of which are strong simplifications of the real world processes. In addition, our analysis is based on simplified initial conditions of globally uni-

form particles at one point in time, which affects transport rates between different regions. A full model of realistic plastic particles would need to take into account a range of processes through parameterisations. For example, in the model of Kooi *et al.* [2], most particles do not stay at a fixed depth but constantly oscillate in the water column. A more realistic approach to study the effect of biofouling would hence have to couple the Lagrangian advection scheme to a model for the vertical transport of particles, such as the one of Kooi *et al.* [2]. However, considering the fact that the mechanisms of vertical transport are still not fully understood, this study gives a first indication to the full complexity of marine plastic transport and provides a simple and plausible mechanism to explain how plastic could be transported from subtropical to polar regions. Our results show that there is a need for a better understanding of the sinking processes of plastic and to incorporate such effects in 3-dimensional transport modelling in order to close the global marine plastic budget. This is crucial to understand the impact plastic debris can have on remote but through subsurface currents connected pristine ecosystems.

2.5. Supplementary material to chapter 2

Figure S2.1: Modelled annual surface distributions. The basin-scale accumulation regions in the subtropics (garbage patches) become visible after 3 years of advection, and approximately stable in time after 7 years. The colour bar represents the number of particles per 2° bin (logarithmic scale).

Figure S2.2: Choices of different regions for the regional transport analysis.

Figure S2.3: Distributions after the 10 simulation years for all simulations not shown in the main text. The colour bar represents the number of particles per 2° bin (logarithmic scale).

Figure S2.4: Initial location (scatter points) and final region (colour) after 10 years, where particles end up in one of the subtropical basins. Shown are the results for the scenarios not presented in the main text.

Figure S2.5: Initial location (scatter points) and final region (colour) after 10 years where particles end up in one of the polar regions. Shown are the results for the scenarios not presented in the main text.

Figure S2.6: Vertical distribution of particles in the 3D simulation after 10 years of advection, in the Pacific (a) and Atlantic (b) basins using the basin definition of fig. S2.2. Subfigure (c) shows a normalised histogram of the depth of all particles, showing that with our initial conditions most of the particles stay relatively close to the surface. The figures are based on the final location of particles after 10 years of advection.

Table S2.1: F-matrix [%]: $z = 0.0$ m

	NP	NA	SP	SA	IO	A	SO
NP	95.1	0.1	0.6	0.2	2.9	1.2	0.0
NA	0.2	92.9	0.0	0.0	0.0	6.9	0.0
SP	2.0	0.0	89.5	4.0	4.5	0.0	0.0
SA	0.0	6.8	6.5	63.8	22.9	0.0	0.0
IO	1.5	0.0	25.0	10.1	63.4	0.0	0.0
A	0.0	31.0	0.0	0.0	0.0	68.9	0.0
SO	0.0	0.0	34.0	14.8	44.2	0.0	7.0

Table S2.2: F-matrix [%]: $z = 2.1$ m

	NP	NA	SP	SA	IO	A	SO
NP	95.7	0.2	0.8	0.1	1.9	1.3	0.0
NA	0.1	91.5	0.0	0.0	0.0	8.4	0.0
SP	2.0	0.0	89.9	3.9	4.2	0.0	0.0
SA	0.0	5.9	5.4	65.0	23.7	0.0	0.0
IO	1.6	0.0	26.0	8.3	64.2	0.0	0.0
A	0.0	32.6	0.0	0.0	0.0	67.4	0.0
SO	0.0	0.0	25.7	16.4	42.8	0.0	15.1

Table S2.3: F-matrix [%]: $z = 4.5$ m

	NP	NA	SP	SA	IO	A	SO
NP	95.3	0.2	1.1	0.1	1.7	1.5	0.0
NA	0.0	88.8	0.0	1.8	0.0	9.3	0.0
SP	1.9	0.0	89.7	3.9	4.5	0.0	0.0
SA	0.0	5.5	4.9	65.6	24.1	0.0	0.0
IO	1.7	0.0	27.1	6.9	64.3	0.0	0.0
A	0.0	38.8	0.0	0.0	0.0	61.2	0.0
SO	0.0	0.0	20.9	17.1	38.9	0.0	23.2

Table S2.4: F-matrix [%]: z = 8.9 m

	NP	NA	SP	SA	IO	A	SO
NP	94.3	0.4	2.1	0.0	1.5	1.7	0.0
NA	0.0	88.4	0.0	2.2	0.0	9.4	0.0
SP	1.7	0.0	90.9	4.0	3.3	0.0	0.0
SA	0.0	4.1	4.1	67.2	24.6	0.0	0.0
IO	1.7	0.1	28.4	5.1	64.7	0.0	0.0
A	0.0	44.9	0.0	0.0	0.0	55.1	0.0
SO	0.0	0.0	16.4	17.8	34.6	0.0	31.1

Table S2.5: F-matrix [%]: z = 15.2 m

	NP	NA	SP	SA	IO	A	SO
NP	93.6	0.2	3.3	0.0	1.2	1.7	0.0
NA	0.1	89.1	0.0	3.2	0.0	7.7	0.0
SP	1.7	0.0	93.0	4.4	0.9	0.0	0.0
SA	0.0	2.6	4.3	68.5	24.5	0.0	0.0
IO	1.4	0.0	30.5	3.7	64.4	0.0	0.1
A	0.0	33.7	0.0	0.0	0.0	66.3	0.0
SO	0.0	0.0	13.7	17.6	31.7	0.0	37.1

Table S2.6: F-matrix [%]: z = 24.6 m

	NP	NA	SP	SA	IO	A	SO
NP	91.2	0.2	5.8	0.0	1.3	1.4	0.0
NA	0.2	88.7	0.0	3.5	0.0	7.7	0.0
SP	1.6	0.0	91.6	5.6	1.1	0.0	0.0
SA	0.0	2.0	7.9	68.2	21.8	0.0	0.1
IO	1.3	0.0	42.7	3.1	52.7	0.0	0.2
A	0.0	43.1	0.0	0.0	0.0	56.9	0.0
SO	0.0	0.0	12.8	16.4	30.4	0.0	40.4

Table S2.7: F-matrix [%]: z = 38.4 m

	NP	NA	SP	SA	IO	A	SO
NP	89.5	0.1	6.9	0.0	2.7	0.7	0.0
NA	0.4	85.7	0.0	5.0	0.0	8.9	0.0
SP	2.2	0.0	89.5	6.2	2.0	0.0	0.0
SA	0.0	3.9	13.1	64.6	18.1	0.0	0.3
IO	1.3	0.0	47.6	4.1	46.6	0.0	0.3
A	0.0	41.4	0.0	0.0	0.0	58.6	0.0
SO	0.0	0.0	13.6	12.8	30.6	0.0	43.0

Table S2.8: F-matrix [%]: z = 57.4 m

	NP	NA	SP	SA	IO	A	SO
NP	81.8	0.1	9.8	0.1	8.2	0.0	0.0
NA	0.4	85.3	0.0	4.1	0.0	10.2	0.0
SP	12.5	0.0	75.6	6.0	5.7	0.0	0.1
SA	0.0	14.9	15.1	49.8	19.4	0.0	0.8
IO	1.2	0.4	32.6	7.5	57.9	0.0	0.4
A	0.0	29.7	0.0	0.0	0.0	70.3	0.0
SO	0.0	0.0	14.8	9.2	30.0	0.0	46.0

Table S2.9: F-matrix [%]: z = 82.2 m

	NP	NA	SP	SA	IO	A	SO
NP	79.4	0.1	11.9	0.2	8.5	0.0	0.0
NA	0.8	85.8	0.0	0.9	0.0	12.5	0.0
SP	13.1	0.0	74.8	4.9	6.8	0.0	0.3
SA	0.0	11.6	18.5	53.4	13.8	0.0	2.6
IO	0.9	0.5	17.3	10.8	69.2	0.0	1.3
A	0.0	22.8	0.0	0.0	0.0	77.2	0.0
SO	0.0	0.0	17.5	9.3	16.4	0.0	56.7

Table S2.10: F-matrix [%]: z = 91.9 m

	NP	NA	SP	SA	IO	A	SO
NP	80.2	0.1	10.6	0.3	8.8	0.0	0.0
NA	0.6	86.1	0.0	0.8	0.0	12.5	0.0
SP	12.9	0.0	75.5	4.6	6.6	0.0	0.5
SA	0.0	11.5	16.9	55.4	12.4	0.0	3.8
IO	0.8	0.5	14.4	11.1	71.1	0.0	2.1
A	0.0	22.8	0.0	0.0	0.0	77.2	0.0
SO	0.0	0.0	14.1	10.2	14.8	0.0	60.9

Table S2.11: F-matrix [%]: z = 113.9 m

	NP	NA	SP	SA	IO	A	SO
NP	82.7	0.0	8.1	0.3	8.9	0.0	0.0
NA	0.5	85.5	0.0	0.9	0.0	13.0	0.0
SP	11.1	0.0	76.9	4.5	6.5	0.0	1.0
SA	0.0	12.3	13.9	56.6	10.5	0.0	6.7
IO	0.6	0.6	11.2	10.4	73.5	0.0	3.7
A	0.0	19.1	0.0	0.0	0.0	80.9	0.0
SO	0.0	0.0	9.1	10.6	13.0	0.0	67.3

Table S2.12: F-matrix [%]: Random Uniform

	NP	NA	SP	SA	IO	A	SO
NP	84.2	0.8	5.0	0.3	6.5	3.2	0.0
NA	0.1	82.3	0.0	3.2	0.0	14.4	0.0
SP	4.0	0.0	82.1	7.3	6.6	0.0	0.0
SA	0.0	13.6	15.0	52.0	18.8	0.1	0.5
IO	1.6	0.2	39.9	8.1	50.0	0.0	0.2
A	0.0	46.3	0.0	0.0	0.0	53.7	0.0
SO	0.0	0.0	19.8	10.1	33.5	0.0	36.6

Table S2.13: F-matrix [%]: Kukulka mixing

	NP	NA	SP	SA	IO	A	SO
NP	95.4	0.1	0.6	0.1	2.0	1.7	0.0
NA	0.2	92.1	0.0	0.0	0.0	7.8	0.0
SP	2.0	0.0	89.5	4.5	4.0	0.0	0.0
SA	0.0	6.3	4.9	65.0	23.8	0.0	0.0
IO	1.7	0.0	24.9	8.7	64.7	0.0	0.0
A	0.0	30.7	0.0	0.0	0.0	69.3	0.0
SO	0.0	0.0	22.2	18.9	46.0	0.0	12.8

Table S2.14: F-matrix [%]: 3D Simulation

	NP	NA	SP	SA	IO	A	SO
NP	75.9	0.8	8.1	0.4	12.1	2.6	0.0
NA	0.1	83.6	0.0	1.8	0.0	14.5	0.0
SP	7.2	0.2	79.9	5.7	7.0	0.0	0.0
SA	0.0	18.7	5.3	49.7	26.1	0.2	0.0
IO	1.6	0.8	17.9	9.4	70.3	0.0	0.0
A	0.0	61.5	0.0	0.1	0.0	38.4	0.0
SO	0.0	0.5	23.9	22.4	38.4	0.0	14.8

3

Mixing of passive tracers at the ocean surface and its implications for plastic transport modelling

The tracking of virtual particles is one of the main numerical tools to understand the global dispersion of marine plastic debris and has been successful in explaining the global-scale accumulation patterns of surface microplastic, often called 'garbage patches'. Yet, the inherent inaccuracies in plastic input scenarios and ocean circulation model results produce uncertainties in particle trajectories, which amplify due to the chaotic property of the surface ocean flow. Within this chaotic system, the subtropical 'garbage patches' correspond to the attractor. These facts make the large scale surface ocean circulation a mixing dynamical system, which means that the information of a particle's initial location is lost over time. We use mixing entropy and Markov chain mixing of the transfer operator associated with surface ocean transport to quantify the time scales of mixing for the global surface ocean in each subtropical basin. In the largest parts of all basins we find mixing times in the order of or below 10 years, which is lower than typical simulation times for surface plastic transport simulations. Maximum mixing times of more than 10 years are found in some parts of the North and South Pacific. Our results have important implications for global dispersion modelling of floating materials on the basin scale: precise initial information has little relevance for long term simulations, and there is a temporal limit after which the backtracking of particles is not meaningful any more.

This chapter has been published as:

Wichmann, D., Delandmeter, P., Dijkstra, H. A., and van Sebille, E. (2019): Mixing of passive tracers at the ocean surface and its implications for plastic transport modelling. *Environmental Research Communications*, 1(11), 115001. <https://doi.org/10.1088/2515-7620/ab4e77>

3.1. Introduction

Plastic debris is very common in today's marine environment. It can be found in the open ocean and on beaches, but also in more remote habitats such as the polar regions and deep sea sediments [39, 48, 63, 71]. Sources of debris can be land based (e.g. river mouths) or marine (e.g. shipping vessels) [47, 64, 71]. Once in the ocean, buoyant plastic debris can be transported over large distances by the surface ocean currents. An important tool to understand these pathways on the global scale has been numerical modelling [18], mostly Lagrangian particle tracking. Most of the currently existing numerical studies restrict themselves to the surface ocean, and they successfully reproduce the large scale features of surface microplastic distributions, which are the large scale accumulation patterns in each individual subtropical gyre [4, 14, 16, 17, 25, 46]. Figure 3.1 shows the result of a typical simulation of an initially uniform particle distribution after 10 years (see section 3.2 for simulation details). The high concentrations in the subtropical gyres are mainly caused by the converging Ekman currents in the subtropics [15, 72] and are often called 'garbage patches'.

Figure 3.1: Concentration of approximately 1 million initially uniformly distributed particles after 10 years. The concentration (logarithmic scale) is computed with a 2° square binning. The converging Ekman currents in the subtropics lead to the accumulation of particles in these regions, called garbage patches. The simulation details are described in section 3.2.

The success of these surface transport simulations is to some extent surprising. Recent experimental and theoretical studies clearly suggest that initially buoyant microplastic particles are likely to sink due to biofouling [53–55], such that the actual particle dynamics is unlikely to be well-represented by surface transport alone. Furthermore, the accumulation in the subtropical gyres is a feature that is present in long term distributions of different modeling studies that use different initial particle input scenarios (e.g. uniform, realistic, release at once or continuous) and different circulation models [4, 14, 16, 17, 25]. The presence of large scale robust features in different models to some extent seems to argue against the usage of realistic, small-scale input scenarios to model long term global plastic distributions, e.g. based on population den-

sities [4, 16]. Particles seem to accumulate in similar regions in the long run, with the particle history only playing a minor role.

In this article, we address this issue: our hypothesis is that the surface ocean current in each basin is *mixing* on time scales of several years, which is in the order of typical time scales of interest for global plastic transport modelling. The mixing property means that individual surface particles ‘forget’ their history, and *particle densities*, i.e. particle distributions normalised to unity, converge to the subtropical accumulation regions almost independently of initial conditions or small perturbations of the flow, such as sub-surface advection. The basic idea is illustrated in fig. 3.2, that shows the initial (a) and final (b) positions of virtual particles as a scatter plot in the North Pacific. In the figure, particles originating from the eastern and western parts of the basin are labelled with different colours. While there are a few regions (mostly close to coasts) that remain dominantly red or blue, the colours apparently have completely mixed within the centre of the basin after 10 years.

Figure 3.2: Initial and final positions of particles that are coloured according to their initial position. The red and blue particles are apparently completely mixed after 10 years in the centre of the subtropical gyre. We refer for simulation details to section 3.2.

For the application we present here, the property of ‘history loss’ of an individual particle results from the combination of two features that are inherent to all simulations using Lagrangian particle tracking on the oceanic scale: the chaotic behaviour of trajectories, and the finite accuracy of ocean circulation models. For plastic modelling, plastic input scenarios are an additional source of uncertainty. The chaotic behaviour implies that two initially close particles will separate exponentially and thus very quickly go independent pathways [13]. Small errors in initial conditions, differences in trajectory integration schemes and numerical errors amplify, such that the correlation between initial and final particle densities decays over time. This *decay of correlations* is expected to be present in a chaotic system such as the surface ocean, but to our knowledge the time scales of this process at the ocean surface have not been analyzed so far.

Understanding these time scales is important due to several reasons. From a practical perspective, specifying detailed initial particle positions is not necessary if particles are advected for much longer than the typical mixing time

scale. This also means that far beyond such a time scale, initial information does not have much relevance for final particle densities, i.e. for particle distributions normalised to unity. Consequently, such a mixing time also poses a natural limit for *backtracking*, i.e. when using Lagrangian particle tracking to investigate where plastic debris or other floating material came from. This is again illustrated in fig. 3.2b, where blue and red particles are mixed to length scales below the typical spatial resolution of ocean circulation models. It is then virtually impossible to accurately track individual particles back to their origin: loosely speaking, almost any origin (here east or west) within a specific basin is possible.

In this paper, we analyse the convergence to the subtropical accumulation regions and the property of mixing at the ocean surface with two methods: mixing entropy for a full time dependent particle transport model, and the concept of mixing time of the Markov chain given by the discretised transfer operator associated with the annual advective transport of particle densities. Both full particle simulations and transfer operator methods are common tools to model long term marine plastic debris transport, such that the relevance of our results can be directly interpreted within these two frameworks. Our methods reveal that on time scales of about 10-15 years, mixing is very relevant for basin scale transport simulations.

3.2. Methods

All methods in this paper are based on the advection of virtual passive particles constrained to the global ocean surface. Surface ocean velocity fields are obtained from a 1/12° global NEMO ORCA-N006 simulation [24] that is forced by the Drakkar forcing consisting of wind, heat and fresh water fluxes derived from reanalysis and observed data [68]. The hydrodynamic data is provided on the ORCA grid [69]. Our dataset starts on January 5, 2000, and has a temporal resolution of 5 days. We use the Parcels framework version 1.1.0 [66] (<http://oceanparcels.org>) for the particle advection with the C-grid interpolation scheme described in [67]. Trajectories are integrated with the 4th-order Runge-Kutta method, with a time step of 10 minutes. Our simulations capture the transport of passive particles, i.e. we do not include any additional mechanisms of beaching, sinking or other removal from the sea surface. The code for advecting particles and for data analysis can be found on https://github.com/OceanParcels/surface_mixing. We note again that we use the term *particle density* to refer to a particle distribution *normalised to unity*.

3.2.1. Entropy of mixing

The entropy of mixing is the Shannon entropy [73] related to probabilities derived from spatial distributions of particles that belong to different species. It describes how well-mixed particles of different species are, i.e. how fast information on particle species is lost when transported by a given flow (see [74] and references therein).

In order to study the mixing process within each basin (see the definition of the basins in section 3.2.3), we start from an initially uniform distribution of particles on a global $0.2^\circ \times 0.2^\circ$ grid (approximately 1 million particles) on the beginning of our dataset (January 5, 2000), and advect them for 10 years. Each particle receives a fixed label i corresponding to its initial position within the bins of a $5^\circ \times 5^\circ$ lattice. This choice is to ensure that there are enough particles for each label to compute spatial densities based on particle statistics. It is coarser than the choice of binning for the Markov chain models, for which we are not limited by the finite number of particles (see section 3.2.2). Let $\rho_{i,k}(t)$ denote the discretised density of particles with label i evaluated in spatial grid cell k (taken to be $5^\circ \times 5^\circ$ as well), computed based on the particle distribution at time t . The (local) entropy of mixing at time t and bin k (called $S_k(\text{species})$ in [74]) is then defined as:

$$S_k(t) = - \sum_i p_{i|k}(t) \ln p_{i|k}(t), \quad (3.1)$$

where

$$p_{i|k}(t) = \frac{\rho_{i,k}(t)}{\sum_i \rho_{i,k}(t)}$$

is the conditional probability to find a particle of label i in a fixed bin k after advection for time t . The entropy is a measure of how strongly particles of different types are mixed within a cell k , larger values of S_k corresponding to stronger mixing. With our definition of particle labels, the entropy of mixing is a measure of the loss of the information regarding the origin of a particle. Note that we only use the densities $\rho_{i,k}$ to compute the entropy, which is independent of the actual number of particles placed in a respective bin i , as long as it is unequal of zero. Consequently, our results also apply to globally non-uniform particle distributions that initially have a non-vanishing number of particles in each bin.

As we are interested in the mixing property in each basin separately, we compute S_k for each individual basin by first dividing the ocean into five ocean basins (see section 3.2.3). This is justified by the fact that the different subtropical gyres are only weakly connected [45] on the time scales of interest for plastic transport modelling (few decades). The maximum entropy is reached for a distribution where all $p_{i|k}$ are the same and therefore equal to $\frac{1}{M}$, where M is the number of labels in a respective basin. This yields a maximum value of $S_k^{\max} = \ln M$ for all k . Note that entirely particle depleted regions have $S_k = 0$ by definition.

3.2.2. Markov chain mixing

The second method is based on approximating ocean surface transport by a finite Markov chain given by a *transition matrix*. Transition matrices have been used to model passive particle advection at the ocean surface based on surface

drifters [4, 14] and based on ocean circulation model output [45]. Our approach to compute the transition matrix is similar to the one of [45].

We compute a global *annual* transition matrix by advecting an initially uniform particle distribution, with particle spacing of 0.1° in longitude and latitude, for a certain time interval Δt . We then divide the ocean into N square-bins of some size $\Delta x \times \Delta x$ in longitude and latitude, and calculate the share $P_{ij}^{t \rightarrow t+\Delta t}$ of particles travelling from a bin $i = 1, \dots, N$ at time t to any other bin $j = 1, \dots, N$ after a time interval Δt . This share is then interpreted as the probability for a particle starting at bin i to travel to bin j in time Δt . Similar to [45], we use $\Delta t = 60$ days and $\Delta x = 2^\circ$ for most parts of this paper, and form the product of six matrices to model the advection for 360 days: $T = P_{t_0 \rightarrow t_1} \cdot P_{t_1 \rightarrow t_2} \cdot \dots \cdot P_{t_5 \rightarrow t_6}$, where $t_0 =$ January 1, 2001, $t_1 = t_0 + 60$ days, $t_2 = t_1 + 60$ days, etc. For computational convenience, we model annual transport as the transport for 360 days. Note that the choice of $t_0 =$ January 1, 2001 is different from the starting time for the entropy method because we want the Markov matrix to correspond to one specific year and our dataset for the ocean current starts only on January 5, 2000. We also vary the parameters Δx , Δt and t_0 to analyse the robustness of our results.

The matrix T models the advection of a discretised particle density for one year as a stochastic Markov chain. If $\rho(t)$ is an N -dimensional row-vector representing such a density at time t (the beginning of the year), the density one year later can be computed based on the transition matrix as

$$\rho(t+1) = \rho(t)T. \quad (3.2)$$

Note that transport modelling with transition matrices is expected to be more diffusive than with individual particles [75]. In addition, modelling the advection of tracers with annual transition matrices assumes the ocean current to be time-periodic. Despite these drawbacks, transition matrices take into account some of the common problems of plastic transport modelling at the ocean surface: initial conditions and the ocean currents are not known to infinite precision. In addition, tools from finite Markov chain theory are readily available to study the mixing behaviour.

From Markov chain theory, it is known (e.g. [28]) that the transition matrix has at least one *stationary density* π satisfying

$$\pi = \pi T \quad \text{with} \quad \pi_i \geq 0, i = 1, \dots, N, \sum_{i=1}^N \pi_i = 1. \quad (3.3)$$

The stationary density of the Markov chain is sometimes also called *invariant measure*. If the Markov chain is irreducible and aperiodic, there is only one such stationary density and an arbitrary initial density will be attracted to this unique stationary density in the course of time [28].

For the ocean, however, each individual subtropical basin has an invariant measure, the ‘garbage patches’, such that the Markov chain is not irreducible. As we are interested here in the mixing process within each individual basin, we

project the transition matrix onto each basin separately (section 3.2.3) and normalise the rows to obtain five Markov chains, one for each subtropical basin. Each of these five Markov chains is irreducible and aperiodic and therefore has one unique stationary density corresponding to the ‘garbage patch’ of the respective basin. It is important to note that the projection is an approximation, as some of the garbage patches are only approximately stationary [4, 76] in finite time (a few decades), see section 3.2.3.

Let Ω denote the ocean surface in a basin. The *total variation distance (TVD)* between a density ρ and π , is defined by (see e.g. [28]):

$$d(\rho, \pi) = \max_{A \subseteq \Omega} |\rho(A) - \pi(A)| = \frac{1}{2} \sum_i |\rho_i - \pi_i|. \quad (3.4)$$

Here, $\rho(A)$ denotes the probability to find a particle in the set $A \in \Omega$ given the probability density ρ of particles over the ocean surface Ω . As both ρ and π are densities, i.e. $\sum_i \rho_i = \sum_i \pi_i = 1$, it is easy to see that $0 \leq d(\rho, \pi) \leq 1$, where $d(\rho, \pi) = 1$ if the two densities ρ and π have disjoint support, i.e. when $\rho_i \pi_i = 0$ for every i . Also, $d(\rho, \pi) = 0$ if the densities are equal to each other. For an irreducible aperiodic Markov chain, i.e. a Markov chain with a unique invariant measure π , the distance $d(\rho(t), \pi)$ of an arbitrary initial density $\rho(t=0)$ to the stationary density π decreases over time and converges to zero. Let δ^k denote the density with $\delta_i^k = 1$ if $i = k$ and 0 otherwise. For each k , we define the mixing time t_k as the smallest integer n (in years) that satisfies

$$d(\delta^k T^n, \pi) < \epsilon. \quad (3.5)$$

As is common in Markov chain theory [28], we take $\epsilon = 1/4$, but also test $\epsilon = 1/10$ to see how the mixing time changes.

Finally, it is easy to see that the maximum mixing time for all the δ^k in a basin is also the maximum for all other possible densities, as a discretised density in a fixed basin can be written as $\rho = \sum_k a_k \delta^k$, with $\sum_k a_k = 1$. Using this definition of ρ and $\sum_k a_k = 1$, it follows from the triangle inequality that

$$\begin{aligned} d(\rho T^n, \pi) &= d\left(\sum_k a_k \delta^k T^n, \pi\right) \\ &\leq \frac{1}{2} \sum_i \sum_k a_k |[(\delta^k T^n)_i - \pi_i]| \\ &\leq \sum_k a_k \max_j d(\delta^j T^n, \pi) \\ &= \max_j d(\delta^j T^n, \pi). \end{aligned} \quad (3.6)$$

Thus, $d(\rho T^n, \pi) \leq \max_k d(\delta^k T^n, \pi)$ for any ρ in a fixed basin, which means that the mixing time of an arbitrary initial density in a basin is lower or equal to

the maximum mixing time of the individual δ^k . The stationary density π can be computed numerically from the eigenvector corresponding to the eigenvalue $\lambda = 1$ of T , cf. eq. (3.3).

We also test how well the individual stationary densities describe the final accumulation pattern of a realistic plastic input scenario. For this, we advect the plastic inputs with the *global* transition matrix and compute basin wide densities (according to the basins defined in section 3.2.3) after each year. In this case, different from restricting the dynamics to the basin wide matrices, the number of particles in a basin can change over time. This is because particles that are initially outside of any basin can end up in one of the basins, and also due to connections between the different basins in the global transport matrix. We can then compute the TVD of a realistic initial plastic density in each basin advected by the Markov matrix according to eq. (3.4).

We finally also compute a mixing entropy for the transition matrix by setting $\rho_{i,k}(t) = (T^t)_{ik}$, where t is now an integer corresponding to the number of years, which circumvents the problem of particle-depleted regions with mixing entropy $S_k = 0$ by definition (see section 3.2.1).

3.2.3. Division of the ocean surface into basins

As we wish to study the attraction to each of the individual garbage patches, we need to split up the ocean into disjoint regions and restrict the dynamics to these regions. Ideally, we would have to find the basins of attraction for each of the garbage patches. Mathematically, the basin of attraction is the region from which any particle is attracted to the attractor, i.e. each garbage patch has its own region of attraction (e.g. the North Atlantic patch has its region of attraction mostly in the North Atlantic, etc.). In our case however, the basins of attraction will only be approximations, as the ‘garbage patches’ themselves are only approximately stationary [4, 76]. We call our basins of attraction simply *basins* and identify them by making use of the particle dynamics instead of drawing arbitrary geographic boundaries. We find the basins through the following heuristic procedure.

For the entropy method, we divide the ocean into bins of size $2^\circ \times 2^\circ$ and compute a transition matrix P_{ij} containing the share of particles traveling from bin i to bin j in the 10-year period considered. Based on this matrix, we define the ocean basins as disjoint sets B_l defined by a set of lattice points I_l , $l = 1, \dots, 5$ (for the five basins) such that $B_{l,i} = 1$ if $i \in I_l$ and $B_{l,i} = 0$ otherwise. These sets are chosen such that they satisfy $\sum_{j \in I_l} P_{ij} \geq 0.5$ for $i \in I_l$ and $\sum_{j \in I_l} P_{ij} < 0.5$ for $i \notin I_l$. In words, this means that starting within a certain basin, the probability for a particle to be in the same basin after 10 years is at least 50%. We find the sets B_l by initialising them in each basin around the garbage patches (see fig. S3.1). We then compute the probability $p_{l,i}$ for each grid cell i to end up in B_l : $p_{l,i} = \sum_j P_{ij} B_{l,j}$ and add the index j to I_l if $p_{l,j} \geq 0.5$, and remove it if $p_{l,j} < 0.5$. We iterate this procedure until no new grid boxes are added to or removed from B_l . After defining these regions, we restrict ourselves to those particles that still are in their respective basin of origin *after each year of simulation*.

For Markov chain mixing, we compute the 10th power of the annual transition matrix, i.e. $P_{ij} = (T^{10})_{ij}$, and apply the same algorithm on P as for the entropy of mixing to determine the different regions. We choose 10 years because this is a typical time of interest for plastic transport simulations, and because for this time the North Atlantic and Arctic regions are not fully connected yet for $\Delta x = 2^\circ$ and $\Delta t = 60$ days (see fig. S3.2 for the basins based on different powers of T). The two algorithms converge for all basins after less than 20 iterations. Figure 3.3 shows the final definitions of the five regions for the two methods. The two figures are very similar, apart from the Arctic which is more strongly connected to the North Atlantic for the Markov chain method than for the entropy method. Note that the white areas in the figure, such as the Mediterranean and the Arctic, do not belong to any region.

Figure 3.3: Definition of the separate regions for a) entropy of mixing method and b) Markov chain method with $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$.

For reference, fig. S3.3 shows the final concentration of particles that leave their basin of origin for the full 10-year Lagrangian simulation. Figure S3.4 shows the share of effectively deleted particles after projecting the transition matrix for each grid cell. To assess the sensitivity of our results to the chosen parameters, we compute the entropy and transition matrix also for other choices of parameter values, shown in table 3.1. The definition of the basins corresponding to other parameter values for the transition matrix are shown in fig. S3.5.

3.3. Results

3.3.1. Entropy of mixing

As mentioned, we restrict our considerations to particles that are in their respective basin of origin (shown in fig. 3.3a) after each of the 10 years of simulation. For the Pacific and Atlantic basins, more than 80% of the particles remain in their respective basins during that time. Only 65% stay in the Indian Ocean, reflecting the fact that the Indian Ocean itself is relatively leaky in the long term (see also [76] and fig. S3.2). This is also a consequence of the definition of the Indian Ocean region, which extends substantially into the Southern Ocean. We compute the spatial entropy S_k given by eq. (3.1) and divide it by the maximum

Table 3.1: Parameters used in the sensitivity analysis

Method	Case
Mixing entropy bin width	$4^\circ \times 4^\circ$; $5^\circ \times 5^\circ$; $6^\circ \times 6^\circ$
Markov chain 2001 ^a ($\Delta x, \Delta t$)	$(1^\circ, 45 \text{ days}), \epsilon = 1/4$ $(1^\circ, 60 \text{ days}), \epsilon = 1/4$ $(2^\circ, 60 \text{ days}), \epsilon = 1/4$ $(2^\circ, 60 \text{ days}), \epsilon = 1/10$ $(3^\circ, 60 \text{ days}), \epsilon = 1/4$ $(3^\circ, 90 \text{ days}), \epsilon = 1/4$ $(4^\circ, 90 \text{ days}), \epsilon = 1/4$ $(4^\circ, 120 \text{ days}), \epsilon = 1/4$
Markov chain 2005 ^b ($\Delta x, \Delta t$)	$(2^\circ, 60 \text{ days}), \epsilon = 1/4$

^a i.e. $t_0 = \text{Jan } 1, 2001$.

^b i.e. $t_0 = \text{Jan } 1, 2005$.

Note: All figures for the bold scenarios are in the paper.

Figures for the other scenarios can be found in the supplementary material.

entropy of each basin, $S_k^{max} = \ln M$. The result is shown in fig. 3.4 for different times during the simulation. Note that the figure shows the independent entropies for each basin (according to the basin definition in fig. 3.3a) in one joint global figure. As can be seen, the entropy increases during the course of the simulation to values close to 1 in a longitudinal band containing the subtropical accumulation zones. The spatial entropy is equal to its maximum in the largest parts of all basins in fig. 3.4c, indicating a mixing time in the order of six years. It should be noted that the entropy is close to its maximum value not only in the regions of highest accumulation, i.e. the points of highest concentration in fig. 3.1, but also in more extensive regions spanning from west to east in each basin (as can also be seen in fig. 3.2 in the introduction for the North Pacific). Note again that a zero entropy does not mean that there is no mixing taking place at the specific location, but that there are simply no particles at these positions after some time. A change of the bin width from 5° to 4° and 6° leads to similar conclusions (see figs. S3.6 and S3.7).

3.3.2. Markov chain mixing

Figure 3.5 shows the stationary densities in the different basins, given by the eigenvectors of T with eigenvalue 1. The regions of largest density correspond to the high accumulation regions shown in fig. 3.1. Through our choice of basins (section 3.2.3), there is one unique stationary density in each of the five regions, and the individual transition matrices do not have the problem of spurious stationary densities close to the coasts, as was the case in [45]. The density in the North Pacific shown in fig. 3.5 roughly corresponds to the high concentration of plastic measurements in the North Pacific, as shown in figure 1 of [25]. The same quantity in the North Atlantic seems to be slightly shifted to the west. Comparison to the other basins are difficult due to the scarcity of the

Figure 3.4: The spatial entropy S_k defined with a bin width 5° , scaled by the maximum entropy for each basin separately according to the definitions in fig. 3.3a.

data. It should be noted, however, that the densities in fig. 3.5 are very similar to the simulation results of the ‘Maximenko model’ in [25], if the latter is considered in each basin separately (see their figure 3a). This is expected, as the stationary density corresponds to the long term evolution of an arbitrary initial density, including the uniform distribution of the ‘Maximenko model’ in [25], without permanent particle release. We note that the second largest eigenvalue modulus is well separated from 1 for all basins and maximal for the North and South Atlantic with $|\lambda_2| = 0.89$. Varying Δt and Δx does not result in a strong obvious change of the stationary densities (see figs. S3.8-S3.14).

Figure 3.6 shows the mixing time t_k defined through eq. (3.5) for each bin of the ocean surface. The mixing time for the other transition matrices are shown in figs. S3.15-S3.22 in the appendix. For $\epsilon = 1/4$, around 70% of the covered ocean surface has mixing times of 10 years or less, and this does not change for other parameter choices of $(\Delta x, \Delta t)$ (see fig. S3.23). Note that this excludes the part of the ocean that is not part of any basin, e.g. the Arctic in fig. 3.3. The time scale of 10 years is comparable to or shorter than the typical time scale of simulation in many plastic transport models, which is typically in the order of a few decades [4, 14, 16, 17, 25, 46]. Some regions in fig. 3.6, in particular parts of the North Pacific and South Pacific, have higher mixing times of more than 10 years. Remarkably, it is also visible in fig. 3.6 that mixing times do not become larger towards land. This means that within the probabilistic description of ocean transport based on a transition matrix, particles with initial positions close to land are equally mixed over the respective attractor as particles starting in the open ocean. The low value of mixing time in the largest parts of the

Figure 3.5: The stationary densities in each basin, corresponding to the eigenvector with eigenvalue 1 of T , with $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$. The colour shows the particle density per $2^\circ \times 2^\circ$ bin, normalised to the maximum density in each respective basin. Note that some bins have a stationary density of zero (e.g. close to the Southern Ocean in panels d and e) and are coloured in grey.

global ocean indicates that the convergence towards the garbage patches happens on relatively fast time scales. Setting $\epsilon = 1/10$ in eq. (3.5) changes the mixing times of course, but still more than 70 % of the covered ocean surface has a mixing time of 15 years or less (see figs. S3.22-S3.23).

We wish to emphasise that the maximum Markov mixing time is also the maximum mixing time for any arbitrary initial particle density in the basin, when the dynamics is restricted to that particular basin, as was demonstrated in eq. (3.6).

We now look at the advection of a realistic plastic input scenario, used in the ‘van Sebille model’ in [25], shown in fig. 3.7a. We first advect this input scenario (given on a 1° grid) with the *basin wide* projected transition matrices of $(\Delta x, \Delta t) = (1^\circ, 60 \text{ days})$ and compute the TVD for each year and each basin to the corresponding stationary densities. By doing so, we exclude all the particles that are not initially placed in one of the five regions (see figure S3b). The resulting TVDs are shown in fig. 3.7b. We see that the TVD decreases over time and tends towards zero for all basins. Figure 3.7c shows the TVDs of the full initial

Figure 3.6: The Markov mixing time t_k for each basin, with $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$.

particle distribution advected by the global transition matrix to the densities in each basin (cf. section 3.2.2). The mixing times in fig. 3.7c are slightly larger than in fig. 3.7b. This is expected, as we also include initial positions of particles that are not part of a basin. In addition, the individual basin wide stationary densities can together only approximate a global stationary density because interconnections between the different basins exist in the global transition matrix, and because some particles. The restriction to the basins for fig. 3.7c leads to a reduction of the total number of particles to about 77%. The remaining 23% are mostly in the Arctic and Mediterranean after 25 years of advection, see fig. S3.24. The mixing time according to fig. 3.7c is still less than 15 years, much smaller than the maximum mixing time for the matrix with $(\Delta x, \Delta t) = (1^\circ, 60 \text{ days})$, which is 25 years (see table S3.1 in the supplementary material). This also shows that our restriction to separately study the mixing property of the individual basins is justified.

To test the sensitivity of our results on the choices of spatial and temporal resolution Δx and Δt , we determine the mean mixing time for $\epsilon = 1/4$, $t_0 = 2001$ for each of the Markov matrices in each basin (table 1 in the supplementary material). In all basins except for the North Atlantic, there is a trend for the mean

Figure 3.7: a) Plastic input scenario of the ‘van Sebille model’ in [25]; b) TVDs for the different basins, with dynamics restricted to the individual basins; c) TVDs of the basin wide densities from the full dynamics, cf. section 3.2.2.

mixing time to become smaller for larger Δx , roughly decreasing by one year or less when increasing Δx by 1° in our models. This is intuitive, as larger spatial binning introduces stronger randomization and thus diffusion in the dynamics, such that densities are expected to mix more quickly [75]. Again in all basins but the North Atlantic, the mean mixing time becomes larger with larger Δt , with a maximum of 0.6 years increase when increasing Δt by 30 days. This is again intuitive, as the construction of an annual matrix with a larger Δt requires less matrix multiplications, causing less randomization across grid cells. The North Atlantic plays a somewhat special role, showing no clear dependence of Δx or Δt . Hence, the average mixing time does not change very strongly (order of 1-2 years) when changing the parameters for the Markov matrices (see also fig. S3.23 in the appendix).

The mixing entropy based on the transition matrix by setting $\rho_{i,k}(t) = T_{i,k}^t$ (cf. section 3.2.2) is shown in fig. 3.8. We see that the mixing entropy is at its maximum in most parts of the ocean after 10 years, also containing regions which were black in fig. 3.4 because of the lack of particles. It is interesting to note that the parts of the Southern Ocean that are included in this study keep a relatively low value for the entropy, as does the southeastern part of the North Pacific. This indicates that these regions have mainly outflow into the rest of the respective basin, but only little inflow.

3.4. Conclusion

The chaotic transport of passive particles by the surface ocean is highly sensitive to the choice of initial particle positions and errors in particle advection models. Tracers transported by the ocean currents are therefore mixed in space, and their trajectories gradually lose the correlation to their initial particle positions. On the time scale of several decades, particle clouds are attracted to the subtropical accumulation zone in each ocean basin and are dispersed

Figure 3.8: Entropy S_k computed from the transport given by the Markov matrix T , with $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$.

over the attraction region, almost independent of their initial position.

We used the methods of mixing entropy and Markov chain mixing computed from an annual transition matrix representing the surface flow to analyse the typical mixing time scales of tracers at the ocean surface. All our results are limited to basin scale long term modelling of a few years to decades, and neglect the possibility of a particle to sink or beach, as is similar to previous studies [4, 14, 25, 46]. We apply our methods to individual subtropical basins in order to capture the main dynamics in this time range. Based on a fully time dependent transport model, the entropy method showed that particles from different origins are almost completely mixed after approximately six years in the largest regions of the subtropical basins. The Markov chain method revealed that, while some regions in particular in the northern South Pacific have mixing times up to 25 years, the mixing times are in the order of or below 10 years for 70% of the covered surface ocean area.

The effect of mixing and history loss of passive tracers is relevant and needs to be considered for global Lagrangian particle simulations at the ocean surface, in particular if initial particle positions are correlated with final distributions. This applies in particular to plastic transport modelling, where simulation times are typically in the order of a few decades, i.e. precisely when mixing is relevant. From an environmental perspective, our findings imply that plastic is transported over the entire accumulation zone within a time scale of about 10-15 years, regardless of the sources. Consequently, mitigating marine plastic pollution needs common efforts of all neighbouring countries. Yet, our 'plastic particles' are passive particles at the ocean surface, which is an incomplete

model of real plastic particles. To model realistic plastic particles and analyse their transport and fate is still one of the major challenges in the plastic community, and the global mixing properties of real plastic might be different from the findings presented here.

It should be noted that our results on mixing apply solely to basin wide long term transport processes, where particle distributions converge to the respective attractors, i.e. we neglect transient behaviour. On more local scales and shorter time scales, initial particle locations will matter, and realistic input scenarios can be very relevant for transport pathways and final distributions.

The mixing property analysed in this study is mainly a consequence of incomplete knowledge of initial particle positions and the ocean circulation. We parameterised these uncertainties in an abstract manner by the choices of labels for the entropy method, and the spatial binning and time step for the Markov chain approximation. These parameters are clearly choices, but we find that the spatial distribution of Markov chain mixing times is only weakly affected by changes in these parameters. Further research is needed to actually quantify the amount of uncertainty that is present in the results of Lagrangian particle modelling, which is a combination of uncertainties from initial conditions and the ocean current data. Yet, regardless of the level of uncertainty present in particle advection models, the ocean surface transport will possess the mixing property due to the chaotic nature of the ocean flow. The presence of mixing needs to be carefully taken into account in the interpretation of results from Lagrangian particle trajectories.

The mixing time that arises from these uncertainties sets a time scale after which initial conditions only have limited influence on final particle densities. Consequently, our results also have important implications for backtracking of individual plastic particles: as any initial particle distribution in a basin is attracted to the respective stationary density, the information of the precise initial distribution is lost over time. On the basin scale it is therefore virtually impossible to infer an initial distribution from a final distribution if the latter is close to the stationary density.

3.5. Supplementary material to chapter 3

Figure S3.1: The initial regions chosen to find approximations of the basins of attraction for the different subtropical gyres (within black rectangles), see methods section of the paper.

Δx [deg]	Δt [days]	North Pacific	North Atlantic	South Pacific	South Atlantic	Indian Ocean
1	45	8.4 [1.0, 15.0]	5.1 [1.0, 17.0]	11.5 [1.0, 34.0]	4.1 [1.0, 8.0]	6.5 [1.0, 20.0]
1	60	8.5 [1.0, 16.0]	5.0 [1.0, 18.0]	11.6 [1.0, 25.0]	4.2 [1.0, 8.0]	6.7 [1.0, 21.0]
2	60	7.5 [1.0, 12.0]	5.1 [1.0, 17.0]	11.0 [1.0, 23.0]	4.1 [1.0, 15.0]	5.8 [1.0, 20.0]
3	60	6.4 [1.0, 11.0]	20.0 ^a [1.0, 43.0]	10.2 [1.0, 20.0]	3.3 [1.0, 7.0]	5.4 [1.0, 21.0]
3	90	7.0 [1.0, 12.0]	4.5 [1.0, 13.0]	10.7 [1.0, 22.0]	3.9 [1.0, 9.0]	5.7 [1.0, 23.0]
4	90	6.5 [1.0, 11.0]	4.3 [1.0, 14.0]	9.8 [1.0, 19.0]	3.3 [1.0, 6.0]	5.4 [1.0, 22.0]
4	120	6.7 [1.0, 11.0]	3.6 [1.0, 11.0]	10.4 [1.0, 21.0]	3.8 [1.0, 8.0]	5.6 [1.0, 18.0]

^a This value is very high due to the inclusion of the Arctic to the North Atlantic in the clustering method.
See the corresponding figures for the clusters and mixing times.

Table S3.1: Mean [min, max] mixing time in years (2001 Markov matrices, $\epsilon = 1/10$.)

Figure S3.2: Different basin definitions for the Markov chain with $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$ for different powers of T , see section 2.3 in the main text. It can be seen that the basin of attraction of the North Atlantic attractor increases towards the Arctic for higher powers. Also, the Indian Ocean attractor ceases to exist for powers higher than 20. No further changes are visible for higher powers.

Figure S3.3: Particle concentration of the particles that leave their initial basin, as defined in fig. 3.3a of the paper with 2° binning. The figure title shows the basin of origin. One can see that the highest concentration of these particles is reached in one of the other garbage patches.

Figure S3.4: Percentage of deleted particles after projecting onto the different basins, computed from the row-sum of the basin-projected matrices for $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$. Note that the share of deleted particles is very low everywhere but outside and at the boundaries of the basins, cf. 3.3b in the paper.

Figure S3.5: Basins for the Markov matrices with other parameters values.

Figure S3.6: Entropy of mixing for a binning of 4° .

Figure S3.7: Entropy of mixing for a binning of 6° .

Figure S3.8: The stationary densities for $(\Delta x, \Delta t) = (1^\circ, 45 \text{ days})$.

Figure S3.9: The stationary densities for $(\Delta x, \Delta t) = (1^\circ, 60 \text{ days})$.

Figure S3.10: The stationary densities for $(\Delta x, \Delta t) = (3^\circ, 60 \text{ days})$.

Figure S3.11: The stationary densities for $(\Delta x, \Delta t) = (3^\circ, 90 \text{ days})$.

Figure S3.12: The stationary densities for $(\Delta x, \Delta t) = (4^\circ, 90 \text{ days})$.

Figure S3.13: The stationary densities for $(\Delta x, \Delta t) = (4^\circ, 120 \text{ days})$.

Figure S3.14: The stationary densities for $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$ and $t_0 = \text{January 1, 2005}$.

Figure S3.15: Markov mixing time for $(\Delta x, \Delta t) = (1^\circ, 45 \text{ days})$.

Figure S3.16: Markov mixing time for $(\Delta x, \Delta t) = (1^\circ, 60 \text{ days})$.

Figure S3.17: Markov mixing time for $(\Delta x, \Delta t) = (3^\circ, 60 \text{ days})$.

Figure S3.18: Markov mixing time for $(\Delta x, \Delta t) = (3^\circ, 90 \text{ days})$.

Figure S3.19: Markov mixing time for $(\Delta x, \Delta t) = (4^\circ, 90 \text{ days})$.

Figure S3.20: Markov mixing time for $(\Delta x, \Delta t) = (4^\circ, 120 \text{ days})$.

Figure S3.21: Markov mixing time for $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$ and $t_0 = \text{Jan 1, 2005}$.

Figure S3.22: Markov mixing time for $(\Delta x, \Delta t) = (2^\circ, 60 \text{ days})$ and $\epsilon = 1/10$.

Figure S3.23: a) Area weighted share and its b) cumulative distribution for the mixing time for different parameter choices of the transition matrix. Note that we only consider the regions that are part of one cluster for the area weighting, compare fig. S3.3. We exclude $(\Delta x, \Delta t) = (3^\circ, 60 \text{ days})$ due to the large difference of the North Atlantic basin to the other models, see figs. 3.3 and S3.5c

Figure S3.24: Global advection of realistic plastic input scenario with the transition matrix of $(\Delta x, \Delta t) = (1^\circ, 60 \text{ days})$, see section 3 in the paper. a) concentration (per 1° bin) of particles that are not in a basin after 25 years of advection. Shown are only concentrations for bins with at least 1 particle per bin; b) share of particles that are in one of the basins over time. The value increases initially because some of the particles start outside of any basin (cf. fig. 3.7a in the paper).

4

Detecting flow features in scarce trajectory data using networks derived from symbolic itineraries: an application to surface drifters in the North Atlantic

The basinwide surface transport of tracers such as heat, nutrients and plastic in the North Atlantic Ocean is organized into large scale flow structures such as the Western Boundary Current and the Subtropical and Subpolar Gyres. Being able to identify these features from drifter data is important for studying tracer dispersal, but also to detect changes in the large scale surface flow due to climate change. We propose a new and conceptually simple method to detect groups of trajectories with similar dynamical behaviour from drifter data using network theory and normalized cut spectral clustering. Our network is constructed from conditional bin-drifter probability distributions and naturally handles drifter trajectories with data gaps and different lifetimes. The eigenvalue problem of the respective Laplacian can be replaced by a singular value decomposition of a related sparse data matrix. The construction of this matrix scales with $O(NM + N\tau)$, where N is the number of particles, M the number of bins and τ the number of time steps. The concept behind our network construction is rooted in a particle's symbolic itinerary derived from its trajectory and a state space partition, which we incorporate in its most basic form by replacing

This chapter has been published as:

Wichmann, D., Kehl, C., Dijkstra, H. A., and van Sebille, E. (2020): Detecting flow features in scarce trajectory data using networks derived from symbolic itineraries: an application to surface drifters in the North Atlantic. *Nonlinear Processes in Geophysics*, 27(4), 501-518. <https://doi.org/10.5194/npg-27-501-2020>.

All code needed to reproduce the results of this chapter is available at https://github.com/OceanParcels/drifter_trajectories_network.

a particle's itinerary by a probability distribution over symbols. We represent these distributions as the links of a bipartite graph, connecting particles and symbols. We apply our method to the periodically driven double-gyre flow and successfully identify well-known features. Exploiting the duality between particles and symbols defined by the bipartite graph, we demonstrate how a direct low-dimensional coarse definition of the clustering problem can still lead to relatively accurate results for the most dominant structures, and resolve features down to scales much below the coarse graining scale. Our method also performs well in detecting structures with incomplete trajectory data, which we demonstrate for the double-gyre flow by randomly removing data points. We finally apply our method to a set of ocean drifter trajectories and present the first network-based clustering of the North Atlantic surface transport based on surface drifters, successfully detecting well-known regions such as the Subpolar and Subtropical Gyres, the Western Boundary Current region and the Caribbean Sea.

4.1. Introduction

The transport of tracers such as heat, nutrients or plastic in the ocean is an important field of research in oceanography [23]. Despite the inherent time dependence of oceanic transport due to turbulence and temporal variations in the forcing, on the large scale, transport is organized into quasi-stationary regions that are characterized by distinct flow properties. Examples include the five major ocean basins, the Subtropical and Subpolar Gyres, the Western Boundary Currents, etc. Clearly, understanding these features is important for studying the dispersal of tracers. At the same time, changes in external conditions such as through climate change might lead to variations in these large scale flow features [77, 78], and it is therefore important to develop methods that identify and characterize them based on oceanographic data sets.

Many methods exist to detect fluid structures such as regions with little fluid exchange, transport boundaries and coherent structures based on Lagrangian trajectory data [79]. While most of these methods are traditionally applied to complete sets of uniformly distributed particle trajectories, recent methods have been successful in identifying coherent structures from incomplete trajectory data [80–82], making them suitable for applications to ocean drifter trajectories with data gaps and different drifter lifetimes. These methods essentially consist of two steps: first, the definition of a measure of similarity $s(n, n')$ or a distance measure $d(n, n')$ that defines how similar or different two trajectories n and n' are; and second, the choice of a clustering algorithm to group trajectories with similar behaviour together. The computational cost and the physical interpretability of trajectory clusters depend on these choices.

[80] embed trajectories in a high-dimensional euclidean space, i.e. they define the distance between trajectories as an abstract euclidean distance (or cosine distance for trajectories on the earth surface) using the entire trajectories, and directly cluster the embedded trajectories with a Fuzzy-c-Means algorithm. [81] define a binary network that indicates if two particles come closer than a cer-

tain distance ϵ , together with spectral clustering, see also Banisch *et al.* [83]. Other methods related to clustering make use of diffusion maps [82] or the dynamical distance of trajectories [35]. The latter three methods all use a spectral relaxation of the normalized cut problem (NCut) for the trajectory classification, which was introduced by [84].

Here, we propose a new and conceptually simple network-based method to identify groups of trajectories that have similar dynamical behaviour. The network is constructed based on ideas from symbolic dynamics, which describes the coarse grained trajectory of a particle given some partition (binning) of the state space. The itinerary of a particle, i.e. the sequence of bins it visited, if known for long times, resolves information much below the bin resolution. Different from previous network-based methods, we make full use of the duality between individual particles and their coarse grained itineraries, which can lead to significant computational advantages. Here, we simplify the itineraries to a minimum: neglecting the time dimension, we represent the trajectory data as a bipartite network connecting particles and bins, with links defined by conditional distributions over symbols. With an appropriate choice of similarity measure, our method allows us to formulate spectral relaxations of the NCut [84] in terms of the singular value decomposition (SVD) of a related data matrix. Setting up this matrix scales with $O(NM + N\tau)$, N being the number of trajectories, M the number of bins and τ the number of time steps. Our method is naturally extendable to incomplete trajectory data and thus readily applicable to ocean drifters. Conceptually, our method is close to detecting minimally mixing fluid regions, so called almost-invariant sets [27], although there are also important differences to this method, cf. section 4.3.4. In our case, almost-invariant sets are represented by the initial conditions of groups of particles, with trajectories of different groups having only little overlap.

We show with a model flow, the periodically driven double-gyre flow, that the method accurately finds almost invariant regions and transport barriers. We also show that the method correctly classifies most of the trajectories in an incomplete data set. Our method can also be used to formulate the clustering problem in a low-dimensional setting by choosing a coarse partition, which can still resolve the leading order flow structures to a high accuracy. We show this for the double-gyre flow with an effectively 9-dimensional formulation of the clustering problem, which still resolves the leading features of the flow up to details much below the coarse graining scale. Our method is designed for detecting quasi-stationary features from ocean drifter trajectories. We therefore emphasize that, owing to the strong simplifications of the particle itineraries, the method is not suitable for the detection of coherent vortices that are transported in a background flow, such as the ‘Bickley Jet’ (discussed e.g. in [79]). Several trajectory based methods have been applied to the ocean drifter data set on the global scale [80, 82]. In addition, transfer operator methods based on virtual particle trajectories have been used to detect almost-invariant sets at the ocean surface [45]. While these methods successfully identified the five major ocean basins, each of these basins has an approximate attractor in its

center [45, 85], such that the long-term dynamics of global drifter trajectories is exceptionally low-dimensional. Using the ocean drifter data set of the Global Drifter Program [11], we apply our method to identify prominent flow features in the North Atlantic Ocean only, and present here the first drifter based clustering result of the North Atlantic surface flow using network theory. Using data from around 8,300 drifter trajectories, we successfully detect well-known features such as the boundary between the Subtropical and Subpolar Gyres as discussed in [31], the Western Boundary Current, the Caribbean Sea and the separation between the Subpolar Gyre and the Nordic Seas [10].

4.2. Drifter data set

We use daily drifter data derived from the six-hourly interpolated data from the NOAA-AOML global drifter program [11], and constrain the data set to those drifters that were released (but not necessarily stay) in the North Atlantic. The restriction to the North Atlantic leaves us with 5,270 drifter trajectories, starting from 1989. A major challenge in analysing the data set is its inhomogeneity in space and time. Figure 4.1a shows the distribution of drifter locations along all trajectories with a square binning of 2° in both longitude and latitude. It is visible that most of the data is located in the centre of the basin, the subtropical gyre. The accumulation of drifters in the gyre centres is a well-known feature of the basin-scale surface ocean, and is attributed to Ekman convergence in the centre of the gyre, also explaining the accumulation of marine debris in these areas [7, 15]. Figure 4.1b shows the distribution of release locations, again with a square binning of 2° . Drifters are mostly released in regions with strongly repelling properties in order to probe the flow that is rarely sampled in the long term. Most notably, many drifters are released in the Western Boundary Current region (along the US east coast), including the Gulf Stream with vigorous mixing and strong currents. Drifter lifetimes (fig. 4.1c) also vary widely, which poses a challenge if trajectories of different lengths need to be compared. Our algorithm is set up in a way that it naturally handles trajectories of different release times, lengths and lifetimes such that all drifter trajectories are available for our analysis, cf. section 4.3.

4.3. Methods

4.3.1. Preliminaries

Suppose we are given a set of N drifter trajectories at τ time instances $x_n(t)$, $n = 1, \dots, N$, $t = 0, \dots, \tau - 1$. We divide the fluid domain Ω into M disjoint sets (bins) $\{B_m\}$ such that $\cup_m B_m = \Omega$. Given such a partition, a particle trajectory can be described by a symbolic sequence of bin labels $m = 1, \dots, M$, called *itinerary*, which is a representation of the trajectory in terms of symbolic dynamics, see fig. 4.2 for an example. The bin labels $m = 1, \dots, M$ are called the *alphabet* of the symbolic dynamics. We define the coarse grained binary position vector $\delta_n(t) \in \{0, 1\}^M$ of particle n as $\delta_{n,m}(t) = 1$ if $x_n(t) \in B_m$, and $\delta_{n,m}(t) = 0$ otherwise. If the data set is incomplete in the sense that for some particles n , x_n is not defined

Figure 4.1: a: Total counts of all drifter locations (6-hourly for each trajectory), computed with a binning of 2° in both longitude and latitude. b: Distribution of drifter release per 2° -bin. c: Scatter plot of initial drifter location, the colour indicating the drifter lifetime in months.

4

at time t , we simply set $\delta_{n,m}(t) = 0$ for all m . As the sets B_m are disjoint, $\delta_n(t)$ has maximally one non-zero entry at any point in time. With this definition, we define the coarse grained data matrix $\mathcal{C}(t) \in \mathbb{R}^{N \times M}$ as

$$C_{nm}(t) = \delta_{n,m}(t). \quad (4.1)$$

As a function of time, the matrices $\mathcal{C}(t)$ describe the coarse-grained dynamics of the entire data set. Note that $\mathcal{C}(t)$ is very sparse, with maximally N non-zero entries at any time t . Next, we define a matrix G as:

$$G = \sum_{t=0}^{\tau-1} \mathcal{C}(t). \quad (4.2)$$

The matrix G has a simple interpretation: G_{nm} is equal to the number of times that particle n visited bin m at the time instances $0, 1, \dots, \tau - 1$. G defines a bipartite graph, i.e. a graph connecting objects of different type, here particles and bins. In the following, we define the degree vector $d[A] \in \mathbb{R}^h$ of an arbitrary matrix $A \in \mathbb{R}^{h \times l}$ by $d[A]_i := \sum_{j=1}^l A_{ij}$, and the degree matrix $D[A] \in \mathbb{R}^{h \times h}$ as $D[A] := \text{diag}(d[A])$.

4.3.2. Definition of the network

Given a partition $\{B_m\}$, the itineraries of a group of particles can be used to slice the state space into smaller regions. For example, all particles having the same itinerary for a certain number of time steps could be grouped together. With increasing trajectory length, this partitions the state space into smaller and smaller sets of initial conditions that have similar dynamic behaviour. We refer to chapter 14 of the open source book by Cvitanović *et al.* [26] for an introduction to symbolic dynamics and the use of particle itineraries to partition the state space. In applications, requiring exactly equal itineraries is not very

Figure 4.2: Two example trajectories on a binned fluid domain, with alphabet $A, \dots, L$ and $\tau = 7$. The markers along the trajectories represent the points in time the location is stored. The itinerary for the blue particle is: $AAFFGCC$. The red particle has itinerary $GGCD_L_$. Here, $_$ denotes a missing data point, the one at time $t = 4$, indicated by an 'X' in the respective marker. The last data point of the red trajectory is also missing, as its lifetime is shorter than τ time instances.

4

practical. This is because in a chaotic flow two particles that start close may separate exponentially. As the number of all possible itineraries is still very large - there are M^τ of them - there would be no two equal itineraries after a short time.

We therefore define another continuous similarity measure $s(n, n')$ for two particles n and n' , with $0 \leq s(n, n') \leq 1$, based on the particles' itineraries. For clarity, we will write the itineraries in terms of letters rather than numbers, and imagine that each symbol is part of an alphabet with M letters. Missing data is represented by $_$. We require the following properties, using an example itinerary $AABC$:

- (Req. 1) Invariance to permutation: $s(AABC, CABA) = s(AABC, AABC)$
- (Req. 2) Sensitivity to missing data: $s(AABC, _ABC) < 1$, $s(AABC, AA_C) < 1$,
 $s(_ABC, _ABC) < 1$, $s(AABC, _ABC) = s(AABC, DABC)$
- (Req. 3) Zero similarity for disjoint itineraries: $s(AABC, DEFF) = 0$.

Requirement 1 essentially discards the time dimension. Requirement 2 takes into account that an itinerary with data gaps contains less information than a full itinerary, and that a missing data point should be treated just as another symbol ('D') that is not part of the example itinerary. Requirement 3 states that completely different itineraries have zero similarity. The easiest way to implement requirements 1-3 is through introducing a conditional symbol distribution $p(m|n) \in \mathbb{R}^M$, defined for each particle n and symbol m , and an appropriate choice of similarity measure between these distributions. We define the distribution by normalization of the individual symbol counts with the *total* trajectory length, i.e. including data gaps. For example, a particle n with itinerary 'AB_FCCH' has

$$p(A|n) = p(B|n) = p(H|n) = p(F|n) = 1/7, \quad p(C|n) = 2/7. \quad (4.3)$$

All requirements are fulfilled with the similarity measure

$$s(n, n') = \sum_m p(m|n)p(m|n'). \quad (4.4)$$

Identifying each letter in the symbolic alphabet with a number $m = 1, \dots, M$, we can directly relate $s(n, n')$ to the matrix G defined in eq. (4.2) as $s(n, n') = \frac{1}{\tau^2} \sum_m G_{nm} G_{n'm} = \frac{1}{\tau^2} (GG^T)_{nn'}$. For spectral clustering with the normalized cut (NCut), the constant $\frac{1}{\tau^2}$ is irrelevant. Thus, we define the following network of similarities on particle trajectories:

$$Q := GG^T \in \mathbb{R}^{N \times N}. \quad (4.5)$$

This is the projection of the bipartite network G onto particle space, see e.g. section 10.4 of the book by Fouss *et al.* [86]. Note that in the case of two non-mixing (invariant) sets, for example as in the autonomous double-gyre flow discussed by [30], network Q defined in eq. (4.5) can be brought into block-diagonal structure. This can be achieved with a partition that is optimal in terms of the invariant sets, i.e. if each set can be completely covered by a part of the alphabet. For non-optimal partitions, we expect that it is still possible to detect an imprint of the two invariant sets with a clustering algorithm, which we quickly present in the following section.

4

4.3.3. Normalized cut and spectral relaxation

In this section we sketch the method of solving a relaxed version of the NCut according to [84]. Our methods are equal to the simultaneous K-way NCut method described in [87] and the hierarchical clustering of [84]. The main difference is that our network defined in eq. (4.5) allows to compute an SVD instead of solving the eigenvalue problem of the Laplacian.

Assume we are given an undirected network defined on a discrete set S containing N vertices, with edges given by the symmetric adjacency matrix $Q \in \mathbb{R}^{N \times N}$. We assume that Q is connected. If it is not connected, we focus on each connected component separately. According to Shi and Malik [84], the normalized cut of a partition of the nodes into K sets $S_1, \dots, S_K$, $S = \cup_{i=1}^K S_i$, is defined as

$$\text{NCut}(S_1, \dots, S_K) := \sum_i^K \frac{Q(S_i, S_i^c)}{Q(S_i, S)}. \quad (4.6)$$

Here, $Q(S_i, S_j)$ is the sum of all weights connecting S_i and S_j , i.e. $Q(S_i, S)$ is the sum of all weights connected to S_i . S_i^c denotes the complement of S_i . The term $\frac{Q(S_i, S_i^c)}{Q(S_i, S)}$ appearing in the definition of the NCut in eq. (4.6) is simply the total weight of all the edges connecting a set S_i to its complement relative to the total weight of the set S_i . Clustering a graph according to the NCut refers to finding a partition $\{S_k\}$ such the objective function in eq. (4.6) is minimized. Note that for an increasing number of sets $\{S_k\}$, the NCut can never decrease.

Minimizing the NCut leads to a clustering result that tries to balance the different terms in eq. (4.6), such that the resulting clusters are of approximately equal size in terms of their total relative weight [84]. While this poses no serious problem for detecting large scale flow features in the ocean, it is certainly a limitation for detecting even smaller structures such as eddies or jets in a large ocean domain, and we explicitly exclude such examples from the scope of our method.

As shown in [84], an approximate solution to the problem can be constructed using the eigenvectors of the *symmetric normalized Laplacian* of Q , defined by

$$L_s[Q] = \mathbb{I} - D[Q]^{-1/2} Q D[Q]^{-1/2}. \quad (4.7)$$

Such a solution is only approximate, as constraints of the optimization problem are neglected, hence the term 'spectral relaxation', see also [88] for a discussion of different relaxations of the NCut. $L_s[Q]$ is positive-semidefinite and its eigenvalues $\{\lambda_{s,i}\}$ satisfy $\lambda_{s,0} = 0 < \lambda_{s,1} \leq \dots \lambda_{s,N-1}$. To identify clusters with the NCut under spectral relaxation, we need the eigenvalues and the right eigenvectors, $\{\lambda_{r,i}, v_{r,i}\}$ of the random walk Laplacian [84]

$$L_r[Q] = D[Q]^{-1} Q, \quad (4.8)$$

which are related to the spectrum of $L_s[Q]$ in the following way:

$$\begin{aligned} \lambda_{r,i} &= 1 - \lambda_{s,i}, \\ v_{r,i} &= D[Q]^{-1/2} v_{s,i}. \end{aligned} \quad (4.9)$$

To find K clusters in Q with the NCut under spectral relaxation, one can either apply a hierarchical clustering procedure as done by Shi and Malik [84], or a simultaneous K-way cut, see Von Luxburg [87] for more information. From a computational perspective, in our case the K-way cut is preferable as we can compute the spectrum of the Laplacian directly from the SVD of a related matrix. This is because for the normalized cut with spectral relaxation, we need the eigenvectors corresponding to the K largest eigenvalues of $D[Q]^{-1/2} Q D[Q]^{-1/2}$. As for any two matrices, $d[AB] = Ad[B]$, it follows that these eigenvectors are equal to the left singular vectors corresponding to the K largest singular values of the matrix

$$R := \text{diag}(Gd[G^T])^{-1/2} G. \quad (4.10)$$

Equation (4.10) also motivates to consider the right singular vectors of R . These are indeed under certain conditions related to the almost-invariant sets based on the transfer operator according to Froyland [27], see appendix 4.6.A. The algorithm for the simultaneous K-way cut for Q in our case is:

Algorithm 1: simultaneous K-way clustering

- S1 Compute the first K left singular vectors of $R = \text{diag}(Gd[G^T])^{-1/2}G$,
 $v_{s,0}, \dots, v_{s,K-1}$.
- S2 Compute $v_{r,i} = D[Q]^{-1/2}v_{s,i}$ for $i = 0, \dots, K - 1$.
- S3 Embed the N nodes of the network in $\mathbb{R}^K$ by setting $y_n = (v_{r,0,n}, v_{r,1,n}, \dots, v_{r,K-1,n})$,
 $n = 1, \dots, N$.
- S4 Perform a standard euclidean-space clustering algorithm (here k-Means)
on the N points $y_n \in \mathbb{R}^K$.

We choose this algorithm for the double-gyre flow (cf. section 4.4.1) for comparison with previous methods such as Padberg-Gehle and Schneide [81], Banisch and Koltai [82] and Hadjighasem *et al.* [35]. For the ocean drifter data set (cf. section 4.4.2), we choose a hierarchical method instead of the simultaneous K-way cut, following [84]. The reason is that the hierarchy preserves the most important boundaries of a clustering solution. This is physically desirable, as several of these main boundaries are known to oceanographers. In addition, it simplifies the presentation of our results and different choices of K when combined with a dendrogram. Note also that if a network is initially disconnected, as is the case for the North Atlantic drifters, it is necessary to compute Q in order to determine the individual connected components.

Algorithm 2: hierarchical clustering

- H1 Compute the network Q defined in eq. (4.5).
- H2 Find the largest connected component of the network and restrict Q to it.
Define $\bar{Q}$ as this restriction.
- H3 Compute the eigenvector $v_{s,1}$ of $L_s[\bar{Q}]$.
- H4 Compute $v_{r,1} = D[\bar{Q}]^{-1/2}v_{s,1}$
- H5 Find a cutoff c such that the sets of nodes defined by $S_1 = \{n \in \{1, \dots, N\} \mid v_{r,1,n} < c\}$ and $S_2 = \{n \in \{1, \dots, N\} \mid v_{r,1,n} \geq c\}$ minimize the NCut for the two sets S_1 and S_2 .
- H6 Split the original network into two networks, with respective adjacency matrices Q_{S_1}, Q_{S_2} defined by the projection of $\bar{Q}$ onto these sets.
- H7 For each sub-network, repeat steps H3-H5.
- H8 Choose to split the sub-network that minimizes the generalized normalized cut in eq. (4.6).
- H9 Repeat the steps H3-H8 up to a certain number of sets K .

At each iteration, only one of the clusters is split into two, and the *global* NCut is minimized at each step. This is different from other hierarchical procedures such as in Ma and Boltt [89], where a local coherence ratio is maximized and the number of clusters can double at each iteration. Algorithm 2 was, apart from H2, suggested by Shi and Malik [84] and is preferable if we do not have a specific bound for the coherence of an individual cluster in mind, ensuring to have, at each step, the most important clusters in terms of eq. (4.6). Note that we do not check for each sub-network if it is connected, as the sub-networks derived from the North Atlantic drifters were connected (only one eigenvalue equal to zero). However, this might be necessary for other data sets. We also apply algorithm 2 to the double-gyre model flow to see how the clustering result is affected by the choice of algorithm.

Note that there is no general rule to determine the number of clusters K in algorithm 1, or where to stop the hierarchical clustering procedure in algorithm 2. A popular heuristic to determine K is to look for a prominent gap in a spectrum of L_s and choose K as the number of smallest eigenvalues before that gap [35]. This is however problematic for systems with no prominent spectral gap, which is the case for the systems considered here (cf. section 4.4). For algorithm 1, we therefore compute the clustering results for different values of K to see how the results depend on this choice. For algorithm 2, one can set a maximum value on the cost function in eq. (4.6) as suggested by Shi and Malik [84]. Yet, the cost function is rather abstract, and there is no general rule what this value should be. Here, we compute the clusters up to a certain (arbitrary) order and compare the results to known structures in oceanography.

4.3.4. Comparison to existing methods

Our method aims to detect groups of particles, with trajectories of different groups having only little overlap. In this sense, our method detects groups of particles with little mixing between each other, which is close to detecting almost-invariant sets according to Froyland [27]. Yet our method is different from detecting almost-invariant sets with the transfer operator in several aspects. First, it is based on similarities between individual particles rather than spatial sets (bins), which allows us to cluster on the particle level rather than the bin level. As we will show in section 4.4.1, this can be used to resolve flow features down to scales much below the bin size. Secondly, our method employs the full trajectory information in terms of a particle's symbolic itinerary, rather than just the start and end points or symbols. In practical applications, this can be an advantage compared to the transfer operator framework, as there is no need in assuming Markovian behaviour of the flow given a state space partition, as done by e.g. [45].

There are also major differences between our method and other existing methods that cluster on the particle level [35, 80–82]. First, these methods only compare particles at equal times, while we disregard the time information. This can be a significant advantage in situations where the major features of a flow are approximately stationary, i.e. can be seen as part of a (noisy) autonomous dy-

namical system. In this case, using the time information of drifter trajectories should not be necessary. Especially for the ocean drifter dataset, containing drifters of different starting times and lengths, it would be very difficult if not impossible to find sub-basin large scale structures when restricting to drifters that necessarily overlap temporally, although this is possible on the global scale to identify the basins themselves [80, 82]. Note that simply placing all drifters at the same initial time and proceeding with one of the existing methods would lead to further problems, as there is ambiguity in which point of a trajectory should be taken for the initial time, probably requiring more data processing such as demanding an initially uniform particle distribution. Our method of simplifying the trajectories does not have these problems by construction, and can be readily applied to scarce drifter datasets.

From a computational perspective, setting up one of the sparse matrices $\mathcal{C}(t)$ in eq. (4.1) is $O(N)$ such that computing the matrix G in eq. (4.2) is $O(N\tau)$. Computing $d[G^T]$ is $O(NM)$ as is the product $Gd[G^T]$. In total, computing R of eq. (4.10) is therefore of computational complexity $O(NM + N\tau)$. If we work with R directly, i.e. we use the simultaneous K-way clustering method described in algorithm 1 in section 4.3.3, computing this network is of lower computational complexity as the computation of the networks used by other studies [35, 81, 82]. These methods typically rely on comparing particle positions between all particles at all time instances, i.e. they scale with $O(N^2\tau)$ in the worst case, although a nearest-neighbour search as applicable to the studies of Padberg-Gehle and Schneide [81] and Banisch and Koltai [82] can reduce the N^2 term to something like $N \log N$. Further, the matrix R in eq. (4.10) is sparser than $L_s[Q]$ or can have column dimension (= number of bins M) significantly lower than row dimension (= number of particles N), cf. section 4.4.1. In these cases, computing the SVD of R instead of the eigenvectors of L_s can lead to computational speed up. Finally, it is interesting to note that the computation of the network is faster for coarser partitions, i.e. when particles are connected in the network even when their trajectories are far apart, as the number of bins M decreases. This is opposite to the methods of Padberg-Gehle and Schneide [81] and Banisch and Koltai [82], where computing the network becomes more costly for larger spatial scale parameters (called ϵ in both studies).

The major drawback of our method is the dependence on a reference frame with respect to which the phase space partition and thus the symbolic itineraries are defined. This can be understood when imagining a time-independent flow from a rotating reference frame. The rotation of the reference frame contributes to a particle's itinerary, and, by averaging over different points in time, non-zero similarities between trajectories can result from the sole rotation of the reference frame. Due to this reason, our method can not be applied to strongly time-dependent systems such as the Bickley jet model flow where coherent vortices are transported in a periodic background flow. It is, however, still possible to detect transport boundaries in time-dependent flows such as the periodically driven double-gyre flow, as we show in section 4.4.1, where particle trajectories belonging to different invariant sets can still be distinguished

with a fixed partition.

4.4. Results

4.4.1. Periodically driven double-gyre flow

To test our method, we choose a model flow that has been used for the detection of coherent structures before [30, 80, 82]. The periodically driven double-gyre flow is defined on a domain $\Omega = [0, 2] \times [0, 1]$, with equations of motion:

$$\begin{aligned}\dot{x} &= -\pi A \sin(\pi f(x, t)) \cos(\pi y), \\ \dot{y} &= \pi A \cos(\pi f(x, t)) \sin(\pi y) \frac{df}{dx}(x, t),\end{aligned}\tag{4.11}$$

where $f(x, t) = \epsilon \sin(\omega t)x^2 + (1 - 2\epsilon \sin(\omega t))x$, and $A = 0.25$, $\epsilon = 0.25$ and $\omega = 2\pi$. Similar to Banisch and Koltai [82], we initially place 20,000 particles on the vertices of a uniform grid on the domain $(0, 2) \times (0, 1)$ and compute trajectory outputs for twenty gyre periods with time steps of 0.1, i.e. we have $\tau = 201$. The eigenvectors are computed with the SVD of the matrix R in eq. (4.10). These eigenvectors are then used for the K-way clustering algorithm with k-Means, cf. algorithm 1 in section 4.3.3.

Figure 4.3 shows the result for the clustering of Q , plotted at $t = 0$ and a binning of $\Delta x = \Delta y = 0.04$, i.e. the column dimension of R is $M = 1,250$. In the figure, we include higher order (larger K) splits as well to show the full range of results obtained by our algorithm. For $K = 2$ (fig. 4.3b, and all even values of K), we see a clear separation between the left and right gyres, which is a common feature found by other studies [30, 80, 82]. The figure also resolves to a certain accuracy the expected transport barrier for the blue and brown particles (the filaments extending into the respective other set). For $K = 3$, we separate the gyre centres from their surrounding. Subsequent uneven values of K further split up the gyre centres in slices. Corresponding results for the autonomous double gyre ($A = 1$, $\epsilon = 0$) illustrating the idea of optimal partitions (cf. section 4.3.2) are shown in figs. 4.7 and 4.8 in the appendix.

We emphasize that the clustering of Q in fig. 4.3 does not show any binning structure. Intuitively, a long itinerary can resolve small-scale structures, i.e. distinguish particles that are initially much closer than the typical bin size, see chapter 14 of [26]. The computational cost can be reduced even more if we make the bins larger. Figure 4.4 shows the result of the clustering for $\Delta x = 2/3$, $\Delta y = 1/3$, i.e. $M = 9$, up to $K = 4$. Note that the choice of Δx prevents a preferable binning along the $x = 1$ line. For this choice, the matrix Q has column dimension equal to nine, i.e. the clustering problem is effectively 9-dimensional. The corresponding clustering result still resolves the most dominant structure up to very high resolution: the split between the left and right side is preserved, and even the structure of the transport barrier is very similar to the one in fig. 4.3. The results for $K = 3$ and $K = 4$ are however different from fig. 4.3, as the gyre centres appear smaller. This indicates that only the most prominent struc-

Figure 4.3: Clustering of Q with $\Delta x = \Delta y = 0.04$ of the full data set, i.e. $N = 20,000$ and $\tau = 201$.

tures, here the separation between the left and right sides, are preserved under coarsening the partition. Nevertheless, figs. 4.4c-d do still give an impression about the flow structures at higher orders, though not completely equal to the high resolution case. Note that the singular values (fig. 4.4a) are strongly suppressed compared to the $M = 1,250$ case in fig. 4.3.

Figure 4.4: Clustering of Q with $\Delta x = 2/3$, $\Delta y = 1/3$ of the full data set, i.e. $N = 20,000$ and $\tau = 201$. The most dominant structure is still visible to high accuracy, although the problem is effectively 9-dimensional.

To test the robustness of our method to missing data, we randomly choose 500 out of the 20,000 particles and for the remaining data set randomly delete 80 % of the data points. This is similar to the approach of [82] and [80]. When the data becomes sparser, different nodes become disconnected, which leads to small, noisy clusters that can be identified by multiple singular values equal to 1. We can remedy that by increasing the bin size such that the network becomes connected again. Therefore, we choose $\Delta x = \Delta y = 0.4$. In doing so, we effectively extend the domain in the y -direction to $y = 1.2$ and disregard the fact that the top row of bins is not completely covered with initial conditions.

Figure 4.5 shows the result for the clustering of Q for the incomplete data set, plotted on top of the corresponding clustering result of the full trajectories (i.e. $\Delta x = \Delta y = 0.04$). The result of the incomplete data set roughly agrees with the expected result of the full clustering. For the results shown in fig. 4.5b, 10 out of the 500 labels were assigned incorrectly compared to the full data case computed with $\Delta x = \Delta y = 0.04$ (fig. 4.3), see fig. 4.9 in the appendix for the number of wrongly assigned labels for different bin sizes. Note again the strong suppression of the singular values for the incomplete data case (fig. 4.5a).

Figure 4.5: Clustering of Q for the incomplete data set, where 500 particles are retained and subsequent deletion of 80 % of the data points. a: singular values of R for the incomplete data set, with $\Delta x = \Delta y = 0.4$. b-d: points representing the initial position of those particles that have data at initial time. Background: result for the full trajectory with $\Delta x = \Delta y = 0.04$ (cf. fig. 4.3).

The results for the double-gyre flow illustrate the robustness of our method in identifying the most dominant structures with incomplete trajectory data. Having control over the bin size enables to tune the network such that it stays connected and the major structures can be resolved. At the same time, small-scale features of the flow seem to be resolved, at least to some extent, independent of the bin size using the K-way simultaneous NCut.

We also tested the algorithm for shorter trajectories, cf. fig. 4.10 in the appendix, showing an expected change of the boundary filaments between the left and right sides of the fluid, which mix less in the shorter period of time. To better understand the differences between algorithms 1 and 2 introduced in section 4.3.3, we also applied the hierarchical NCut method to the non-autonomous double gyre flow. A problem arises here for the first split into two clusters, as there is no unique minimum in the objective function in eq. (4.6), i.e. determining the cutoff c in algorithm 2, H5 (cf. section 4.3.3), is ambiguous for $\Delta x = \Delta y = 0.04$. The lack of a unique minimum for the NCut has been observed before for the same model flow by Froyland and Padberg [30] in the transfer operator framework (see their fig. 15) corresponding to the lack of a unique maximum of the coherence ratio there (see proposition 1 in appendix 4.6.A). Yet, for $\Delta x = \Delta y = 0.04$, in our case, the split between the left and right sides along the transport boundary (the $K = 2$ split in fig. 4.3) does not even have a local minimum for the NCut, cf. fig. 4.11, as opposed to the local maximum of Froyland and Padberg [30]. This however changes to a local minimum for $\Delta x = \Delta y = 0.1$ (fig. 4.12) and finally to a global minimum for $\Delta x = \Delta y = 0.2$ (fig. 4.13). A possible explanation of this dependence on bin size is that the addition

of noise (i.e. larger bins) decreases the coherence of the gyre centres compared to the transport boundary, making the latter easier to be detected. Due to the sensitivity of the hierarchical clustering result to the bin size, we test different bin sizes for the clustering of the North Atlantic drifters in section 4.4.2 (cf, fig. 4.15).

4.4.2. Surface drifters in the North Atlantic

We compute the matrix G for the drifter data set with a square binning of 1° and a time step of one day and maximum time of 365 days, i.e. $\tau = 365$. The binning is the same as the one chosen by van Sebille *et al.* [4] for the computation of drifter-derived transition matrices. We discard the time dimension for the trajectories and place every trajectory at $t = 0$ at the location of drifter release. All trajectories exceeding one year are cut into smaller pieces, each of length smaller or equal to one year, which expands our data set from 5,270 to 8,334 trajectories. The network defined by the drifters is not connected, such that we identify the largest connected component prior to the clustering (with the python networkx package). This restriction removes seven out of the 8,334 trajectories, one of which is a short trajectory of nine time steps at the Strait of Gibraltar, and the other six corresponding to invalid drifter data. Note that the size of the largest connected component is expected to decrease with decreasing bin size, as trajectory overlaps are less likely for smaller bins. We cluster the data with the hierarchical NCut algorithm for Q , cf. section 4.3.3.

Figure 4.6a shows the particle labels at $t = 0$. The dendrogram in fig. 4.6c shows which groups of particles are split, the y-axis corresponding to the value of the global NCut before the split. See the caption of fig. 4.6 for the oceanographic names of the different regions. Nine out of the 20 clusters consisted of one or two particles only. These are coloured in grey in the figure, and have a label '.' in the dendrogram.

The first major split separates the Subpolar Gyre and Nordic Seas from the subtropical and tropical North Atlantic. This splits essentially the Subpolar Gyre from the Subtropical Gyre, which compose together a double-gyre system, having some similarity to the one in section 4.4.1. The scarce transport of drifters between these two regions has been studied before [31, 75, 90], and it is promising that we identify this separation first. Compared to Rypina *et al.* [90], the separation is slightly shifted northwards at the western side of the basin, the Western Boundary Current region particles, labelled 'E', extending slightly into the Labrador Sea. This changes if we plot the drifters at their final time, see fig. 4.6b, showing that some particles of group 'E' from the western Labrador Sea are transported into the northern Western Boundary Current region. Note here that the trajectories in the different clusters do have small but non-zero overlap, such that the spatial extent of the clusters can be different at initial and final time.

Next, our algorithm separates the Subpolar Gyre from the Nordic Seas. A relatively clear cut is seen along the Iceland-Scotland ridge. The strength of the transport over the ridge is in fact an old topic in oceanography [10]. Compared

to fig. 4.6a, this separation is slightly less prominent in figure 4.6b, indicating some slow but non-zero flow across the ridge. The next separation that our algorithm detects is between the southern and northern parts of the Subtropical Gyre, which can be explained by the slow clockwise rotation of the gyre. The next split separates out the Caribbean Sea (label 'F'), a region well known to be rich of eddies and able to trap water masses, after which the Western Boundary Current region (label 'E') extends northwards along the east coast of the US into the northern part of the subtropics. We eventually also identify the northern part of the Subpolar Gyre (Irminger Current), and separate the Barents Sea from the Norwegian Sea, as well as the Bay of Biscay as last separation of the hierarchical clustering algorithm.

We also tested our clustering algorithm without constraining the trajectory length of the data set, see fig. 4.14 up to comparable values of the NCut. We still resolve many of the major features of the flow in the North Atlantic such as the Western Boundary Current region and the Caribbean Sea, the Subpolar Gyre and the Norwegian Sea. Others disappear, e.g. the Bay of Biscay and the Irminger Current. This is likely an effect of longer trajectories receiving more weight in the definition of our network, cf. section 4.3.2, or due to the fact that some particle groups mix with each other on longer time scales, cf. section 4.4.1. It could, however, also be an effect of the fewer trajectories compared to the case when we set a cutoff on trajectory length. For example, there is almost no drifter starting in the Barents sea in the full data set, see fig. 4.14.

To test the sensitivity of the clustering result in fig. 4.6 on the bin size, we applied the same method to square bins of 2° and 4° , see fig. 4.15. The main changes to the result in fig. 4.6 occur in the centre and the south of the Subtropical gyre, where no specific structures were expected in the first place. Only the Bay of Biscay is not detected in the 2° case (fig. 4.15a). All other structures such as the boundary between the Subtropical and Subpolar Gyres, the Western Boundary Current region, the Caribbean Sea and the structures in the Nordic Seas are still detected very similar to the result in fig. 4.6.

4.5. Conclusion

We introduce a new and conceptually simple method that enables the fast construction and clustering of particle based networks to detect quasi-stationary regions with similar flow properties. Our method is based on ideas from symbolic dynamics, where a coarse but long particle itinerary can still resolve very detailed structures below the partition size. We implement a conceptually simple form of this idea and construct a bipartite graph that connects particles and bins, with links corresponding to the time-averaged conditional symbol distribution of each particle's trajectory. We use this bipartite graph to define a similarity graph on particle trajectories, to which we apply normalized cut spectral clustering. The bipartite fundament of our method enables us to use singular vectors of a related data matrix to construct a simultaneous K-way clustering solution under the normalized cut with spectral relaxation instead of computing eigenvectors of the normalized Laplacian.

Figure 4.6: a: clustering of Q for the drifter data set with $\Delta x = \Delta y = 1^\circ$. Clusters corresponding to one or two particles are coloured in grey. Oceanographic regions of the individual clusters (if existing): D: Norwegian Sea, E: Western Boundary Current region, F: Caribbean Sea, I: Irminger Current, J: Barents Sea, K: Bay of Biscay. Names of groups of clusters: Subtropical Gyre (A,F,G,E,H,K), Subpolar Gyre (B, I), Nordic Seas (D, J). b: clustering result plotted at final time. c: hierarchy of the different splits. For each split, the (inverted) y-axis shows the respective NCut after the split. Clusters corresponding to one or two particles are labelled by '. '. The initial conditions are randomly shuffled prior to plotting such that no colour dominates another one at the region boundaries.

Our results show that although we reduce the amount of processed data to a minimum by considering distributions over particle itineraries only, our method is powerful in handling incomplete trajectory data and is computationally efficient to implement. The basic idea of our algorithm is rooted in dynamical systems theory and symbolic dynamics, where long and coarse particle itineraries slice the state space up to scales much below the partition size. The construction of the sparse data matrix used for the singular value decomposition (SVD) has computational complexity $O(NM + N\tau)$, where N is the number of particles, M the number of bins and τ the number of time steps. The linear scaling with the particle number is promising for applications to large trajectory data set, although the complexity of the corresponding SVD depends on the sparsity structure of the resulting matrix, which under some parameter choices (such

as very long time scales with fine binning) could become problematic.

Despite the performance and the low computational complexity of our method, the construction of the networks defined with itinerary distributions is to a certain extent ad-hoc. The construction is mostly motivated by practical requirements, i.e. the need to define a reasonable similarity measure between particles that is not too exclusive, satisfies some reasonable behaviour regarding missing data and decomposes into block-diagonal structure for invariant flow regions in ideal cases. Due to completely discarding the time dimension, our method is dependent on a fixed reference frame with respect to which the state space partition is defined. Therefore, it can not detect moving Lagrangian vortices, such as those in the ‘Bickley Jet’ (discussed e.g. in [79]). The method presented here is the most basic way itineraries can be dealt with. When refining the definition of similarities, it is likely that graphs constructed from symbolic itineraries have a large potential for fast and reliable coherent structure detection.

For the double-gyre flow, our method successfully identifies known flow features to relatively high detail in the known transport boundaries. We demonstrate that our algorithm performs relatively well under deleting a large part of the trajectory data, making it suitable for real-world applications. We also show that an a priori low-dimensional definition of the clustering problem through a coarse binning can still detect the major flow features with an accuracy down to scales well below the bin resolution. We finally apply hierarchical clustering to the network constructed from drifter data in the North Atlantic, and successfully detect major flow regions such as the Western Boundary Current region, the Subpolar-Subtropical Gyres and the Caribbean Sea, providing the first drifter based clustering of the North Atlantic surface transport using network theory.

4.6. Appendices

4.6.A. Relation of Q to almost-invariant sets with the transfer operator

For an introduction to finding almost-invariant sets with the transfer operator, see Dellnitz and Junge [91], Froyland [27] and Froyland and Padberg [30]. Assume $\tau = 2$ and denote by $C_0 = C(0)$ and $C_1 = C(1)$ the two required data matrices defined in eq. (4.1). The transition matrix P that approximates the transfer operator from time $t = 0$ to $t = 1$ is by definition related to these matrices by: $P = D[C_0^T]^{-1}C_0^T C_1$. As described by Froyland [27], one can find almost-invariant sets with the eigenvectors of the matrix $\hat{P} = \frac{1}{2}(P + \Pi^{-1}P^T\Pi)$, where $\Pi = \text{diag}(\pi_1, \dots, \pi_N)$ and π corresponding to the invariant measure of the flow, i.e. $\pi P = \pi$. $\hat{P}$ can be seen as a reversible Markov chain with stationary density π , and is the random walk Laplacian $L_r[A_P]$ of the adjacency matrix

$$A_P = \frac{1}{2}(\Pi P + (\Pi P)^T). \quad (4.12)$$

Froyland [27] then proposes to find sets $\{S_1, \dots, S_K\}$, defined by index sets $\{I_1, \dots, I_K\}$ referring to the contained bin labels in each set, such that a generalized coherence ratio $\rho(S_1, \dots, S_K)$ (see eq. (4.13) for a definition) is maximized. We first show that this is equivalent to the NCut problem applied to A_p .

Proposition 1: Minimizing the generalized normalized cut of A_p defined in eq. (4.6) is equal to maximizing the generalized coherence ratio ρ defined by [27].

Proof: By definition,

$$\begin{aligned}
 \rho(S_1, \dots, S_K) &= \sum_{k=1}^K \frac{\sum_{i \in I_k, j \in I_k} \pi_i P_{ij}}{\sum_{i \in I_k} \pi_i} \\
 &= \sum_{k=1}^K \frac{\sum_{i \in I_k, j \in I_k} A_{p,ij}}{\sum_{i \in I_k, j} A_{p,ij}} \\
 &= \sum_{k=1}^K \frac{\sum_{i \in I_k, j} A_{p,ij} - \sum_{i \in I_k, j \notin I_k} A_{p,ij}}{\sum_{i \in I_k, j} A_{p,ij}} \\
 &= K - \text{NCut}(S_1, \dots, S_K). \square
 \end{aligned} \tag{4.13}$$

Note that this relies on the fact that π is invariant under right multiplication by P , from which follows that $\sum_j A_{p,ij} = \pi_i$ in the denominator of lines 2-3 in eq. (4.13). We now show that the right singular vectors of R defined in eq. (4.10) are equal to the right eigenvectors of $\hat{P}$ if the particle measure is invariant and uniform.

Proposition 2: If the particle measure is invariant and uniform, the right eigenvectors of the matrix $\hat{P} = \frac{1}{2} (P + \Pi^{-1} P^T \Pi)$ used by Froyland [27] are the same as the right singular vectors of R defined in eq. (4.10) with $G = C_0 + C_1$ as defined in eq. (4.2).

Proof: First, note that if the particle measure is invariant and uniform in a complete data set, $R = \beta G$ for a constant $\beta > 0$. We therefore have to show that the eigenvectors of $G^T G$ are under the specified conditions the same as the right eigenvectors of $\hat{P}$. As the transition matrix is defined by $\Pi P = C_0^T C_1$, the symmetric adjacency matrix is $A_p := \Pi \hat{P} = \frac{1}{2} (G^T G - C_0^T C_0 - C_1^T C_1) = \frac{1}{2} G^T G - \Pi$. The last equality follows from the fact $C_0^T C_0 = C_1^T C_1 = \Pi$, as the particle measure is invariant. As Π is assumed to be uniform, this proves the result. $\square$

4.6.B. Supplementary figures double-gyre flow

Figure 4.7: Clustering of Q for the autonomous double gyre with $\Delta x = \Delta y = 0.1$, i.e. an optimal partition, and the simultaneous K -way NCut (algorithm 1 in section 4.3.3). The singular values are doubly degenerate as the network Q consists of two disconnected sets.

4

Figure 4.8: Clustering of Q for the autonomous double gyre with $\Delta x = \Delta y = 0.15$, i.e. a non-optimal partition, and the simultaneous K -way NCut (algorithm 1 in section 4.3.3). The main separation at $x = 1$ is still resolved, but only at every even clustering step. The degeneracy of the singular values, see fig. 4.7, is lifted.

Figure 4.9: Share of incorrectly assigned particle labels for the double-gyre flow (simultaneous K -way clustering) for varying $\Delta x = \Delta y$, $K = 2$ and the incomplete ($N=500$) data set, compared to the baseline of the complete data set with $\Delta x = \Delta y = 0.04$, cf. fig 4.3. The values were obtained with 100 random incomplete data sets for each value of $\Delta x (= \Delta y)$. The filled areas show the minimum and maximum ranges, one standard deviation, and the points the corresponding means. It is visible that for too small bin spacing, the number of wrongly assigned labels is very high. This is most likely because the network becomes less connected, such that it is more difficult to find structure in the eigenvector corresponding to the second smallest eigenvalue of $L_s[Q]$. As explained in the main text, for bin sizes that do not exactly fit the domain, the x - and y -directions were artificially extended.

Figure 4.10: Clustering of Q (simultaneous K -way clustering) for the non-autonomous double gyre with $\Delta x = \Delta y = 0.04$ and shorter trajectories of 10 gyre periods, i.e. $\tau = 101$.

Figure 4.11: Hierarchical clustering result of the double gyre using algorithm 2 of section 4.3.3 and $\Delta x = \Delta y = 0.04$. a: NCut for the first split as a function of the cutoff c , where $c = 0$ corresponds to a cut along the transport boundary, the red line indicating the minimum found by our algorithm. b: result of the hierarchical clustering. c: dendrogram representing the hierarchy, where the horizontal lines correspond to the NCut value after each split. There is no global or local minimum at $c = 0$ in (a), which is why the transport boundary is not detected. Instead, two global minima are present in (a), but our implementation selects the left minimum for the first split due to the choice of sampling points.

Figure 4.12: Hierarchical clustering result of the double gyre using algorithm 2 of section 4.3.3 and $\Delta x = \Delta y = 0.1$. a: Ncut for the first split as a function of the cutoff c , where $c = 0$ corresponds to a cut along the transport boundary, the red line indicating the minimum found by our algorithm. b: result of the hierarchical clustering. c: dendrogram representing the hierarchy, where the horizontal lines correspond to the Ncut value after each split. There is only a local minimum at $c = 0$, which is why the transport boundary is still not detected. This situation is similar to the one described by Froyland and Padberg [30] for the transfer operator framework.

Figure 4.13: Hierarchical clustering result of the double gyre using algorithm 2 of section 4.3.3 and $\Delta x = \Delta y = 0.2$. a: Ncut for the first split as a function of the cutoff c , where $c = 0$ corresponds to a cut along the transport boundary. b: result of the hierarchical clustering. c: dendrogram representing the hierarchy, where the horizontal lines correspond to the Ncut value after each split. There is a global minimum at $c = 0$, so that the transport boundary is detected as the first split.

4.6.C. Supplementary figures ocean drifters

Figure 4.14: Clustering of Q for the drifter data set with $\Delta x = \Delta y = 1^\circ$ and no restriction on trajectory length, plotted at the drifter release location. Some main features as seen in fig. 4.6 are still visible, but some details disappear. The hierarchical clustering was stopped at a NCut value comparable to the maximum NCut in fig. 4.6, i.e. about 3.8.

Figure 4.15: Clustering of Q for the drifter data set and different bin sizes, with trajectory lengths restricted to one year, 20 clusters as in fig. 4.6, plotted at initial time and the drifter release location. a: $\Delta x = \Delta y = 2^\circ$. b: $\Delta x = \Delta y = 4^\circ$. The main features of fig. 4.6 are still visible, although some structures change: the Bay of Biscay is not detected in (a), and the structure of the clusters in the centre of the Subtropical Gyre change compared to fig. 4.6. There are also new clusters in black (a, b) and dark green (b). Note that the main features of the North Atlantic Ocean such as the Subtropical-Subpolar Gyre boundary, the Western Boundary Current region, the Caribbean Sea and the Nordic Seas are still detected, see the result in fig. 4.6.

5

Ordering of trajectories reveals hierarchical finite-time coherent sets in Lagrangian particle data: detecting Agulhas rings in the South Atlantic Ocean

The detection of finite-time coherent particle sets in Lagrangian trajectory data using data clustering techniques is an active research field at the moment. Yet, the clustering methods mostly employed so far have been based on graph partitioning, which assigns each trajectory to a cluster, i.e. there is no concept of noisy, incoherent trajectories. This is problematic for applications in the ocean, where many small coherent eddies are present in a large, mostly noisy fluid flow. Here, for the first time in this context, we use the density-based clustering algorithm OPTICS [38] to detect finite-time coherent particle sets in Lagrangian trajectory data. Different from partition-based clustering methods, derived clustering results contain a concept of noise, such that not every trajectory needs to be part of a cluster. OPTICS also has a major advantage compared to the previously used DBSCAN method, as it can detect clusters of varying density. The resulting clusters have an intrinsically hierarchical structure, which allows one to detect coherent trajectory sets at different spatial scales at once. We apply OPTICS directly to Lagrangian trajectory data in the Bickley jet model flow and successfully detect the expected vortices and the jet. The

This chapter has been published as:

Wichmann, D., Kehl, C., Dijkstra, H. A., and van Sebille, E. (2021): Ordering of trajectories reveals hierarchical finite-time coherent sets in Lagrangian particle data: detecting Agulhas rings in the South Atlantic Ocean, *Nonlin. Processes Geophys.*, 28, 43–59. <https://doi.org/10.5194/npg-28-43-2021>

All code needed to reproduce the results of this chapter is available at https://github.com/OceanParcels/coherent_vortices_OPTICS. The data for the virtual particles in the South Atlantic is available on Zenodo [92].

resulting clustering separates the vortices and the jet from background noise, with an imprint of the hierarchical clustering structure of coherent, small-scale vortices in a coherent, large-scale, background flow. We then apply our method to a set of virtual trajectories released in the eastern South Atlantic Ocean in an eddying ocean model and successfully detect Agulhas rings. We illustrate the difference between our approach and partition-based k-Means clustering using a 2-dimensional embedding of the trajectories derived from classical multidimensional scaling. We also show how OPTICS can be applied to the spectral embedding of a trajectory-based network to overcome the problems of k-Means spectral clustering in detecting Agulhas rings.

5.1. Introduction

Understanding the transport of tracers in the ocean is an important topic in oceanography. Despite large-scale transport features of the mean flow, on smaller scales, mesoscale eddies and jets play an important role for tracer transport [7]. Such eddies can capture large amounts of a tracer, and, while transported in a background flow, redistribute them in the ocean. Eddies have been shown to play an important role for the accumulation of plastic [93] and the transport of heat and salt [94]. To quantify the effects of eddies on tracer transport in the ocean, it is necessary to develop methods that are able to detect and track them. Many methods exist to detect such *finite-time coherent sets* of fluid parcels based on different mathematical or heuristic principles [79]. The term ‘finite-time coherent set’ is based on the work of Froyland *et al.* [95], and is in our context defined as a set of particles that stay, in a sense to be made more specific, close to each other along their entire trajectories. Here, for the first time in this context, we make use of the density-based clustering algorithm OPTICS [38] to detect finite-time coherent sets in Lagrangian trajectory data.

The detection of coherent Lagrangian vortices using abstract embeddings of Lagrangian trajectories together with data clustering techniques has received significant attention in the recent literature [35, 37, 80–82, 96, 97]. Using embedded trajectories for the detection of finite-time coherent sets is interesting as it allows one to use sparse trajectory data, and it can in principle be applied to ocean drifter trajectories, as demonstrated by Froyland and Padberg-Gehle [80] and Banisch and Koltai [82] for the detection of the five ocean basins. Yet, most of these methods cluster trajectory data with graph partitioning, which does not incorporate the difference between coherent, clustered trajectories and noisy trajectories that should not belong to any cluster. Graph partitioning has been shown to work in situations where the finite-time coherent sets are not too small compared to the fluid domain [35, 80–82, 97]. For applications to Lagrangian trajectory datasets on basin-scale ocean domains, where multiple small-scale coherent sets (eddies) coexist with noisy trajectories in the background, graph partitioning is however likely to fail. Similar observations were made by Froyland *et al.* [37] for the partition-based clustering approaches based on transfer and dynamic Laplace operators [97]. Although some attempts have been made to accommodate such concepts in hard partitioning,

e.g. by incorporating one additional cluster corresponding to noise [35], this approach is likely to fail for large ocean domains, as discussed by Froyland *et al.* [37] and shown in section 5.4 of this paper. Froyland *et al.* [37] have developed a special form of trajectory embedding based on sparse eigenbasis decomposition given the eigenvectors of transfer operators and dynamic Laplacians. By superposing different sparse eigenvectors, they successfully separate coherent vortices from unclustered background noise.

Motivated by the results Froyland *et al.* [37] obtained by developing a new form of trajectory embedding, we here explore the potential of another clustering algorithm to overcome the inherent problems of partition-based clustering. We use the density-based clustering method OPTICS (Ordering Points To Identify the Clustering Structure) developed by Ankerst *et al.* [38] to detect finite-time coherent sets in large ocean domains, using a very simple choice of embedding (cf. section 5.2). Density-based clustering aims to detect groups of data points that are close to each other, i.e. regions with high data *density*. Our data points correspond to entire trajectories, and groups of trajectories staying close to each other over a certain time interval correspond to such regions of high point density. Different from partition-based methods such as k-Means or fuzzy-c-means, OPTICS does not require to fix the number of clusters beforehand. Further, density-based clustering has an intrinsic notion of a noisy data point: a point does not belong to any cluster (i.e. a finite-time coherent set) if it is not part of a dense region. A more detailed comparison of the method presented here to existing related methods can be found in section 5.4.

Another desirable property of the OPTICS algorithm is its ability to capture coherence hierarchies. In the ocean, coherent sets of trajectories naturally come with a notion of such a hierarchy. For example, the surface flow in the North Atlantic Ocean can be seen as approximately coherent [45], while mesoscale eddies and jets are also finite-time coherent sets of trajectories at smaller scales *within* the North Atlantic Ocean. Froyland *et al.* [37] show how their leading eigenvectors resolve coherent sets at large scales, while small-scale results can be obtained with a sparse eigenbasis approximation of a set of eigenvectors. Similarly, clustering results obtained from OPTICS are typically hierarchical. The main result of OPTICS, the reachability plot, provides this hierarchical information in a simple 1-dimensional graph.

In section 5.4, we first show how OPTICS detects finite-time coherent sets at different scales for the Bickley jet model flow (also discussed e.g. by Hadjighasem *et al.* [79]), successfully detecting the six coherent vortices and the jet as the steepest valleys in the reachability plot. The general structure of the reachability plot also reveals the large-scale finite-time coherent sets, i.e. the northern and southern parts of the model flow, separated by the jet. We then apply our method to Lagrangian particle trajectories released in the eastern South Atlantic Ocean, where large rings detach from the Agulhas Current (e.g. Schouten *et al.* [98]). We detect several Agulhas rings, and on the larger scale also separate the eastward and westward moving branches of the South Atlantic Subtropical Gyre. While the traditional approach to study Agulhas rings is based

on sea surface height analysis (see e.g. Dencausse *et al.* [99]), several methods based on virtual Lagrangian trajectories have been applied to Agulhas ring detection before [32–36]. Our method is different from these approaches in that it is directly applicable to a trajectory dataset, i.e. without much pre-processing of the data. As the OPTICS algorithm is readily available in the scikit-learn library in Python, the detection of finite-time coherent sets can be done without much effort and with only a few lines of code. A further difference is the mentioned intrinsic notion of coherence hierarchy, which allows for simultaneous analysis of trajectory data at different scales. While we mainly focus on the direct embedding of trajectories in an abstract high-dimensional Euclidean space, we also show in appendix 5.6.C that OPTICS can be used to overcome the limits of k-Means clustering in the context of spectral clustering of the trajectory-based network of Padberg-Gehle and Schneide [81].

5.2. Trajectory datasets

5

5.1. Quasi-periodically perturbed Bickley jet

We apply our method to a model system that has been used frequently in studies to detect finite-time coherent sets [35, 79, 81, 82, 97]. The velocity field of the quasi-periodically perturbed Bickley jet [100, 101] is defined by a stream function $\psi(x, y, t)$, i.e. $\dot{x} = -\frac{\partial\psi}{\partial y}$ and $\dot{y} = \frac{\partial\psi}{\partial x}$, with $\psi(x, y, t) = \psi_0(y) + \psi_1(x, y, t)$ consisting of a stationary eastward background flow

$$\psi_0(y) = -UL \tanh(y/L), \quad (5.1)$$

and a time-dependent perturbation

$$\psi_1(x, y, t) = UL \operatorname{sech}^2(y/L) \operatorname{Re} \left[\sum_{n=1}^3 f_n(t) \exp(ik_n x) \right], \quad (5.2)$$

where $\operatorname{Re}(z)$ denotes the real part of the complex number z . We use the same parameter values as Hadjighasem *et al.* [79], with $U = 62.66$ m/s the characteristic velocity of the zonal background flow, and $L = 1770$ km. The parameters in eq. (5.2) are given by $k_n = 2n/r_0$, $f_n(t) = \epsilon_n \exp(-ik_n c_n t)$ with $\epsilon_1 = 0.075$, $\epsilon_2 = 0.4$, $\epsilon_3 = 0.3$, $c_1 = 0.1446U$, $c_2 = 0.205U$, $c_3 = 0.461U$. The domain of interest is $\Omega = [0, \pi r_0] \times [-3000 \text{ km}, 3000 \text{ km}]$, where $r_0 = 6371$ km is the radius of the Earth, and the left and right edges of Ω are identified, i.e. the flow is periodic in x -direction with period πr_0 . Similar to Banisch and Koltai [82], we seed the domain with an initial number of 12,000 particles on a uniform 200×60 grid. For this choice, the initial particle spacing is slightly above 100 km in both directions. We compute the trajectories for 40 days with a time step of one second using the SciPy integrate package. We output the trajectories every day, i.e. we have $T = 41$ data points in time for each trajectory.

5.2. Agulhas rings in the South Atlantic

To test the OPTICS algorithm with a more realistic ocean flow, we simulate surface particle trajectories in a strongly eddying ocean model. Surface velocities are derived from a NEMO ORCA-N006 run [24], which has a horizontal resolution of $1/12^\circ$ and velocity output for every five days. The model is forced by reanalysis and observed data of wind, heat and fresh water fluxes [68], i.e. the currents do not only contain the geostrophic component, as is the case in altimetry-derived currents [33, 37]. For the advection of virtual particles, we use version 1.11 of the open source Parcels framework [66], see oceanparcels.org. The 2-dimensional surface current velocity is interpolated in space and time with the C-grid interpolation scheme of Delandmeter and van Sebille [67], using a 4th order Runge-Kutta method with a time step of 10 minutes. We initially distribute particles uniformly in the ocean on the vertices of a $0.2^\circ \times 0.2^\circ$ grid in the domain $[30^\circ W, 20^\circ E] \times [40^\circ S, 20^\circ S]$, which corresponds to a total number of 23,821 particles. At $30^\circ S$, a spacing of 0.2° corresponds to roughly 20 km. The particles start at January 5, 2000 and are advected for two years. We output the trajectories with a time interval of five days. We only use the first 100 days as data to detect the finite-time coherent sets, i.e. we have $T = 21$ data points for each trajectory, but also look at later times to see how long the rings need to disperse. We provide the used trajectory data for the Agulhas flow as NumPy file on Zenodo [92].

5.3. Methods

5.1. Detecting coherent structures in Lagrangian trajectory data

For N trajectories of dimension D and length T , the trajectory information can be stored in a *data matrix* $X \in \mathbb{R}^{N \times DT}$, where each row results from a particle trajectory by concatenating the different spatial dimensions. The analysis of trajectory data to detect finite-time coherent sets of trajectories [35, 80–82, 96, 97, 102] can be split into two essential steps:

Step 1 **Embedding** of the trajectories in an abstract (metric) space, i.e. $X \rightarrow \bar{X} \in \mathbb{R}^{N \times M}$, where $M \leq DT$. If one uses a dimensionality reduction method, $M < DT$.

Step 2 **Clustering** of the embedded data with a clustering algorithm.

The embedding is necessary to represent the trajectories as points in a metric space. Different options for embedding the trajectories exist, e.g. a direct embedding of the data points along the trajectories [80], or embeddings based on the eigenvectors derived from networks that are defined by physically motivated trajectory similarities [81, 82, 82, 97]. Once an embedding of each trajectory as a point in a metric (typically Euclidean) space is established, one can apply a clustering algorithm. Roughly speaking, clustering algorithms try to identify groups of points that are close to each other as a cluster. Partition-based clustering methods divide the entire data into a (typically fixed) number

of K clusters, such that each data point belongs to a cluster. The most popular method in this category is the k-Means algorithm, which tries to find a given number of K clusters such that the sum of pairwise squared distances of points within a cluster is minimized. Other clustering algorithms contain a concept of 'noisy' data, i.e. data points that do not belong to any cluster, or belong to a cluster only with a certain probability. Examples for the former case are DBSCAN [103], discussed by Schneide *et al.* [96] in the fluid dynamics context, and the here presented OPTICS [38] algorithm. For the latter case, the most popular method is fuzzy-c-means clustering, as discussed by Froyland and Padberg-Gehle [80] in the context of finite-time coherent sets.

Figure 5.1 shows a few possible options for trajectory embedding and clustering that have partially been explored before (see the footnotes in the figure for the combinations used in related studies). For a given trajectory dataset, one can in principle apply an arbitrary combination of embedding and clustering methods. Only a few of the different combinations have been explored so far, and many more options for embedding and clustering as those shown in fig. 5.1 exist. It is important to note that a good choice of embedding and clustering might well depend on the specific problem at hand, and there might be no combination that performs well for all possible situations.

Figure 5.1: Different steps to detect coherent trajectories in Lagrangian data with trajectory clustering. The figure is non-exhaustive, and many more options for embedding and clustering exist. Footnotes: ¹ Froyland and Padberg-Gehle [80]. ² Hadjighasem *et al.* [35], Padberg-Gehle and Schneide [81] and Banisch and Koltai [82] all define networks with spectral embedding and subsequent k-Means clustering. Froyland *et al.* [37] define spectral embeddings defined on dynamic Laplacian and transfer operators. ³ Schneide *et al.* [96].

Most of the studies that use clustering techniques to detect finite-time coherent sets have focused on developing new forms of trajectory embeddings. For example, Hadjighasem *et al.* [35], Padberg-Gehle and Schneide [81], Banisch and Koltai [82] and Froyland and Junge [97] all use different forms of spectral embeddings, together with k-Means clustering. Froyland *et al.* [37] have developed a powerful form of embedding, based on a sparse eigenbasis ap-

proximation. Here, we focus on the clustering step in fig. 5.1, and propose the OPTICS clustering algorithm in the fluid dynamics context. We test the algorithm for three different kinds of embeddings:

- E1 A direct embedding of the trajectory data in a high dimensional Euclidean space, i.e. $M = DT$ (cf. section 5.2).
- E2 A reduction of the trajectory data to a 2-dimensional embedding space using classical multidimensional scaling (MDS, cf. section 5.2). This is mainly to visualize the difference to partition-based k-Means clustering.
- E3 A spectral embedding of the network proposed by Padberg-Gehle and Schneide [81].

In the following sections, we explain in detail the embeddings E1 and E2 and the OPTICS algorithm. We introduce the network embedding E3 together with the corresponding results in appendix 5.6.C.

5.2. Trajectory embedding

Direct embedding

The direct embedding of each trajectory in $\mathbb{R}^{DT}$ is the most straightforward embedding as it requires no further pre-processing of the trajectory data. For simplicity, assume we are given a set of N trajectories in a 3-dimensional space, i.e. $(x_i(t), y_i(t), z_i(t))$ where $i = 1, \dots, N$ and $t = t_1, \dots, t_T$. We then simply define the embedding of trajectory i in the abstract $3T$ -dimensional space as

$$u_i = (x_i(t_0), x_i(t_1), \dots, x_i(t_T), y_i(t_0), y_i(t_1), \dots, y_i(t_T), z_i(t_0), z_i(t_1), \dots, z_i(t_T)) \in \mathbb{R}^{3T}, \quad (5.3)$$

and impose an Euclidean metric in $\mathbb{R}^{3T}$ to measure distances between different embedded trajectories. The resulting embedded data matrix $\bar{X}$ is then simply given by the vertical concatenation of the different embedding vectors. This kind of embedding was also explored by Froyland and Padberg-Gehle [80], together with a fuzzy-c-means clustering. Intuitively, if two trajectories i and j belong to the same finite-time coherent set, the corresponding particles follow very similar pathways, i.e. the Euclidean distance of the embedding vectors $d_{ij} = \|u_i - u_j\|$ is expected to be small. On the other hand, a particle i that belongs to a coherent set is expected to have a larger distance to a particle j that is not part of the set. In other words, groups of particles that form a finite-time coherent set are *dense* in the embedding space. This motivates to use a density-based clustering algorithm to detect finite-time coherent sets.

To take into account the πr_0 -periodicity in x-direction of the Bickley jet flow, we first put the individual 2-dimensional data points on the surface of a cylinder with radius $r_0/2$ in $\mathbb{R}^3$, and interpret the resulting trajectories in a 3-dimensional Euclidean space. The resulting data matrix is $\bar{X} \in \mathbb{R}^{N \times 3T}$, with $N = 12,000$ and $T = 41$. For the Agulhas particles, we put the single data points on the earth

surface in a 3-dimensional Euclidean embedding space by the standard coordinate transformation of spherical to Euclidean coordinates. The resulting data matrix is thus $\bar{X} \in \mathbb{R}^{N \times 3T}$ with $N = 23,821$ and $T = 21$.

Dimensionality reduction with classical multidimensional scaling

To get an intuition for what the OPTICS algorithm does, and the differences to k-Means, we wish to visualize the data structure in the plane. For this, it is necessary to reduce the embedding dimension of each trajectory from $3T$ to two in a way that the density structure, and hence the individual Euclidean distances between embedded trajectories $d_{ij} = \|u_i - u_j\|$, cf. eq. (5.3), are preserved. We do so by a common method of nonlinear dimensionality reduction, called classical multidimensional scaling (MDS), see e.g. chapter 10.3 of Fouss *et al.* [86]. Classical MDS tries to find an embedding of the high-dimensional data points in a low dimensional space such that the pairwise distances are approximately preserved. Similar to a principal component analysis, classical MDS makes use of the eigenvectors corresponding to the largest eigenvalues of a kernel matrix, which is in this case defined by

$$B = -\frac{1}{2}H\Delta^2H, \quad (5.4)$$

where $\Delta^2 \in \mathbb{R}^{N \times N}$ is a matrix containing all squared distances between the points, $\Delta_{ij}^2 = \|u_i - u_j\|^2$, and H is the centring matrix with $H_{ij} = \delta_{ij} - 1/N$, where δ_{ij} denotes the Kronecker delta. The matrix B in eq. (5.4) is called the centred inner product matrix. If $\tilde{B}$ is the matrix of inner products of the embedded data points, i.e. $\tilde{B}_{ij} = u_i \cdot u_j$ with Euclidean scalar product, then B can be obtained by removing the mean of all rows and columns of $\tilde{B}$, cf. chapter 10.3 of Fouss *et al.* [86]. An embedding of the data points using the eigenvectors corresponding to the leading non-negative eigenvalues of B in eq. (5.4) ensures to capture the main variance of the (squared) distance structure, similar to a principal component analysis.

We compute Δ^2 with the Euclidean embedding described in section 5.2 and restrict ourselves to the first two dimensions to visualize the data structure in the plane, i.e. the embedding is defined by

$$u_i = (w_{0,i}, w_{1,i}), \quad i = 1, \dots, N, \quad (5.5)$$

where $Bw_j = \lambda_j w_j$, and $\lambda_0 \geq \lambda_1 \geq \lambda_k$ for all $k = 2, \dots, N-1$. This choice of embedding ensures to capture the main variance of the data points, and we therefore also expect to capture the main structure in terms of data density. For large particle sets however, computing the spectrum of B in eq. (5.4) is computationally not feasible, as the matrix B is dense and computing the spectrum scales with $O(N^3)$. We apply classical MDS to the 12,000 particles of the Bickley jet model flow, and a random selection of the equal number of particles for the Agulhas flow. In our context, the method is most useful for visualization purposes, as it provides a good 2-dimensional approximation of the point distances, i.e. also the density structure of the embedded trajectories.

5.3. Clustering with OPTICS

The detection of dense accumulations of points that are separated from each other by non-dense regions (noise) is the main goal of density-based clustering. We use the OPTICS (Ordering Points To Identify the Clustering Structure) algorithm by Ankerst *et al.* [38] to detect these regions. The OPTICS algorithm can be seen as an extension of DBSCAN [103]. As we have no prior information on the density structure of the embedded nodes, we set the ‘generating distance’ of OPTICS to infinity and our presentation here is limited to this case. The general OPTICS algorithm with finite generating distance is computationally more efficient and slightly more complicated, and we refer to [38] for more details. For $\delta \in \mathbb{R}$, the δ -neighbourhood of a point $p \in \mathbb{R}^M$ is defined as the M -dimensional ball of radius δ around p . Define $M_\delta(p)$ as the number of points that is in the δ -neighbourhood of p , including p itself. OPTICS requires one parameter, an integer s_{min} (called MinPts by Ankerst *et al.* [38]), that defines the *core-distance* of a point p as

$$c(p) = \{\min(\delta) \mid M_\delta(p) \geq s_{min}\}. \quad (5.6)$$

The core distance is simply the minimum radius of a ball around p , such that the ball contains s_{min} points. Note that the generating distance that we set to infinity is a maximum cut off distance for the computation of the core distance in eq. (5.6), beyond which the core distance is not defined. As we do not have an intuition for a good value of such a cut off, we remove it by setting it to infinity.

The ordering of the points is based on the *reachability distance* of a point p w.r.t. another point q , defined as

$$r(p|q) = \max(c(q), \|p - q\|), \quad (5.7)$$

where $\|p - q\|$ in our case denotes the Euclidean distance between p and q . The ordering of points is then constructed with the following scheme:

- Step 1 Pick a point p_1 . This is the first point in the order, and is arbitrary.
- Step 2 Compute the core-distance $c(p_1)$ of p_1 .
- Step 3 Define an ordered seed list containing all other points, p_l , $l = 2, \dots, N$. For each point p_l , define the reachability value $r(p_l)$ as the reachability distance (eq. (5.7)) w.r.t. p_1 , $r(p_l) = r(p_l|p_1)$. Order the list in ascending order of the $r(p_l)$.
- Step 4 Pick the first point on the ordered seed list as p_2 and compute the core-distance $c(p_2)$. For all remaining points p_l , $l = 3, \dots, N$, update the reachability value $r(p_l) \rightarrow \min(r(p_l), r(p_l|p_2))$.
- Step 5 Update the ordered seed list according to the new reachability.
- Step 6 Repeat steps 4-5 to obtain p_3 . Continue until all points are processed.

Note that the ordering of points is achieved by constantly updating the ordered seed list, cf. step 3. In this way, the algorithm iterates through groups of dense points one after the other, and only continues with other points once a dense region has been fully explored. Note also that the entire algorithm depends on the choice of the parameter s_{min} . The value of s_{min} should be chosen roughly as a minimum value of the expected cluster size. In the examples presented in this paper, we take values for s_{min} that correspond to the estimated minimum size of the coherent sets.

The main result of the OPTICS algorithm is a *reachability plot*. This plot is the graph defined by $(i, r(p_i))$, where $r(p_1) = \infty$ by definition. The reachability plot is a powerful presentation of the global and local distribution of a set of points at once. The valleys in this plot correspond to dense regions, which we relate to finite-time coherent sets. We show examples of reachability plots in section 5.4. Given the reachability plot $(i, r(p_i))$, we use two common ways to derive a clustering result:

5

1. **DBSCAN clustering:** Choose a cut-off parameter ϵ and define all points p_i with $c(p_i) \leq \epsilon$ as core points. All points that are not in the ϵ -neighbourhood of a core point are defined as noise. This set of noisy data points is equivalent to all points p_i that are not core points and have a reachability value $r(p_i)$ with $r(p_i) > \epsilon$. A cluster of size L is then defined as a consecutive set (in the sense of the ordering) of non-noise points $(p_j, p_{j+1}, \dots, p_{j+L-1})$, with adjacent points p_{j-1} and p_{j+L} being noise. This is similar to the clustering result of a DBSCAN run with equal values for s_{min} and ϵ . All possible realizations of DBSCAN clusters, with the same value for s_{min} , can therefore be derived from the reachability values, core distances and the ordering determined by OPTICS. Up to boundary points, a DBSCAN clustering result can be obtained by drawing horizontal lines in the reachability plot, cf. section 5.4.
2. **ξ -clustering:** While the DBSCAN clustering method looks for deep valleys in the reachability plot, this method looks for valleys with steep boundaries. In short, the larger a parameter ξ with $0 < \xi < 1$, the steeper the boundary of a valley has to be to be classified as a cluster. In more detail, a ξ -cluster is defined as a consecutive set of points $(p_j, p_{j+1}, \dots, p_{j+L-1})$ that has steep boundaries in the sense that for a parameter ξ , $0 < \xi < 1$:
 - (a) The start of the cluster p_j is in a ξ -steep downward area. A ξ -steep downward area is a maximal set of consecutive points $(p_l, p_{l+1}, \dots, p_{l+k})$, $k \in \{1, \dots, N - l\}$ where: 1. p_l and p_{l+k} are ξ -steep downward points, i.e. $r(p_l) \leq (1 - \xi)r(p_{l-1})$ and $r(p_{l+k}) \leq (1 - \xi)r(p_{l+k-1})$, 2. $p_{l+i} \leq p_l$ for all $i = 1, \dots, k$ and 3. not more than s_{min} consecutive points in the set are no ξ -steep downward points.
 - (b) The end of the cluster p_{j+L-1} is a ξ -steep upward area. The definitions are the reverse of the ξ -steep downward area, with the definition of a ξ -steep upward point as $r(p_j) \leq (1 - \xi)r(p_{j+1})$.

- (c) The cluster contains at least s_{min} points, i.e. $L \geq s_{min}$.
- (d) Every point in the inside of the cluster is at least a factor of $(1 - \xi)$ smaller than the boundary points p_j and p_{j+L-1} . All points that do not belong to a cluster are classified as noise.

We refer to Ankerst *et al.* [38] for a more detailed discussion of the ξ -clustering method with illustrations for example data. Note that the full ξ -clustering method presented by [38] does contain some more details related to the choice of the start and end points, which we did not mention here.

The OPTICS algorithm as well as functions to derive both clustering results from an OPTICS output are available in the Python scikit-learn package. Note that the implementation in sklearn allows for a minimum cluster size different from s_{min} for the ξ -clustering method (item 2 c above), but we will not make use of this additional freedom to reduce the number of parameters. Note that, different from k-Means, both clustering methods do not require an a priori determination of the number of clusters. For the ξ -clustering method, a larger ξ requires steeper boundaries to form a cluster, i.e. will typically lead to a reduction of the number of resulting clusters. For DBSCAN clustering with very large ϵ , one will detect one large global cluster. Making ϵ smaller leads then to consecutive splits of this cluster, forming (up to noise) a cluster hierarchy. We will demonstrate the properties for both clustering methods in section 5.4 for different situations. In the following applications, we use an estimation of the minimum number of particles per finite-time coherent set for the parameter s_{min} .

Intuitively, the two clustering methods can be understood as follows. DBSCAN detects those groups of points that have a certain minimum density defined by the minimum reachability distance ϵ . Clusters detected by DBSCAN are therefore defined by a global density criterion. This assumes no structural differences in the type of coherent sets in different regions of the fluid. Different from that, the ξ -clustering method detects clusters by finding strong changes in the density of the data points, and not based on absolute densities. This has the advantage that clusters of different absolute density can be detected. Such a situation can arise if the distribution of particles is inhomogeneous over the fluid domain, or if the spatial extend of the fluid domain is very large such that the properties of finite-time coherent sets vary significantly. It is important to note that the main result of OPTICS is the reachability plot itself. The DBSCAN- and ξ -clustering methods should be seen as useful tools to identify the most important features of that plot.

5.4. Comparison to related methods

Our method is closely related to existing methods to detect finite-time coherent sets with clustering techniques. Most notably, Froyland and Padberg-Gehle [80] also use a direct embedding of individual trajectories similar to eq. (5.3), together with fuzzy-c-means clustering. Hadjighasem *et al.* [35], Banisch and Koltai [82], Padberg-Gehle and Schneide [81] and Froyland and Junge [97] use spectral embeddings of graphs that are defined on some form of physical in-

tuition or of dynamical operators, together with k-Means clustering. These studies show applications of their methods to example flows where the size of almost-coherent sets is not too small compared to the fluid domain. Such examples are the Bickley jet flow, which we also study in section 5.1, the five major ocean basins [80, 82], or few individual eddies in an ocean or atmospheric flow [35, 81, 97]. In such situations, noisy background trajectories can be detected as individual clusters by the partitioning method, as discussed by Hadjighasem *et al.* [35]. For applications in large ocean domains, where the number of eddies is not known beforehand and where there are many more noisy trajectories than coherent trajectories, such an approach is likely to fail, see also the discussion by Froyland *et al.* [37]. OPTICS does not require to fix the number of clusters beforehand, and also contains an intrinsic concept of noisy trajectories that do not belong to any cluster, making OPTICS suitable for challenging flows in large domains.

As mentioned, OPTICS also contains an intrinsic notion of cluster hierarchy, i.e. coherent sets that are themselves part of coherent sets at larger scales. Ma and Bollt [89] studied hierarchical coherent sets in the transfer operator framework of Froyland *et al.* [95], in the spirit of the hierarchical clustering method proposed by Shi and Malik [84]. Their approach is also partition-based, i.e. there is no concept of noisy trajectories. In addition, at each stage of the hierarchy, a fixed cut-off has to be chosen based on minimizing an objective function [89]. Different from that approach, the main result of OPTICS, the reachability plot, contains such hierarchical information in a smooth and intrinsic manner.

As described in section 5.3, clustering results of the DBSCAN algorithm [103] can be derived from the reachability plot of OPTICS. DBSCAN has been used in the context of coherent sets before by Schneide *et al.* [96], although not to identify specific clusters, but to distinguish noisy from clustered trajectories. The potential of density-based clustering for applications in the ocean and its comparison to other existing clustering methods for flow examples such as the Bickley jet (cf. section 5.1) has not been explored so far. Different from OPTICS, DBSCAN detects clusters with a certain fixed minimum density, although clusters with varying densities might be present in a dataset [38]. More specifically, the value for the cut-off parameter ϵ , cf. section 5.3, has to be set beforehand. Choosing a good value for the density parameter in DBSCAN is challenging if there is no underlying physical intuition for the density structure. As described in section 5.3, OPTICS allows one to derive any DBSCAN clustering result, with the same value for the parameter s_{min} , after computing the reachability plot, i.e. after one can get first insights into the clustering structure of the dataset to make an appropriate choice for ϵ . Furthermore, it also allows one to use the ξ -clustering method instead of DBSCAN (cf. section 5.3).

A more recent and powerful technique to detect finite-time coherent sets in sparse trajectory data was presented by Froyland *et al.* [37], based on dynamic Laplacian and transfer operators [97]. Froyland *et al.* [37] apply their method to a trajectory dataset in the Western Boundary Current region in the North Atlantic Ocean, and successfully detect many eddies by superposing individual

eigenvectors. The methods presented there are based on a form of spectral embedding, derived from discretized dynamical operators. Based on this embedding, clustering results have also been derived with k-Means by Froyland and Junge [97] and with individual thresholding by Froyland *et al.* [37]. Froyland *et al.* [37] also show how the low-order eigenvectors correspond to large-scale coherent features, while the individual eddies are derived by a sparse eigenbasis approximation of a number of eigenvectors. The latter approach is essentially a transformation of the embedding to represent the most reliable features, such that a superposition of the eigenvectors alone yields the information about the location and size of finite-time coherent sets (without a clustering step). This is essentially an optimized form of embedding, i.e. the second step in fig. 5.1. Our aim here is to focus on the third step in fig. 5.1, i.e. to demonstrate the potential of the density-based clustering algorithm OPTICS, together with a very simple embedding of eq. (5.3).

A downside of our method compared to other approaches is the rather ad-hoc choice of embedding, cf. eq. (5.3). Different from many other methods, most notably the ones of Banisch and Koltai [82], Froyland and Junge [97] and Froyland *et al.* [37], this type of embedding is not derived from a meaningful dynamical operator. It could be fruitful to explore a combination of these more meaningful embeddings together with OPTICS as a clustering algorithm in future research.

5.4. Results

5.1. Bickley jet flow

We start with the direct embedding of the Bickley jet flow trajectories, cf. section 5.2. The data matrix has dimension $X \in \mathbb{R}^{12,000 \times 123}$. We apply the OPTICS algorithm to the resulting points together with DBSCAN clustering, choosing $s_{min} = 80$ as a minimum size of the finite-time coherent sets. In the following, all axis units are in multiples of 1000 km. Figure 5.2 shows the reachability plot, together with the DBSCAN clustering result of three different choices of ϵ . The six vortices and the jet are clearly visible as the major valleys in the reachability plot. The hierarchical structure of the DBSCAN clustering with decreasing ϵ is visible in the figures from top (large-scale coherence) to bottom (small-scale coherence). Note that for the DBSCAN clustering results, boundary points of the clusters can be above the horizontal line at $y = \epsilon$. This is because of the definition of the DBSCAN clustering in section 5.3.

To illustrate the difference between OPTICS and k-Means, we use the embedded trajectories and apply classical MDS to obtain a 2-dimensional embedding. As described in section 5.2, this assures to capture the major variance along the embedding axes. The spectrum of B in eq. (5.4) is shown in fig. 5.11 in the appendix, with two clearly dominant eigenvalues. The fact that there are two very dominant eigenvalues assures that the illustration of the data in the plane captures the major variance of the data points. Figure 5.3a shows the corresponding embedding of the trajectories in the 2-dimensional Euclidean

Figure 5.2: Result of the OPTICS algorithm applied to the direct embedding of the trajectories. (a), (d) and (g) show the reachability plot with different DBSCAN clustering results, indicated by the black horizontal line. The corresponding clustering results of each choice of DBSCAN parameter ϵ is shown on the right of the reachability plots for different times. Grey particles correspond to noise. Axis units in the centre and right column are in 1000 km.

space. The star-shaped distribution of data points reflect the strong symmetries of the underlying idealized Bickley jet flow. Such symmetry is not expected to be present for more realistic flows. Figures 5.3b and 5.3c show the cluster labels for OPTICS with DBSCAN clustering at $\epsilon = 10^6$ km, and for a k-Means clustering with $K = 8$ clusters, respectively. $K = 8$ corresponds to the six vortices, the jet, and one noise cluster as suggested by [35].

The corresponding clustering results in real space are shown in figs. 5.4 and 5.5 for OPTICS and k-Means, respectively. The jet and the six vortices are clearly recognizable as dense accumulations of points in the 2-dimensional space of fig. 5.3b, see fig. 5.4 for the corresponding colours. The clustering result with k-Means in fig. 5.5 shows that the clusters corresponding to the vortices are much less focussed. In addition, each of the eight clusters in fig. 5.3c contains some of the noisy points of fig. 5.3b, which shows that using one additional cluster for noise does not work in this situation. It is interesting to note that capturing the noisy data points of fig. 5.3b by an additional cluster in k-Means is geometrically impossible, simply because k-Means clusters are circular. Covering all noisy points without including the centre, i.e. the jet in fig. 5.3b, is not possible for k-Means.

It should be noted here that the poor performance of k-Means in figs. 5.3c and 5.5 is not representative for other methods that use k-Means. For example, the method of Banisch and Koltai [82] captures the coherent structures in the Bickley jet rather well, including the jet in the middle. We emphasize again that we

Figure 5.3: a: 2-dimensional embedding of the classical MDS method (cf. section 5.2) of the trajectories. b: with labels according to the DBSCAN result of fig. 5.4. The six vortices and the jet are clearly visible as dense regions. Grey particles correspond to noise. c: k-Means clustering result for $K=8$, see fig. 5.5 for the spatial clustering result of k-Means.

use classical MDS here mostly for visualization purposes, as the computation of the classical MDS embedding is difficult for large particle sets. In our case, a dense $12,000 \times 12,000$ symmetric matrix has to be diagonalized, which already takes a significant amount of computation time.

5

Figure 5.4: Result of DBSCAN clustering of the 2-dimensional embedding of the classical MDS method. a: reachability plot with black line representing the DBSCAN parameter ϵ . b-c: corresponding clustering results at different times. Grey particles represent noise. Axis units are in 1000 km.

Figure 5.5: Result of $K = 8$ k-Means clustering of the 2-dimensional embedding from classical MDS, cf. fig. 5.4. Axis units are in 1000 km.

We finally also tested the performance of our algorithm with a random subset of 2,000 particles, using data for every five days instead of every day, cf. fig. 5.12 in the appendix. OPTICS still detects the six vortices and the jet, although the cluster boundaries are less clearly defined compared to fig. 5.2. Froyland and Junge [97] detect the vortices and the jet by using data of 3,000 particles only at initial and final times ($t = 0$ and $t = 40$ days). Our method is not able

to detect the expected finite-time coherent sets with using only initial and final particle data. This is likely to be a result of the ad-hoc direct embedding, cf. eq. (5.3), see the discussion at the end of section 5.4.

5.2. Agulhas rings

We next apply OPTICS to the Agulhas trajectories. As described in section 5.2, we have $\tilde{X} \in \mathbb{R}^{N \times 63}$ with $N = 23,821$. We choose $s_{min} = 100$ in the following, which corresponds initially to a square cell of $2^\circ \times 2^\circ$, i.e. a reasonable minimum size of an Agulhas ring. Figure 5.6 shows the result of the direct embedding. The reachability plot in fig. 5.6a is much more jagged than for the Bickley jet model flow (cf. fig. 5.2a). The narrow deep valleys and the wider valleys in the reachability plot indicate the presence of large- and small-scale coherence patterns. Figure 5.6a-c show the DBSCAN clustering result for a relatively large value of ϵ . The main separation of fluid domains is between the red and the blue particles, with a few vortices at their boundary. These two water masses are the northern and southern parts of the subtropical gyre in the South Atlantic, the red particles moving to the west, the blue particles to the east. The second and third rows of fig. 5.6 show other clustering results for the DBSCAN- and the ξ -clustering method, respectively. The valleys in fig. 5.6g with steepest boundaries as detected by the ξ -clustering method mostly correspond to eddy-like structures, separated by background noise. Note that not all clusters in the figure correspond to eddies. For example, the blue cluster in figs. 5.6g-i stays approximately coherent over the considered time interval, although it is certainly not an Agulhas ring. An animation of the detected finite-time coherent sets for the full two years of trajectory data based on the ξ -clustering method as in the last row of fig. 5.6 can be found on Zenodo [104], showing that many of the sets stay coherent for significantly longer times than the first 100 days.

Figure 5.6 shows that for this situation, the ξ -clustering method detects more Agulhas rings than DBSCAN. While the clustering results shown in the figure all depends on the parameter values for ξ and ϵ , it is visible in the reachability plot of fig. 5.6g that the definition of some eddies includes the entire boundary of the valleys, i.e. up to very high reachability values. At the same time, the detection of the large-scale clusters as in 5.6a-c is not possible with the ξ -clustering method. These findings are in fact expected, cf. the discussion of the two clustering methods at the end of section 5.3. DBSCAN is best to detect global density structures, i.e. when the reachability values of all points are compared to the same cut-off ϵ . Regions that are dense locally but not necessarily globally are better detected with the ξ -clustering method. Despite these differences between the two clustering methods, we again emphasize that the main result of OPTICS is the reachability plot itself. Fig. 5.7 shows a colour map at initial time of the reachability values. We clearly see Agulhas rings as the dark regions corresponding to lowest values of reachability. The regions of large reachability correspond to trajectories that are relatively noisy compared to all the other trajectories.

In order to illustrate again the difference between OPTICS and k-Means for

Figure 5.6: Result of the OPTICS algorithm applied to the direct embedding of the trajectories, with different clustering methods. Grey particles correspond to noise.

5

this example, we choose 12,000 random trajectories and again embed the trajectories in a 2-dimensional space with classical MDS (cf. section 5.2). The reduction of the particle set is necessary to simplify the eigendecomposition of the matrix B in eq. (5.4), and we therefore choose $s_{min} = 30$. The corresponding spectrum of B is shown in fig. 5.13 in the appendix, showing that there are again two dominant eigenvectors, i.e. visualizing the network in the plane captures the main variance of the data. Figure 5.8 shows the embedded trajectories together with OPTICS / DBSCAN clustering (fig. 5.8b) and k-Means (fig. 5.8c) for $K=40$. Figs. 5.9 and 5.10 show the corresponding clustering results in the fluid domain. It is visible that k-Means does not detect a single vortex, but splits the fluid domain into regions of approximately similar size. OPTICS detects multiple Agulhas rings by finding the deepest valleys in the reachability plot.

It is interesting to note that the use of classical MDS in fig. 5.9 has lead to the detection of many of the vortices of fig. 5.6d-f with DBSCAN instead of the ξ -clustering method. The transformation to the reduced 2D space has hence lead to a simplification of the reachability plot, which now represents the major variations in the distances of the embedded trajectories. At the same time, the large-scale structure of 5.6a is not visible any more in fig. 5.9. This indicates that exploring more dimensionality reduction techniques could be useful for future research, in particular those that are computationally more efficient than classical MDS.

Spectral embeddings derived from networks together with partition-based clustering have a similar problem as the one illustrated in figs. 5.8c and 5.10 [37].

Figure 5.7: Reachability values at initial time, resulting from the OPTICS algorithm applied to the direct embedding of the trajectories. The regions with lowest values clearly correspond to Agulhas rings. The colour bar is cut off at a reachability of 1000 km to show the relevant structure of variations.

5

Figure 5.8: Embedding of the Agulhas trajectories in the 2-dimensional space defined by the leading eigenvectors of the MDS Kernel matrix B . a: no labels. b: clustering labels of OPTICS / DBSCAN, see fig. 5.9 for the corresponding plot in the Agulhas region. Grey particles represent noise. c: k-Means with $K = 40$, see fig. 5.10 for the corresponding plot in the Agulhas domain.

Similar to the case discussed here, OPTICS can be used to overcome the problems of k-Means. We show this in appendix 5.6.C for the network proposed by [81] for the Agulhas region, together with a brief introduction of the network and how to construct spectral embeddings. In summary, k-Means again fails to detect any of the vortices, while OPTICS detects many of the coherent vortices in the spectrally embedded network. Yet, other flow features are also present that result from the physical motivation of the network definition, see the results in appendix 5.6.C.

5.5. Conclusion

The abstract embedding of particle trajectories in a metric space with subsequent clustering is a promising field of research for the detection of finite-time coherent sets in oceanography. Yet, most of the existing methods have been based on graph partitioning, which has no concept of noisy, unclustered tra-

Figure 5.9: Result of OPTICS applied to the 2-dimensional embedding of 12,000 randomly selected particles with the classical MDS method, cf. fig. 5.8b, and $s_{min} = 30$. The corresponding spectrum is shown in fig. 5.13 in the appendix, showing that there are two dominant eigenvectors. Grey particles are classified as noise.

Figure 5.10: Result of the k-Means clustering with $K = 40$ applied to the 2-dimensional embedding with classical MDS, cf. fig. 5.8c.

jectories. This is a problem for applications in the ocean, where many eddies are transported in a noisy background flow on large domains. This study is motivated by the success of Froyland *et al.* [37] in overcoming the problem of graph partitioning by a sophisticated form of trajectory embedding. Here, we show how the density-based clustering algorithm OPTICS [38] can be used instead of graph partitioning, in order to detect small-scale eddies in large ocean domains. Different from partition-based clustering methods such as k-Means, OPTICS does not require to fix the number of clusters beforehand. Clusters are detected by identifying dense accumulations of points, i.e. groups of trajectories that are close to each other in embedding space. Coherent groups of particle trajectories can be identified as valleys in the reachability plot computed by the OPTICS algorithm. This plot also has a natural interpretation in terms of cluster hierarchies, i.e. finite-time coherent sets that are by themselves part of a larger scale finite-time coherent set. Such hierarchies are present in the surface ocean flow, where the subtropical basins are approximately coherent and at the same time contain other finite-time coherent structures such as eddies and jets.

We apply OPTICS to Lagrangian particle trajectories directly, in the spirit of Froyland and Padberg-Gehle [80]. OPTICS successfully detects the expected coherent structures in the Bickley jet model flow, separating the six vortices and the jet from background noise. We also apply OPTICS to simulated trajectories in the eastern South Atlantic and successfully identify Agulhas rings, separated by noise. We visualize the difference between OPTICS and k-Means with a 2-dimensional embedding of the trajectories based on classical multidimensional scaling. We also show how OPTICS can be applied to the spectral embedding

of the particle-based network proposed by Padberg-Gehle and Schneide [81], providing a necessary amendment to their method to detect coherent vortices in a large ocean domain, i.e. when k-Means fails. Our method is very simple to implement in Python, as OPTICS is available in the Python scikit-learn package. While we here present the results of OPTICS with three different kinds of embeddings, it is likely that OPTICS also works for other trajectory embeddings, such as the spectral embeddings of Banisch and Koltai [82] or Froyland and Junge [97]. Using such dynamically motivated embeddings instead of the ad-hoc direct embedding presented here could be a promising direction for future research.

Extending our method to datasets with more trajectories can be made more efficient by choosing a finite generating distance for OPTICS [38]. While this is better from a computational point of view, it requires some knowledge or intuition about the spatial distribution of the embedded trajectories. A major challenge for the method proposed here is the embedding dimension. For long trajectories, it is necessary to reduce the dimensionality of the trajectories before applying OPTICS. A complication here is the desired property of an embedding to preserve both local and global distances in order to make full use of the hierarchical properties of OPTICS. This means, for example, that the popular method of a locally linear embedding [105] is not suitable, unless only the small-scale (densest finite-time coherent sets) are to be detected. Using classical multidimensional scaling (MDS), as we did here to visualize the clustering results, in principle preserves local and global distances, although our results indicate that the large-scale coherence structure in the Agulhas flow is less pronounced for the classical MDS embedding compared to the full embedding of trajectories. In any case, classical MDS is not an option for very large datasets, as it requires the diagonalization of a dense symmetric square matrix of size equal to the particle number. Spectral embeddings of derived networks such as the ones of Hadjighasem *et al.* [35], Padberg-Gehle and Schneide [81] and Banisch and Koltai [82] are useful to achieve lower-dimensional embeddings, but they come with the introduction of additional parameters for the network construction and heuristics to truncate the embedding dimension. Further research into other non-linear dimensionality reduction techniques that have not been explored in the context of finite-time coherent sets can lead to more efficient and robust methods.

5.6. Appendices

5.6.A. Additional figures for the Bickley jet flow

Figure 5.11: Spectrum of the classical MDS kernel matrix B for the Bickley jet flow. It is visible that there are two dominant eigenvalues. We choose the vectors corresponding to these first two eigenvalues as embedding vectors in section 5.1.

5

Figure 5.12: Result of the OPTICS algorithm for a random subset of 2,000 particles in the Bickley jet flow, with particle data every 5 days instead of every day. To account for the smaller number of particles, we set $s_{min} = 15$ for this case. The six vortices and the jet are still clearly visible.

5.6.B. Additional figures for the Agulhas flow

Figure 5.13: Spectrum of the classical MDS kernel matrix B for the Agulhas flow, where we first constrain the particle data to 12,000 randomly selected trajectories. There are again two dominant eigenvalues, for which we choose the corresponding vectors for the embedding in section 5.2.

5.6.C. Detecting Agulhas rings with a particle-based network

To demonstrate that OPTICS can also be applied to the spectral embedding of a particle-based network, we use the network proposed by Padberg-Gehle and Schneide [81]. If we have a set of particle trajectories $x_i(t)$, where $i = 1, \dots, N$, $t = t_1, t_2, \dots, t_T$ with N the number of particles and T the number of time steps, the network $A \in \mathbb{R}^{N \times N}$ is defined as:

$$A_{ij} = \begin{cases} 1, & \text{if } \exists t \in \{t_1, t_2, \dots, t_T\} \text{ s.t. } \|x_i(t) - x_j(t)\| < d, \\ 0, & \text{otherwise.} \end{cases} \quad (5.8)$$

Here, $\|\cdot\|$ denotes the Euclidean norm and $d > 0$ is a fixed pre-determined cut-off parameter, see Padberg-Gehle and Schneide [81] for a discussion on the choice of d (called ϵ in Padberg-Gehle and Schneide [81]). Similar to Padberg-Gehle and Schneide [81], we embed the nodes in a lower dimensional space $\mathbb{R}^K$ by means of the eigenvectors of its random walk Laplacian, (see e.g. Von Luxburg [87])

$$L_r = D^{-1}A, \quad (5.9)$$

where D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$. The embedding of node i is defined by

$$y_i = (v_{1,i}, v_{1,i}, \dots, v_{K,i}) \in \mathbb{R}^K, \quad (5.10)$$

where v_i , $i = 0, \dots, N-1$ are the right eigenvectors corresponding to the largest eigenvalues λ_i of L_r . The eigenvalues are assumed to be ordered in descending order, i.e. $1 = \lambda_0 > \lambda_1 \geq \dots \geq \lambda_N$. The classical simultaneous K-way normalized cut proceeds with applying the k-Means algorithm to the embedding defined in eq. (5.10) to detect K clusters [87], resulting in an approximate solution to the normalized cut problem [84].

Figure 5.14 shows the spectrum of the resulting random walk Laplacian with $d = 200$ km. No obvious spectral gap is visible that would suggest a truncation of the embedding space. Figure 5.15 shows the clustering result if we apply a k-Means algorithm as suggested by Padberg-Gehle and Schneide [81] to detect $K = 40$ clusters. It is visible that the partition-based k-Means clustering method does not detect any individual Agulhas rings, but partitions the state space into regions of approximately equal size.

Applying OPTICS instead of k-Means with a subsequent ξ -clustering detects some of the Agulhas rings, see fig. 5.16, where we choose $s_{min} = 100$ as in section 5.2. Note that also other structures than typical circular eddies are detected. While this depends on the clustering parameter ξ (or ϵ for DBSCAN), this is also a consequence of the *physically motivated* network defined by eq. (5.10), where particles are connected equally if they are close to each other at least once in time. This is different from the direct embedding, where we require particles to stay close to each other along the entire trajectory.

Figure 5.14: Spectrum of the random walk Laplacian, cf. eq. (5.9) of the network proposed by Padberg-Gehle and Schneide [81] applied to the Agulhas trajectory data. No clear gap exists that suggest a truncation of the embedding.

5

Figure 5.15: Result of k-Means clustering applied to the 40 leading eigenvectors of the random walk Laplacian, cf. eq. (5.9), looking for 40 clusters. No individual vortices are detected.

Figure 5.16: Result of OPTICS applied to the $K = 40$ spectral embedding of the network defined in eq. (5.8) with $d = 200$ km and $s_{min} = 100$. Grey particles are classified as noise.

6

Conclusion and outlook

This thesis explored methods to understand the large-scale structure of the ocean flow with Lagrangian trajectories. With regards to the three main goals of this thesis (see section 1.5), chapter 2 addressed goal 1 by including some of the known behaviour of microplastic in the vertical direction, albeit in a very idealized form. Goal 2 was achieved by the results of chapter 3, where the limits of surface plastic modelling were successfully quantified and shown to be relevant for existing models. Chapters 4 and 5 addressed goal 3, providing two methods for the data-driven analysis of large amounts of Lagrangian particles for ocean drifters and simulated trajectories.

6.1. Research questions and answers

How does our understanding of global plastic transport change if we take into account near-surface currents?

In chapter 2, we computed transport rates between large-scale oceanographic regions for different models of particle transport. The long term distribution of particles constrained to the ocean surface successfully reproduced previous results of the accumulation of microplastic in the Subtropical Gyres. Constraining particles not to the surface but to different fixed depths up to 120 m in the water column leads to a gradual decay of this ‘garbage patch’ structure, where the accumulation regions disappear at a depth of around 60 m. This is a consequence of the decay of the Ekman velocity with depth. The results for the global long-term distribution for particles being randomly displaced in the upper water column, and for particles that follow the full three-dimensional currents are similar. The most important result of chapter 2 is that the transport between subtropical and polar regions becomes more likely in deeper layers compared to the surface. As buoyant plastic is known to lose its buoyancy due to biofouling, this could provide a possible explanation for the occurrence of microplastic in subpolar and polar regions. Moving slightly away from the

ocean surface does change our understanding of the global transport of buoyant microplastic, although it is important to develop more accurate transport models that reflect the actual biofouling behaviour of particles in the ocean in the future.

What is the mixing time of the surface ocean flow, and how does that affect our current understanding of plastic transport?

In chapter 3, we show that the mixing time scale for large parts of the ocean is at the order of or below 10 years, which is a typical simulation time chosen in studies on the accumulation of surface microplastic. Our method is based on assuming imperfect knowledge of the initial plastic distribution in space, reflected by the choice of binning for the mixing entropy and Markov chain mixing time. Our results have two important implications in the plastic context. First, it is not necessary to develop complicated models for plastic release if the integration time is on the order of 10 years. The information of any initial distribution will be lost after such a long time of particle advection. Secondly, as the correlation between an initial particle distribution and its final distribution decays over time, there is a limit to the backtracking of Lagrangian particles, i.e. for determining where particles came from.

6

Can we identify the major large-scale features in the North Atlantic Ocean based on drifter data with network theory?

We showed in chapter 4 that a network based on the symbolic itineraries of individual particles together with spectral clustering can indeed detect the most popular features of the large-scale surface flow in the Atlantic Ocean, such as the Western Boundary Current region, the Caribbean Sea, as well as the approximate transport barrier between the Subpolar and Subtropical Gyres. Our method presented in chapter 4 is designed for the application to incomplete trajectory data, and maximises the possible use of data by neglecting all time-dependence of the drifter trajectories.

Can we accommodate the concept of hierarchy and noise in detecting coherent structures in Lagrangian particle data?

The OPTICS clustering technique of Ankerst *et al.* [38] used in chapter 5 accommodates the concepts of hierarchy and noisy data points. In the chapter, we show that OPTICS can be used to overcome problems of partition based clustering. OPTICS is a powerful tool for the detection of finite-time coherent sets, i.e. eddies and jets, on challenging ocean domains. As the clustering method is fully implemented in the scikit-learn library of Python, the method is very easy to apply to Lagrangian data, and naturally comes with a hierarchical interpretation of coherence.

6.2. Future perspectives

There are several concrete potential topics for future research related to the results presented in this thesis. In the context of surface mixing, cf. chapter 3,

it could be useful to study how the global mixing structure changes with more realistic particle advection models, e.g. by including more realistic forms of biofouling, such as the one proposed in [2]. It is likely that the mixing of particles that are not restricted to the strongly converging surface currents is less strong.

In the context of data clustering and the identification of structure in drifter data, further research into suitable dimensionality reduction techniques, such as the locally linear embedding [105], could yield important information on trajectory patterns. More specifically, one could test the combination of the spectral embedding of Froyland and Junge [97] with OPTICS as a clustering algorithm. This is a straightforward extension of chapter 5, with large potential in terms of performance and physical interpretability of the resulting coherence structure.

Finally, an extension of the density-based clustering technique of chapter 5 to the three-dimensional ocean flow could yield valuable insights into the transport patterns of the deep ocean. This could also provide useful insights into the performance of an ocean model for reproducing the global thermohaline circulation. In this context, particularly the ξ -clustering technique of the OPTICS algorithm of chapter 5 could be useful for the three-dimensional ocean flow, as it provides a detection of clusters based on a local density criterion. Such a local criterion is necessary as the time scales of the ocean flow varies significantly with depth. To implement this in practice, it is necessary to first significantly reduce the dimensionality of the Lagrangian particle data set with dimensionality reduction techniques.

The simulation of virtual particles in geophysical flows is about to reach a considerable scale, producing very large amounts of data with large computing facilities and powerful numerical implementations of the Lagrangian dynamics. Yet, in the overall picture, the analysis of such large Lagrangian data sets seems to lag behind the current state-of-the-art data analysis techniques. Existing studies in the field have only explored a very limited range of available data reduction and clustering algorithms. This is very well reflected by the fact that most of the published algorithms employing data clustering techniques, cf. chapters 4 and 5, still rely on k-means clustering. The research presented in this thesis shows the large potential that existing methods from data science can have for the analysis of Lagrangian particle data. In chapter 5, we have shown how an established algorithm developed more than 20 years ago outperforms many recent algorithms in the detection of ocean eddies on large ocean domains, even without an abstract pre-processing of the data. It is likely that other data reduction techniques exist that perform even better, potentially yielding valuable insights into the dynamics and the emergent behaviour of large sets of Lagrangian particles. Future advances in the field of Lagrangian data analysis will have to move forward by employing more sophisticated data reduction and analysis techniques.

Bibliography

- [1] J. Scutt Phillips, A. Sen Gupta, I. Senina, E. van Sebille, M. Lange, P. Lehodey, J. Hampton, and S. Nicol, *An individual-based model of skip-jack tuna (*katsuwonus pelamis*) movement in the tropical pacific ocean*, Progress in Oceanography **164**, 63 (2018).
- [2] M. Kooi, E. H. Van Nes, M. Scheffer, and A. A. Koelmans, *Ups and Downs in the Ocean: Effects of Biofouling on Vertical Transport of Microplastics*, Environmental Science and Technology **51**, 7963 (2017).
- [3] J. D. Everett, E. van Sebille, M. D. Taylor, I. M. Suthers, C. Setio, P. Cetina-Heredia, and J. A. Smith, *Dispersal of eastern king prawn larvae in a western boundary current: New insights from particle tracking*, Fisheries Oceanography **26**, 513 (2017).
- [4] E. van Sebille, M. H. England, and G. Froyland, *Origin, dynamics and evolution of ocean garbage patches from observed surface drifters*, Environmental Research Letters **7**, 044040 (2012).
- [5] H. A. Dijkstra, *Dynamical Oceanography* (Springer-Verlag Berlin Heidelberg, 2008).
- [6] D. Castellana, S. Baars, F. W. Wubs, and H. A. Dijkstra, *Transition probabilities of noise-induced transitions of the atlantic ocean circulation*, Scientific reports **9**, 20284 (2019).
- [7] E. van Sebille, S. Aliani, K. L. Law, N. Maximenko, J. M. Alsina, A. Bagaev, M. Bergmann, B. Chapron, I. Chubarenko, A. Czar, 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. Khatmulina, A. A. Koelmans, T. Kukulka, C. Laufktter, L. Lebreton, D. Lobelle, C. Maes, V. Martinez-Vicente, M. A. Morales Maqueda, M. Poulain-Zarcos, E. Rodrguez, 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*, Environmental Research Letters **15**, 023003 (2020).
- [8] B. F. Jnsson and J. R. Watson, *The timescales of global surface-ocean connectivity*, Nature Communications **7**, 11239 (2016).
- [9] C. I. Fraser, A. K. Morrison, A. M. C. Hogg, E. C. Macaya, E. van Sebille, P. G. Ryan, A. Padovan, C. Jack, N. Valdivia, and J. M. Waters, *Antarctica's ecological isolation will be broken by storm-driven dispersal and warming*, Nature Climate Change **8**, 704 (2018).

- [10] A. Bower, S. Lozier, A. Biastoch, K. Drouin, N. Foukal, H. Furey, M. Lankhorst, S. Rhs, and S. Zou, *Lagrangian views of the pathways of the atlantic meridional overturning circulation*, Journal of Geophysical Research: Oceans **124**, 5313 (2019).
- [11] R. Lumpkin and L. Centurioni, *Global Drifter Program quality-controlled 6-hour interpolated data from ocean surface drifting buoys*, (2019), accessed: March 18, 2020.
- [12] L. C. Laurindo, A. J. Mariano, and R. Lumpkin, *An improved near-surface velocity climatology for the global ocean from drifter observations*, Deep Sea Research Part I: Oceanographic Research Papers **124**, 73 (2017).
- [13] J. LaCasce, *Statistics from lagrangian observations*, Progress in Oceanography **77**, 1 (2008).
- [14] N. Maximenko, J. Hafner, and P. Niiler, *Pathways of marine debris derived from trajectories of lagrangian drifters*, Marine Pollution Bulletin **65**, 51 (2012).
- [15] M. Kubota, *A Mechanism for the Accumulation of Floating Marine Debris North of Hawaii*, Journal of Physical Oceanography **24**, 1059 (1994).
- [16] L.-M. Lebreton, S. Greer, and J. Borrero, *Numerical modelling of floating debris in the world's oceans*, Marine Pollution Bulletin **64**, 653 (2012).
- [17] M. Eriksen, L. C. Lebreton, H. S. Carson, M. Thiel, C. J. Moore, J. C. Borrero, F. Galgani, P. G. Ryan, and J. Reisser, *Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea*, PLoS ONE **9**, e111913 (2014).
- [18] B. D. Hardesty, J. Harari, A. Isobe, L. Lebreton, N. Maximenko, J. Potemra, E. van Sebille, A. D. Vethaak, and C. Wilcox, *Using numerical model simulations to improve the understanding of micro-plastic distribution and pathways in the marine environment*, Frontiers in Marine Science **4**, 30 (2017).
- [19] G. A. Viglione and A. F. Thompson, *Lagrangian pathways of upwelling in the southern ocean*, Journal of Geophysical Research: Oceans **121**, 6295 (2016).
- [20] P. D. Nooteboom, P. K. Bijl, E. van Sebille, A. S. von der Heydt, and H. A. Dijkstra, *Transport bias by ocean currents in sedimentary microplankton assemblages: Implications for paleoceanographic reconstructions*, Paleoceanography and Paleoclimatology **34**, 1178 (2019).
- [21] M. R. Maxey and J. J. Riley, *Equation of motion for a small rigid sphere in a nonuniform flow*, The Physics of Fluids **26**, 883 (1983).

- [22] A. Lasota and M. C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*, 2nd ed. (Springer-Verlag New York, 1998).
- [23] E. van Sebille, S. M. Griffies, R. Abernathey, T. P. Adams, P. Berloff, A. Bistoch, 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. Rührs, S. H. Shah, M. D. Thomas, J. Wang, P. J. Wolfram, L. Zanna, and J. D. Zika, *Lagrangian ocean analysis: Fundamentals and practices*, Ocean Modelling **121**, 49 (2018).
- [24] G. Madec, *NEMO ocean engine*, Note du Pôle de modélisation **27** (2008).
- [25] E. van Sebille, C. Wilcox, L. Lebreton, N. Maximenko, B. D. Hardesty, J. A. Van Franeker, M. Eriksen, D. Siegel, F. Galgani, and K. L. Law, *A global inventory of small floating plastic debris*, Environmental Research Letters **10**, 124006 (2015).
- [26] P. Cvitanović, R. Artuso, R. Mainieri, G. Tanner, and G. Vattay, *Chaos: Classical and Quantum* (Niels Bohr Inst., Copenhagen, 2016).
- [27] G. Froyland, *Statistically optimal almost-invariant sets*, Physica D: Nonlinear Phenomena **200**, 205 (2005).
- [28] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times* (American Mathematical Society, 2008).
- [29] B. Cushman-Roisin and J.-M. Beckers, *Introduction to Geophysical Fluid Dynamics* (Academic press, 2011).
- [30] G. Froyland and K. Padberg, *Almost-invariant sets and invariant manifolds - Connecting probabilistic and geometric descriptions of coherent structures in flows*, Physica D: Nonlinear Phenomena **238**, 1507 (2009).
- [31] E. Brambilla and L. D. Talley, *Surface drifter exchange between the north atlantic subtropical and subpolar gyres*, Journal of Geophysical Research: Oceans **111**, C07026 (2006).
- [32] G. Haller and F. J. Beron-Vera, *Coherent Lagrangian vortices: The black holes of turbulence*, Journal of Fluid Mechanics **731**, R4 (2013).
- [33] F. J. Beron-Vera, Y. Wang, M. J. Olascoaga, G. J. Goni, and G. Haller, *Objective Detection of Oceanic Eddies and the Agulhas Leakage*, Journal of Physical Oceanography **43**, 1426 (2013).
- [34] G. Froyland, C. Horenkamp, V. Rossi, and E. van Sebille, *Studying an Agulhas ring's long-term pathway and decay with finite-time coherent sets*, Chaos: An Interdisciplinary Journal of Nonlinear Science **25**, 083119 (2015).

- [35] A. Hadjighasem, D. Karrasch, H. Teramoto, and G. Haller, *Spectral-clustering approach to lagrangian vortex detection*, Phys. Rev. E **93**, 063107 (2016).
- [36] 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 **130**, 15 (2018).
- [37] G. Froyland, C. P. Rock, and K. Sakellariou, *Sparse eigenbasis approximation: Multiple feature extraction across spatiotemporal scales with application to coherent set identification*, Communications in Nonlinear Science and Numerical Simulation **77**, 81 (2019).
- [38] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, *OPTICS: Ordering Points to Identify the Clustering Structure*, ACM Sigmod record **28**, 49–60 (1999).
- [39] A. Cózar, F. Echevarría, J. I. González-Gordillo, X. Irigoien, B. Úbeda, S. Hernández-León, Á. T. Palma, S. Navarro, J. García-de Lomas, A. Ruiz, M. L. Fernández-de Puellas, and C. M. Duarte, *Plastic debris in the open ocean*, Proceedings of the National Academy of Sciences **111**, 10239 (2014).
- [40] L. C. Woodall, A. Sanchez-Vidal, M. Canals, G. L. Paterson, R. Coppock, V. Sleight, A. Calafat, A. D. Rogers, B. E. Narayanaswamy, and R. C. Thompson, *The deep sea is a major sink for microplastic debris*, Royal Society Open Science **1**, 140317 (2014).
- [41] R. W. Obbard, S. Sadri, Y. Q. Wong, A. A. Khitun, I. Baker, and R. C. Thompson, *Global warming releases microplastic legacy frozen in arctic sea ice*, Earth's Future **2**, 315 (2014).
- [42] I. Peeken, S. Primpke, B. Beyer, J. Gütermann, C. Katlein, T. Krumpfen, M. Bergmann, L. Hehemann, and G. Gerdt, *Arctic sea ice is an important temporal sink and means of transport for microplastic*, Nature Communications **9**, 1505 (2018).
- [43] C. L. Waller, H. J. Griffiths, C. M. Waluda, S. E. Thorpe, I. Loaiza, B. Moreno, C. O. Pachterres, and K. A. Hughes, *Microplastics in the antarctic marine system: An emerging area of research*, Science of The Total Environment **598**, 220 (2017).
- [44] M. Kubota, K. Takayama, and D. Namimoto, *Pleading for the use of biodegradable polymers in favor of marine environments and to avoid an asbestos-like problem for the future*, Applied Microbiology and Biotechnology **67**, 469 (2005).
- [45] G. Froyland, R. M. Stuart, and E. van Sebille, *How well-connected is the surface of the global ocean?* Chaos: An Interdisciplinary Journal of Nonlinear Science **24**, 033126 (2014).

- [46] L. Lebreton, B. Slat, F. Ferrari, B. Sainte-Rose, J. Aitken, R. Marthouse, S. Hajbane, S. Cunsolo, A. Schwarz, A. Levivier, K. Noble, P. Debeljak, H. Maral, R. Schoeneich-Argent, R. Brambini, and J. Reisser, *Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic*, Scientific Reports **8**, 4666 (2018).
- [47] J. R. Jambeck, R. Geyer, C. Wilcox, T. R. Siegler, M. Perryman, A. Andrady, R. Narayan, and K. L. Law, *Plastic waste inputs from land into the ocean*, Science **347**, 768 (2015).
- [48] D. K. A. Barnes, F. Galgani, R. C. Thompson, and M. Barlaz, *Accumulation and fragmentation of plastic debris in global environments*, Philosophical Transactions of the Royal Society B: Biological Sciences **364**, 1985 (2009).
- [49] A. Isobe, K. Uchiyama-Matsumoto, K. Uchida, and T. Tokai, *Microplastics in the southern ocean*, Marine Pollution Bulletin **114**, 623 (2017).
- [50] A. Cincinelli, C. Scopetani, D. Chelazzi, E. Lombardini, T. Martellini, A. Katsoyiannis, M. C. Fossi, and S. Corsolini, *Microplastic in the surface waters of the ross sea (antarctica): Occurrence, distribution and characterization by ftir*, Chemosphere **175**, 391 (2017).
- [51] A. L. Lusher, V. Tirelli, I. O'Connor, and R. Officer, *Microplastics in Arctic polar waters: The first reported values of particles in surface and sub-surface samples*, Scientific Reports **5**, 14947 (2015).
- [52] S. Allen, D. Allen, V. R. Phoenix, G. Le Roux, P. Durántez Jiménez, A. Simonneau, S. Binet, and D. Galop, *Atmospheric transport and deposition of microplastics in a remote mountain catchment*, Nature Geoscience **12**, 339 (2019).
- [53] F. M. Fazey and P. G. Ryan, *Biofouling on buoyant marine plastics: An experimental study into the effect of size on surface longevity*, Environmental Pollution **210**, 354 (2016).
- [54] D. Kaiser, N. Kowalski, and J. J. Waniek, *Effects of biofouling on the sinking behavior of microplastics*, Environmental Research Letters **12**, 124003 (2017).
- [55] S. Ye and A. L. Andrady, *Fouling of floating plastic debris under biscayne bay exposure conditions*, Marine Pollution Bulletin **22**, 608 (1991).
- [56] T. Kukulka, G. Proskurowski, S. Morét-Ferguson, D. W. Meyer, and K. L. Law, *The effect of wind mixing on the vertical distribution of buoyant plastic debris*, Geophysical Research Letters **39**, L07601 (2012).
- [57] T. Kukulka and K. Brunner, *Passive buoyant tracers in the ocean surface boundary layer: 1. influence of equilibrium wind-waves on vertical distributions*, Journal of Geophysical Research: Oceans **120**, 3837 (2015).

- [58] J. Reisser, B. Slat, K. Noble, K. du Plessis, M. Epp, M. Proietti, J. de Sonnevile, T. Becker, and C. Pattiaratchi, *The vertical distribution of buoyant plastics at sea: an observational study in the north atlantic gyre*, Biogeosciences **12**, 1249 (2015).
- [59] G. K. Batchelor, *An Introduction to Fluid Dynamics* (Cambridge University Press, Cambridge, 2012).
- [60] P. G. Ryan, *Does size and buoyancy affect the long-distance transport of floating debris?* Environmental Research Letters **10**, 084019 (2015).
- [61] M. Kooi, J. Reisser, B. Slat, F. F. Ferrari, M. S. Schmid, S. Cunsolo, R. Brambini, K. Noble, L. A. Sirks, T. E. Linders, R. I. Schoeneich-Argent, and A. A. Koelmans, *The effect of particle properties on the depth profile of buoyant plastics in the ocean*, Scientific Reports **6**, 33882 (2016).
- [62] G. Haller and T. Sapsis, *Where do inertial particles go in fluid flows?* Physica D: Nonlinear Phenomena **237**, 573 (2008).
- [63] *Microplastics as contaminants in the marine environment: A review*, Marine Pollution Bulletin **62**, 2588 (2011).
- [64] L. Lebreton, J. Van Der Zwet, J. W. Damsteeg, B. Slat, A. Andrady, and J. Reisser, *River plastic emissions to the world's oceans*, Nature Communications **8**, 15611 (2017).
- [65] C. A. Choy, B. H. Robison, T. O. Gagne, B. Erwin, E. Firl, R. U. Halden, J. A. Hamilton, K. Katija, S. E. Lisin, C. Rolsky, and K. S. Van Houtan, *The vertical distribution and biological transport of marine microplastics across the epipelagic and mesopelagic water column*, Scientific Reports **9**, 7843 (2019).
- [66] M. Lange and E. van Sebille, *Parcels v0.9: prototyping a lagrangian ocean analysis framework for the petascale age*, Geoscientific Model Development **10**, 4175 (2017).
- [67] P. Delandmeter and E. van Sebille, *The parcels v2.0 lagrangian framework: new field interpolation schemes*, Geoscientific Model Development **12**, 3571 (2019).
- [68] R. Dussin, B. Barnier, and L. Brodeau, *DRAKKAR/MyOcean Report 01-04-16*, Tech. Rep. (LGGE, Grenoble, France., 2016).
- [69] G. Madec and M. Imbard, *A global ocean mesh to overcome the North Pole singularity*, Climate Dynamics **12**, 381 (1996).
- [70] P. P. Niiler and J. D. Paduan, *Wind-Driven Motions in the Northeast Pacific as Measured by Lagrangian Drifters*, Journal of Physical Oceanography **25**, 2819 (1995).

- [71] J. G. Derrai, *The pollution of the marine environment by plastic debris: a review*, Marine Pollution Bulletin **44**, 842 (2002).
- [72] E. van Sebille, *The oceans' accumulating plastic garbage*, Physics Today **68**, 60 (2015).
- [73] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal **27**, 379 (1948).
- [74] M. Camesasca, M. Kaufman, and I. Manas-Zloczower, *Quantifying fluid mixing with the shannon entropy*, Macromolecular Theory and Simulations **15**, 595 (2006).
- [75] R. McAdam and E. van Sebille, *Surface Connectivity and Interocean Exchanges From Drifter-Based Transition Matrices*, Journal of Geophysical Research: Oceans **123**, 514 (2018).
- [76] M. van der Mheen, C. Pattiaratchi, and E. van Sebille, *Role of indian ocean dynamics on accumulation of buoyant debris*, Journal of Geophysical Research: Oceans **124**, 2571 (2019).
- [77] L. Wu, W. Cai, L. Zhang, H. Nakamura, A. Timmermann, T. Joyce, M. J. McPhaden, M. Alexander, B. Qiu, M. Visbeck, P. Chang, and B. Giese, *Enhanced warming over the global subtropical western boundary currents*, Nature Climate Change **2**, 161 (2012).
- [78] L. M. Beal and S. Elipot, *Broadening not strengthening of the agulhas current since the early 1990s*, Nature **540**, 570 (2016).
- [79] A. Hadjighasem, M. Farazmand, D. Blazeovski, G. Froyland, and G. Haller, *A critical comparison of lagrangian methods for coherent structure detection*, Chaos: An Interdisciplinary Journal of Nonlinear Science **27**, 053104 (2017).
- [80] G. Froyland and K. Padberg-Gehle, *A rough-and-ready cluster-based approach for extracting finite-time coherent sets from sparse and incomplete trajectory data*, Chaos: An Interdisciplinary Journal of Nonlinear Science **25**, 087406 (2015).
- [81] K. Padberg-Gehle and C. Schneide, *Network-based study of lagrangian transport and mixing*, Nonlinear Processes in Geophysics **24**, 661 (2017).
- [82] R. Banisch and P. Koltai, *Understanding the geometry of transport: Diffusion maps for lagrangian trajectory data unravel coherent sets*, Chaos: An Interdisciplinary Journal of Nonlinear Science **27**, 035804 (2017).
- [83] R. Banisch, P. Koltai, and K. Padberg-Gehle, *Network measures of mixing*, Chaos: An Interdisciplinary Journal of Nonlinear Science **29**, 063125 (2019).

- [84] J. Shi and J. Malik, *Normalized Cuts and Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888 (2000).
- [85] D. Wichmann, P. Delandmeter, H. A. Dijkstra, and E. van Sebille, *Mixing of passive tracers at the ocean surface and its implications for plastic transport modelling*, Environmental Research Communications **1**, 115001 (2019).
- [86] F. Fouss, M. Saerens, and M. Shimbo, *Algorithms and Models for Network Data and Link Analysis* (Cambridge University Press, 2016).
- [87] U. Von Luxburg, *A Tutorial on spectral clustering*, Statistics and Computing **17**, 395 (2007).
- [88] N. Fan and P. M. Pardalos, *Multi-way clustering and biclustering by the ratio cut and normalized cut in graphs*, Journal of combinatorial optimization **23**, 224 (2012).
- [89] T. Ma and E. M. Bollt, *Relatively Coherent Sets as a Hierarchical Partition Method*, International Journal of Bifurcation and Chaos **23**, 1 (2013).
- [90] I. I. Rypina, L. J. Pratt, and M. S. Lozier, *Near-Surface Transport Pathways in the North Atlantic Ocean: Looking for Throughput from the Subtropical to the Subpolar Gyre*, Journal of Physical Oceanography **41**, 911 (2011).
- [91] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM Journal on Numerical Analysis **36**, 491 (1999).
- [92] D. Wichmann, *Lagrangian particle dataset (2 years) for Agulhas region surface flow*, (2020).
- [93] L. Brach, P. Deixonne, M. F. Bernard, E. Durand, M. C. Desjean, E. Perez, E. van Sebille, and A. ter Halle, *Anticyclonic eddies increase accumulation of microplastic in the North Atlantic subtropical gyre*, Marine Pollution Bulletin **126**, 191 (2018).
- [94] C. Dong, J. C. McWilliams, Y. Liu, and D. Chen, *Global heat and salt transports by eddy movement*, Nature Communications **5**, 3294 (2014).
- [95] G. Froyland, N. Santitissadeekorn, and A. Monahan, *Transport in time-dependent dynamical systems: Finite-time coherent sets*, Chaos: An Interdisciplinary Journal of Nonlinear Science **20**, 043116 (2010).
- [96] C. Schneide, A. Pandey, K. Padberg-Gehle, and J. Schumacher, *Probing turbulent superstructures in Rayleigh-Bénard convection by Lagrangian trajectory clusters*, Physical Review Fluids **3**, 113501 (2018).
- [97] G. Froyland and O. Junge, *Robust FEM-based extraction of finite-time coherent sets using scattered, sparse, and incomplete trajectories*, SIAM Journal on Applied Dynamical Systems **17**, 1891 (2018).

- [98] M. W. Schouten, W. P. M. de Ruijter, P. J. van Leeuwen, and J. R. E. Lutjeharms, *Translation, decay and splitting of Agulhas rings in the southeastern Atlantic Ocean*, Journal of Geophysical Research: Oceans **105**, 21913 (2000).
- [99] G. Dencausse, M. Arhan, and S. Speich, *Routes of Agulhas rings in the southeastern Cape Basin*, Deep-Sea Research Part I: Oceanographic Research Papers **57**, 1406 (2010).
- [100] W. Bickley, *Lxxiii. the plane jet*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **23**, 727 (1937).
- [101] D. del Castillo-Negrete and P. Morrison, *Chaotic transport by rossby waves in shear flow*, Physics of Fluids A: Fluid Dynamics **5**, 948 (1993).
- [102] D. Wichmann, C. Kehl, H. A. Dijkstra, and E. van Sebille, *Detecting flow features in scarce trajectory data using networks derived from symbolic itineraries: an application to surface drifters in the north atlantic*, Nonlinear Processes in Geophysics **27**, 501 (2020).
- [103] 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*, KDD'96 (AAAI Press, 1996) p. 226–231.
- [104] D. Wichmann, *Animation of finite-time coherent sets in the Agulhas region*, (2020).
- [105] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290**, 2323 (2000).

Acknowledgement

First, I want to thank Erik for the exciting time in his research group. For me, Erik has set a very high bar when it comes to effectively managing and leading a group of people with a positive and encouraging attitude. I appreciate that his door was always open for questions and discussions, and that he invests a lot of time into supervision and guidance of his students. The atmosphere he creates in his group was very stimulating for my research, giving me enough space to develop and discuss my own ideas, allowing me to align my PhD program with my interests and my private life.

Next, I want to thank Henk for his support during my PhD. I really enjoyed the mathematical approach of his lecture on dynamical oceanography, which made the topic accessible to me. Thank you also for the discussions we had on dynamical systems concepts related to my research, and for your input to the chapters of this thesis.

I had a great time at IMAU with all its different kind of people. Thank you Floor, Sandra, Clara and Marcel for your constant support. Thanks to all the people in the ocean group for sharing ideas and for the discussions in the werkberepreken. I also want to thank the team of the Centre for Complex Systems Studies, in particular Qingyi, for organizing talks and research on complexity which really broadened my horizon. Thanks also to Mara, Alje and Mark for involving me in an exciting joint research project on ecology at the Centre. I also want to thank Jörg Lippold from Heidelberg University for introducing me to the oceanography topic during my B.Sc., and for writing an uncountable amount of recommendation letters for most of the different research-related positions I applied for after my undergraduate studies. I also want to thank Claus Kiefer, my M.Sc. supervisor at the University of Cologne, for the joint work in theoretical physics and for raising my interest in theoretical and mathematical physics through his lectures on relativity.

Next, I want to thank the entire Parcels group I interacted with during my time as PhD student, including Mikael, Daan, Victor, Cleo, Delphine, Steffie, and Reint for the discussions and the time we had together, the Lagrangian viewpoint, and for making my time at IMAU a good and enjoyable time.

I also want to thank Philippe for his non-exhaustible effort in making the Parcels code better, discussing mathematical concepts and for helping me out with all sorts of technical problems in the numerical simulation of Lagrangian particles for chapters 2 and 3.

Many thanks also goes to Christian for implanting the seed of density-based clustering in my head, which resulted in chapter 5, as well as for his down-to-earth attitude when it comes to techniques commonly known as machine learning.

I also want to mention Peter, my office mate in pre-Corona times, and thank

him for his enjoyable company in the office. Never before have I met someone with such a high peanut butter turnover.

Although I do not know him in person, I also want to thank Andrew Coward from NOC Southampton for providing the NEMO ORCA-N006 simulation data, without which chapters 2, 3 and 5 would not have been possible.

Special thanks goes to Tjebbe for taking care of my daughter regularly, as well as for his tolerance regarding my coffee taste. I also want to thank Daniele, as the way he managed his office desk made me very proud of my own desk every day. Thanks also goes to André. In his presence, I never needed my phone to get any sort of useless information, which really reduced my energy bill and brought me closer to nature. Finally, I want to thank Martina for the regular supply with foodstuffs, and her true and honest commitment to science and rice.

Special thanks goes to my parents for their support during my entire educational career that lead me to this PhD and made me to what I am today. I also appreciate their permanent willingness to move around all sort of furniture from one place to the other.

Finally, I want to thank Verena for her unlimited support in this PhD project. You understand me better than anyone else, and being with you is what really kept me going and motivated during these years. I enjoyed our daily discussion about our PhD journeys, and I am at the same time happy that they come to an end after four years. Thank you in particular for taking care of Clara most of the time when the pandemic started and I had to finish the last two chapters. I would never have managed to finish this thesis otherwise. I also want to thank my children Clara and ? for permanently reminding me what really matters to me. In some way, they are the major outcome of this PhD journey, with many more chapters to come.

Curriculum Vitæ

David Wichmann

25-05-1990 Born in Balingen, Germany.

Education

2000–2009 Grammar School
Gymnasium Ebingen

2010–2013 B.Sc. in Physics
Heidelberg University, Germany

2010–2013 B.Sc. in Economics
Heidelberg University, Germany

2013–2015 MPA in International Development
Tsinghua University, China

2015–2017 M.Sc. Physics
University of Cologne, Germany

2017–2021 PhD. Physical Oceanography
Utrecht University

List of Publications

1. van Dam, A. et al. (2021): Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Sci Rep* 11, 8926.
<https://doi.org/10.1038/s41598-021-87971-9>
2. Wichmann, D., Kehl, C., Dijkstra, H. A., and van Sebille, E. (2021): Ordering of trajectories reveals hierarchical finite-time coherent sets in Lagrangian particle data: detecting Agulhas rings in the South Atlantic Ocean, *Nonlin. Processes Geophys.* 28, 43–59.
<https://doi.org/10.5194/npg-28-43-2021>
3. Wichmann, D., Kehl, C., Dijkstra, H. A., and van Sebille, E. (2020): Detecting flow features in scarce trajectory data using networks derived from symbolic itineraries: an application to surface drifters in the North Atlantic, *Nonlin. Processes Geophys.*, 27, 501–518.
<https://doi.org/10.5194/npg-27-501-2020>
4. Van Sebille, Erik, et al. (2020): The physical oceanography of the transport of floating marine debris. *Environmental Research Letters*, 15(2), 023003.
<https://doi.org/10.1088/1748-9326/ab6d7d>
5. Wichmann, D., Delandmeter, P., Dijkstra, H. A., and van Sebille, E. (2019): Mixing of passive tracers at the ocean surface and its implications for plastic transport modelling. *Environmental Research Communications*, 1(11), 115001.
<https://doi.org/10.1088/2515-7620/ab4e77>
6. Wichmann, D., Delandmeter, P., and van Sebille, E. (2019): Influence of near-surface currents on the global dispersal of marine microplastic. *Journal of Geophysical Research: Oceans*, 124, 6086–6096.
<https://doi.org/10.1029/2019JC015328>
7. Onink, V., Wichmann, D., Delandmeter, P., and van Sebille, E. (2019): The role of Ekman currents, geostrophy, and stokes drift in the accumulation of floating microplastic. *Journal of Geophysical Research: Oceans*, 124, 1474– 1490.
<https://doi.org/10.1029/2018JC014547>
8. Kiefer, C., Wichmann, D. (2018): Semiclassical approximation of the Wheeler–DeWitt equation: arbitrary orders and the question of unitarity. *Gen Relativ Gravit* 50, 66.
<https://doi.org/10.1007/s10714-018-2390-4>

9. Mulitza, S., et al. (2017): Synchronous and proportional deglacial changes in Atlantic meridional overturning and northeast Brazilian precipitation, *Paleoceanography*, 32, 622– 633.
<https://doi.org/10.1002/2017PA003084>