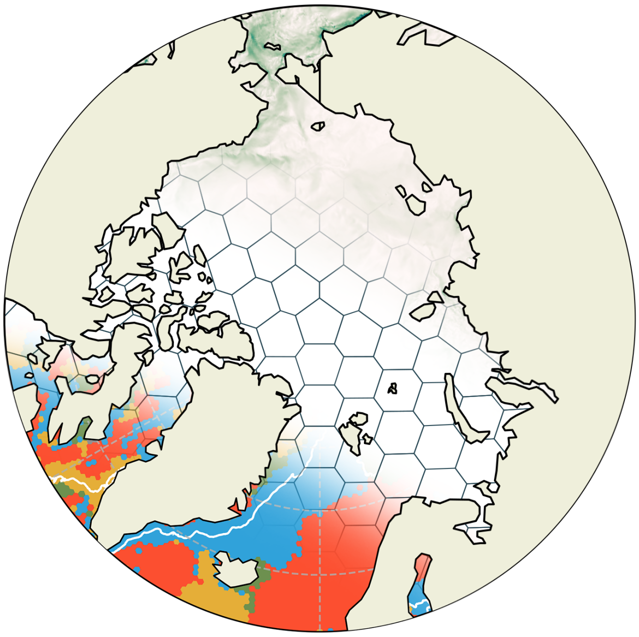
*Master's thesis:*

# Assessing Ocean Surface Connectivity in the Arctic

## Capabilities and caveats of community detection in Lagrangian Flow Networks



Daan Reijnders (4302001)

Supervisors:
Dr. Erik van Sebille
Dr. Erik Jan van Leeuwen
David Wichmann, MSc

Utrecht University, January 2020

# Abstract

Community detection algorithms from the field of network theory have been used to divide a fluid domain into clusters that are sparsely connected with each other and to identify barriers to transport, for example in the context of larval dispersal. Communities detected by the community detection algorithm *Infomap* have barriers that have been shown to often coincide with well-known oceanographic features. Thus far, this method has only been applied to closed domains such as the Mediterranean. We apply this method to the surface of the Arctic and subarctic oceans and show that it can be applied to open domains. First, we construct a Lagrangian flow network by simulating the exchange of Lagrangian particles between different bins in an icosahedral-hexagonal grid. Then, *Infomap* is applied to identify groups of well-connected bins. The resolved transport barriers include naturally occurring structures, such as the major currents. As expected, clusters in the Arctic are affected by seasonal and decadal variations in sea-ice concentration. We also discuss several caveats of this method. Firstly, there is no single definition of what makes a cluster, since this is dependent on a preferred balance of internally high connectivity, sparse connectivity between clusters, and the spatial scale of investigation. Secondly, many different divisions into clusters may qualify as good solutions and it may thus be misleading to only consider the solution that optimizes a certain quality parameter the most. Finally, while certain cluster boundaries lie consistently at the same location between different good solutions, other boundary locations vary significantly, making it difficult to assess the physical meaning of a single solution. Particularly in the context of practical applications like planning Marine Protected Areas, it is important to consider an ensemble of qualifying solutions to find persistent boundaries.

# Layman's summary

Currents and eddies in the surface of the Arctic Ocean move water and particles between different regions of the Arctic and thus determine the connectivity different regions. To assess which regions are connected to one another, we try to divide the Arctic into different regions. To do so, we first simulate the movement of particles in the ocean surface, incorporating observations of the ocean and the equations that drive ocean flow. Then, we divide the ocean into boxes and investigate the exchange of particles between different boxes. We use the computer algorithm *Infomap* to find groups of boxes that exchange relatively many particles among each other and relatively few particles with other boxes. Knowledge about which regions in the Arctic Ocean are connected to one another is important for planning areas of conservation, such that marine species can travel between different areas of conservation.

It is important to be careful with the interpretation of regions identified by *Infomap*, since the division into regions can differ each time that *Infomap* is run. Each division is good in principle, since particles tend to stay within each region and the exchange of particles between regions is low. However, since the boundaries between regions can differ, it is important to run *Infomap* multiple times and see which boundaries occur persistently. *Infomap* does not use the ocean flow directly to find connected regions, but instead uses a simplified mathematical representation of the flow. It is important to take this into account when interpreting a division into connected regions.

We find that the boundaries between different connected regions often coincide with ocean currents, meaning that the presence of an ocean current can hinder the exchange of particles between two regions. We also find that the division into connected regions is affected by sea ice. Since the amount of sea ice in the Arctic differs across seasons and since it is decreasing over the years due to climate change, the division into connected regions also changes seasonally and over multiple years.

# Comment with respect to thesis requirements

This thesis serves to fulfill the graduation requirements for the master's program Climate Physics at Utrecht University, as well as the requirements for the Complex Systems profile. It does so by combining aspects from oceanography, network theory and information theory. Mainly related to oceanography are the simulation of Lagrangian particles, the assessment of the quality of hydrodynamic provinces in terms of coherence and mixing, the application to the Arctic domain, the assessment of transport barriers, connections to surface velocities and sea ice, and the investigation of trends. Aspects that mainly pertain to network and information theory are the construction of Lagrangian flow networks, a description of how *Infomap* works, experiments related to *Infomap*'s configuration, and the assessment and discussion of degeneracy. However, this research is interdisciplinary and the different topics treated in this thesis are inherently intertwined.

# Acknowledgements

Just like this thesis relies on insights from multiple disciplines, I want to extend my gratitude to the many people that helped shape this thesis. In particular, my three supervisors each played instrumental roles, for which I am very thankful. While their support extends through many aspects of this thesis, I want to highlight specific areas where their help was indispensable. Specifically, Dr Erik van Sebille introduced me to the topic of connectivity in the Arctic and stressed the implications for planning Marine Protected Areas. Together with a vibrant open-source community, Erik also develops the software package *Parcels*, which greatly simplified the particle simulations carried out in experiments. Dr Erik Jan van Leeuwen contributed through fruitful discussions on how *Infomap* functions and underlined important computational aspects of the methods used throughout this thesis. David Wichmann was always available for critical -and thus fruitful- discussions on the clustering of transition matrices, solution degeneracy and implications to mixing.

I also want to thank the other members of the *Physical Oceanography* and *OceanParcels* groups at Utrecht University for their discussions and feedback. The support of Dr Philippe Delandmeter with running *Parcels* was invaluable. I have greatly benefited from the discussions with Dr Erwin Lambert on the influence of major surface currents in the Arctic. I am also grateful to Michael Kliphuis for his support with hydrodynamic data. The many discussions with Anneke Vries and Reint Fischer about numerous topics related to the Arctic, Lagrangian particle simulations and plotting techniques have also proved invaluable.

Multiple people have also supported me outside of Utrecht University. I want to thank Dr Martin Rosvall for providing support with *Infomap* and for letting me use his illustrative figure that explains the core notion behind the algorithm. I also want to express my gratitude to Prof Louis Moresi for his help with constructing the icosahedral-hexagonal grid used throughout this thesis. Lastly, I have received the generous support of Jurre Wijnhoven, who provided me with vital computational resources.

# Contents

# 1 | Introduction

Different regions of the global ocean are connected by flowing currents and eddies. Looking through the Lagrangian lens, these currents and eddies facilitate the exchange of fluid parcels, that move along chaotic trajectories which change through space and time. The ocean pathways of many objects suspended in fluid can be studied using Lagrangian analysis [70], including the larvae of marine species [32, 60]. With knowledge of how particles travel through different geographical areas of a fluid domain, we can investigate the exchange of particles between different areas, in order to assess spatial connectivity.

Connectivity is a widely-used term in marine ecology and marine spatial planning, where it is used in the context of the exchange of individuals of a species between different, geographically separated sub-populations [11, 60]. In this context, connectivity is important for safeguarding the genetic exchange and productivity of marine species [55]. Therefore, connectivity between regions is taken into account when planning marine protected areas (MPAs) [55, 9]. The exchange between sub-populations is determined by many factors, such as spawning behavior, larval dispersal, predator-prey survival, habitat availability and larval conditions. Larval dispersal is the dominant process contributing to the spatial aspect of population connectivity [11]. Larval dispersal may be a largely passive process for some species, while other species are capable of orienting and navigating their movements through directed horizontal swimming [11]. Larval traits such as spawning time, swimming behavior and survival are incorporated in some connectivity modeling studies [32, 7], while other studies simply model larvae as passive particles [1], in certain cases also neglecting vertical effects by modeling them as buoyant particles [60].

If we simplify larval dispersal as a completely passive process, the definition of *connectivity* in the context of marine ecology becomes generalized as the exchange of any passive particle between geographical regions. Through this generalization, any quantification of connectivity thus becomes applicable for any object that can be modeled as a Lagrangian particle, such as marine debris [43], phytoplankton [5], or fluid parcels themselves. Globally, this definition of connectivity of the ocean surface has been investigated using the Lagrangian approach in the context of identifying basins of attraction [22].

Several modeling studies have aimed to quantify connectivity between existing MPAs or localized population sites in order to cluster sets of regions that have an internally high connectivity [32, 1, 75]. Rossi et al. [60] generalized this approach by not just clustering spatially separated regions, but instead considering an entire fluid domain. To do so, they describe flow in the Mediterranean Sea as a *Lagrangian flow network* and use clustering methods from network theory to divide the network into groups of boxes that are sparsely connected with one another. These groups are referred to as *hydrodynamic provinces.* This approach was first presented in the context of MPAs, since the boundaries between hydrodynamic provinces can be understood as barriers to larval transport, thus hindering the connectivity between MPAs. In certain cases, these boundaries have been shown to coincide with well-known oceanographic features [60]. Since Rossi et al. modeled larvae as passive Lagrangian particles, the retrieved hydrodynamic provinces bear relevance not only to larvae, but to Lagrangian particles in general, and boundaries between clusters can be interpreted as barriers to flow itself [71].

We note that hydrodynamic provinces differ from *Lagrangian coherent structures.* Much work has been put into the detection of Lagrangian coherent structures, which are delineated by material lines that are linearly stable or unstable for longer times than surrounding regions [28]. For a comparison between Lagrangian coherent structure detection methods, see Hadjighasem et al. [27]. While Lagrangian coherent structures may move in space with the mean flow, by identifying hydrodynamic provinces we instead aim to partition an entire domain into time-invariant localized coherent regions [71]. Moreover, hydrodynamical provinces are not only characterized by small fluid exchange across their boundaries, but also by high internal mixing. Both properties are particularly useful in the context of planning networks of MPAs, where one seeks to maintain connectivity between fixed areas in space [60, 71].

This thesis aims to investigate surface connectivity in the Arctic and subarctic oceans through identification of hydrodynamic provinces using the Lagrangian flow network approach. Studying the Arctic through this approach is interesting for multiple reasons. Firstly, community detection has only been applied to Lagrangian flow networks in the Mediterranean, which can be approximated as a closed domain, while the Arctic and subarctic Oceans comprise a domain that is open at the southern boundary. We therefore investigate whether this approach is successful at identifying meaningful communities in open domains. Secondly, the Arctic ocean experiences strong seasonal variations in the strength and location of ocean currents, as well as seasonal variations in the sea ice extent, with sea-ice affecting surface flow [24]. These variations in domain topology and hydrodynamics will therefore influence the location of barrier to flow. We compare connectivity between different seasons and years, in order to see which physical mechanisms are governing barriers to transport. Thirdly, the average Arctic sea ice extent has been decreasing over the past couple decades and is very likely to decrease in the future [10, 76]. Since this decrease will cause a potentially irreversible shift into a new climatic state [56], it is insightful to see whether these developments are reflected back in the topology of hydrodynamic provinces. Lastly, while efforts for planning networks of marine protected areas in the Arctic ocean are currently underway [57], this study provides the first assessment of the connectivity of the surface ocean in the Arctic.

We also aim to describe important considerations when using this approach and to raise caveats that have not been previously discussed. This includes a discussion of what should make a good community, as well as what should be the physical interpretation of communities found by the community detection algorithm *Infomap* [61]. Moreover, community detection algorithms in complex networks have been shown to be sensitive to degenerate solutions, meaning that many good solutions may exist, while their topology may significantly differ [23, 6]. We therefore assess which structures are persistently found between different solutions.

With these aims in mind, this thesis is structured as follows. First, we provide a theoretical description of Lagrangian flow networks, community detection using *Infomap* and hydrodynamic provinces in chapter 2. This theoretical breakdown closely follows Rossi et al. [60] and Ser-Giacomi et al. [71], who first presented this approach to studying geophysical fluids. We elucidate step-by-step how *Infomap* functions, including important tuning mechanisms, and relate these to the underlying flow when possible. Then, we give a brief overview of oceanographic structures in the Arctic Ocean and hypothesize how these may influence community structures. With this theoretical basis in mind, chapter 3 describes our methodology for assessing Arctic ocean surface connectivity, expanding on the methodology from Rossi et al. [60] and Ser-Giacomi et al. [71]. We provide a detailed description of our data, parameters and method for assessing the quality of communities returned by *Infomap*. Next, different experiments related to *Infomap*'s configuration and connections to the ocean surface and sea ice are presented and their results are reported in chapter 4. These are discussed in chapter 5, where we also present several important considerations and caveats of using community detection for assessing connectivity. The thesis is concluded in chapter 6.

Annotated code for running the experiments in this thesis is available on `https://github.com/daanreijnders/arctic-connectivity`.

# 2 | Theory

The characterization of a fluid as a network constructed from Lagrangian trajectories was first introduced by Rossi et al. [60] and later described from a more technical perspective by Ser-Giacomi et al. [71]. By mapping flow onto a network, the dynamics of the fluid system are captured by the topology of the network [46]. This enables us to analyze these dynamics using the vast toolkit of network science that has become available in the past couple of decades. This toolbox is rich due to the fact that many problems throughout different disciplines can be approached by representing systems as networks, enabling an interdisciplinary cross-pollination of problem-solving methods [52]. Examples of such systems, specifically in which a quantity flows between the components of a system, include the flow of passengers between airports [26], transactions between banks [73], and the spread of innovations between individuals [47]. A frequently recurring problem is the division of a network into communities of nodes that are well connected among each other, with only sparse connections between distinct communities [51] and many approaches have been put forward for tackling this problem. For comparisons, see Danon et al. [13] and Fortunato [20]. In the context of a flow network, ideally such a division yields barriers to fluid transport, with fluid being unlikely to cross these barriers. Simultaneously, high connectivity within a network community should correspond to the fluid within one community being well-mixed.

## 2.1 Lagrangian Flow Networks

A network representation of a system comprises a graph $G = (V, E)$, consisting of a set of *nodes*, $V$, and a set of *edges*, $E$, where an edge $(i, j) \in E$ forms a connection between nodes $i, j \in V$. Additionally, these edges may be directed, so that an edge from node $i$ to node $j$ is distinct from an edge from $j$ to $i$. Moreover, edges may take on weights $w_{ij}$, which may correspond to the importance of a connection. For example, they can represent quantities like the flow of passengers between airports in an airport network or the number of citations in a citation network. In our practical application, $V$ and $E$ are finite sets.

When mapping fluid flow as a network, the fluid domain needs to be discretized in order to represent the continuous flow by the finite sets $V$ and $E$. We can divide the flow domain into a set of $N_B$ bins, $B = \{B_i,\ i = 1, \ldots, N_B\}$, and consider the flow between different bins. These bins then form the nodes of the network, while the flow between bins is captured by the edges between nodes. Since the flow between bins is directional and may differ in magnitude, edges should be weighted and directed.

Fluid flow in our ocean is subject to chaos and many external forcings, some of which are seasonal. The flow between bins is therefore dependent on the initial state of the fluid at time $t_0$ and the time interval $\tau$ in which the flow is considered. In Lagrangian flow networks, we establish a connection between node $i$ and node $j$ if there is exchange of fluid (or equivalently, Lagrangian particles) from the corresponding bin $B_i$ to bin $B_j$ in the time interval $[t_0, t_0 + \tau]$. The weight of edge $(i, j)$ will then be proportional to the amount of fluid that is transported from $B_i$ to $B_j$.

From a Lagrangian perspective, the fluid transport can be determined from the initial and final positions in the trajectories of ideal fluid particles. These trajectories can be determined through integration of the equations of motion of particles. Final particle positions $\mathbf{X}(t_0 + \tau)$ are then given by

$$\mathbf{X}(t_0 + \tau) = \mathbf{X}(t_0) + \int_{t_0}^{t_0+\tau} \mathbf{v}(\mathbf{x}(t), t)dt, \tag{2.1}$$

where $\mathbf{v}(\mathbf{x}, t)$ is the time-dependent Eulerian velocity field [70]. Then, the right-hand side of (2.1) defines the flow map $\Phi_{t_0}^{\tau}$, which maps the initial location $x$ of a fluid particle to its final location.

Given $m(B_i)$ Lagrangian particles initially being distributed in bin $B_i \in B$, we can approximate the flow probability between bins by considering the fraction of particles traveling from bin $B_i$ to bin $B_j$ in time window $[t_0, t_0 + \tau]$ by

$$\mathbf{P}(t_0, \tau)_{ij} = \frac{\#\{x : x \in B_i \text{ and } \Phi_{t_0}^{\tau}(x) \in B_j\}}{m(B_i)}, \tag{2.2}$$

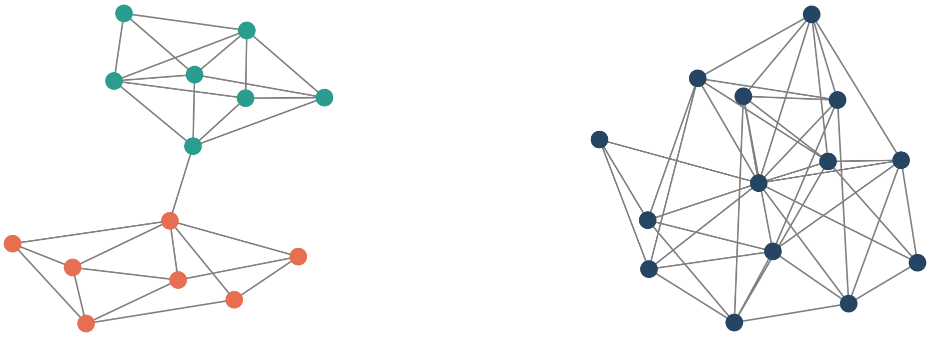
which allows us to construct a transition matrix $\mathbf{P}(t_0, \tau)$ [22, 60, 71]. Therefore, $\mathbf{P}(t_0, \tau)$ defines the approximation of our flow as a Markov chain. As long as fluid parcels, or equivalently, particle trajectories are conserved, $\mathbf{P}(t_0, \tau)$ is row-stochastic, such that for each bin $B_i$, $\sum_{j=1}^{N_B} \mathbf{P}(t_0, \tau)_{ij} = 1$ and each element is non-negative. Note that the definition of our transition matrix (2.2) only considers the initial and final location of the particles, and thus contains no information about fluid exchange between bins at intermediate times.

The transition matrix $\mathbf{P}(t_0, \tau)$ can be used as an adjacency matrix to generate a graph $G_{t_0}^{\tau}$, which is our network representation of the flow. Each row and column index corresponds to a node, and the weight of an edge $w(i, j)$ is given by the entry $\mathbf{P}(t_0, \tau)_{ij}$. In the network representation of the fluid, edge weights thus correspond to the probability that a particle travels between bins and the row-stochastic property of $\mathbf{P}(t_0, \tau)$ ensures that the sum of the weights of outgoing edges for any given node is 1.

## 2.2 Finding Hydrodynamic Provinces through Community Detection

Having obtained a discrete description of our time-dependent flow, we can now look for coherent regions. Hydrodynamic provinces obtained by applying community detection to Lagrangian flow networks should only be sparsely connected with one another. Simultaneously, we require high connectivity within a hydrodynamic province, meaning that the interior of each province should be well-mixed, such that we optimize for fluid from one location in the hydrodynamic province to be exchanged evenly to other locations in the province. Put differently, the corresponding region in the graph should be well connected [71]. This is the goal of community detection in network theory, and many approaches have been proposed to divide a network into distinct communities satisfying this requirement [20, 52].

There is no single definition of what constitutes a community. Given our objective of finding communities that have few edges running to their neighbors and that have many edges in their interior, both requirements may easily be satisfied for certain graphs (see figure 2.1). However, other graphs may have structures in which we cannot easily infer a



**(a)** A graph in which a community structure (orange versus turquoise) can readily be inferred.

**(b)** A graph without an obvious community structure.

**Figure 2.1:** Communities may or may not naturally occur in networks.

division into communities (see figure 2.1b). In the second case, one may wonder whether we should be looking for such a division at all, since it is unclear whether we may find a balance of our requirements that still yields a meaningful community division. Before applying a community detection strategy, it is therefore important to first investigate the characteristics of the network at hand and, if community detection should be applied, think about how high internal connectivity and sparse external connectivity can be balanced in a meaningful manner. Moreover, it is important to determine a scale at which the investigation of communities can lead to meaningful results, since communities may exhibit a nested, multilevel structure [37]. Different aims for community detection have lead to the development of different approaches. The definition of what constitutes a community implicitly depends on the underlying detection strategy, causing different algorithms to detect community structures of different nature [64].

Previous work on partitioning transition matrices in order to find almost-invariant sets [21] actually draw on classical spectral partitioning methods methods from network theory [52]. While these methods satisfy the criterion of minimal fluid exchange along structure boundaries, they do not impose the criterion of strong mixing [71]. Furthermore, this method identifies structures of similar sizes, while communities in flow networks may well exhibit many different sizes [71].

One popular measure for detecting communities is the modularity maximization method [51, 52]. This method relies on the comparison of a given network with a random network or other null model, in order to determine which regions of the network exhibit more connections than would be expected in a random network. Modularity maximization has a couple of limitations. Although there exist implementations of modularity maximization that take the direction of edges into account, most methods neglect directionality of edges [38]. Moreover, the null models used by modularity maximization carry no obvious meaning with respect to flow networks, such that this method lacks a physical interpretation [71]. Lastly, these methods suffer from a *resolution limit*, preventing us from detecting communities smaller than a specific scale that is determined by the size of the total network [19].

With these limitations in mind, Rossi et al. [60] and Ser-Giacomi [71] instead propose using the *Infomap* [61] community detection method for our purposes. *Infomap* uses an information-theoretic approach to identify communities based on the flow within a network. This approach was first introduced by Rosvall and Bergstrom [62] and was expanded on and presented as a software package by Rosvall et al. [61].

*Infomap* is a community detection algorithm that takes edge directionality and weights into account and its solutions are less likely to be impacted by a resolution limit than other methods [35]. In addition, it can find communities that may differ in size. *Infomap* also allows the study of community structures at different scales, either through identifying nested communities [17], or through a tuning parameter that affects community sizes [37]. The information-theoretic approach used by *Infomap* does not have a direct physical meaning, but connections between this approach and our criteria for hydrodynamic provinces are explained in the next section.

### 2.2.1 How does *Infomap* work?

*Infomap* aims to identify communities by considering the dynamics that are governed by the edge structure of a network. Specifically, it aims to partition the network into communities such that it minimizes the average length of encoded trajectories of random walkers, which traverse the network with probabilities corresponding to the local importance of edges. For our flow network, edge weights directly correspond to the transition probabilities of Lagrangian particles. The transition probabilities of Lagrangian particles will thus be used by *Infomap* to drive the movement of the random walkers. While many community detection methods, like modularity maximization, focus on the topology of a network compared to a null model, the random walkers used by *Infomap* simulate flow on a network. The following account of how *Infomap* works closely follows Rosvall et al. [61] and subsequent extensions [39, 37]

*Infomap* capitalizes on the fact that trajectories of random walkers can be encoded using Huffman codes [31], which constitute an optimally efficient encoding. The trajectory of a random walker consists of the sequence of nodes that it traverses. Visiting a node can be regarded as an event in the trajectory, and each possible event can be encoded by assigning it a unique string of bits. Correspondingly, each node in the network is given its unique codeword. Huffman codes are optimally efficient by assigning short codewords to common events and long codewords to rare ones, with no codeword being the prefix of another. This implies that short codewords are assigned to nodes that are visited frequently, while longer codes are assigned to nodes with a low ergodic visiting frequency. Through this method, we can communicate a node-to-node trajectory of a random walker using a concatenated sequence of codes, which we refer to as a path description. The average length of a codeword will grow in size as the number of nodes in the network increases, as more bits are needed to describe a node using a unique codeword. This in turn leads to longer path descriptions.

Rather than describing a network only in terms of nodes and edges, we can consider the nodes in a network to be divided among communities. When a network is made up of communities that are characterized by few edges between communities and many edges within a community, a random walker is likely to spend a long time within a community before moving to another. This allows the construction of a two-level encoding system, where we have a separate encoding for the events of entering each community and for the movement of a random walker within a community, such that each community has its own encoding. Since nodes in different communities may then be assigned the same codewords, codewords are shorter. If a random walker does not switch between communities often such that the corresponding codewords are not used often, path descriptions can in turn become shorter on average. This is visualized in figure 2.2. Finding a good division into communities will then yield shorter path descriptions. This implies a minimization strategy for finding good communities, namely to find a partition that minimizes the average length of codewords used in a random walk.

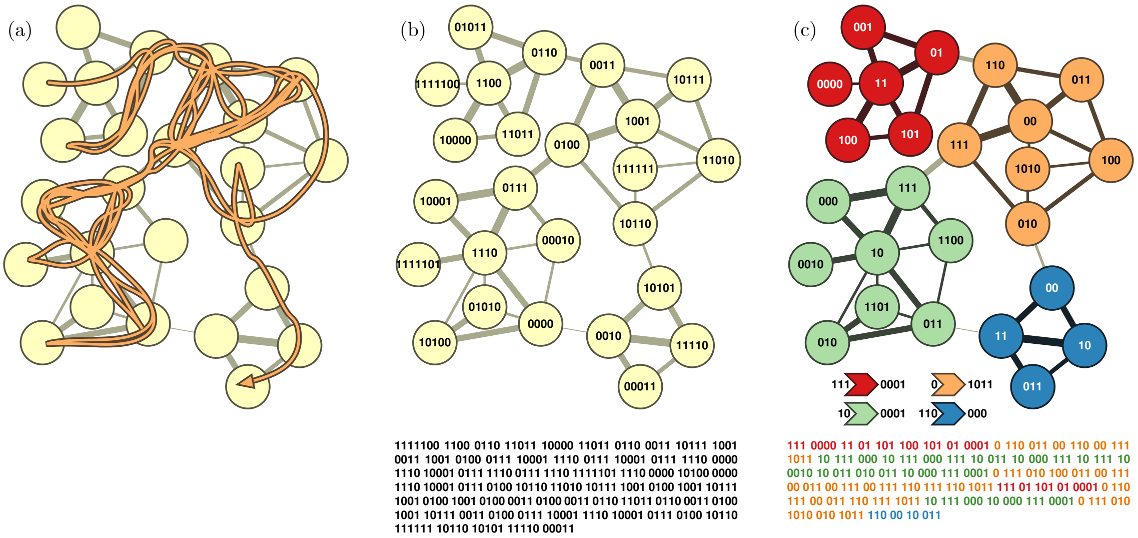


**Figure 2.2:** Finding a good community division can reduce the length of the path description of a random walker traversing a network (adapted from Rosvall and Bergstrom [62] with permission from the authors). (a) Flow within a network is simulated using a random walker that moves from node to node. The orange line shows one sample trajectory. An optimally efficient encoding is given by the codewords depicted in (b), with codeword lengths varying from 4 to 7 bits. The trajectory of the random walker is encoded in 314 bits, starting with 1111100 for the first node in the random walk in the top left, 1100 for the second node, etc. Codewords are separated by spaces for visualization purposes, but in principle they can be sent concatenated. In (c), the same trajectory is encoded, now using a two-level description. The codewords used for movement between communities are indicated using colored arrows (entering a community on the left, exiting a community on the right). The lengths of codewords for movement within a community range between 2 and 4 bits. The walk from (a) can now be encoded in 243 bits. The first three bits 111 indicate the walk begins in the red community, followed by 0000 indicating the starting node, 11 indicating the second node, and so forth.

Rather than explicitly determining the most efficient encoding for a given community division, *Infomap* instead takes advantage of concepts from information theory to find the theoretical lower bound of the average codeword length. Given a partition $\mathsf{P}$ that divides the $n$ nodes in $V$ into $c$ communities $\alpha = 1, 2, \ldots, c$, this lower bound is denoted by $L(\mathsf{P})$. To find an expression for this lower bound, Shannon's source coding theorem is used [72], which implies that the average length of a codeword is bounded from below by the entropy of the random variable $X$, the $n$ states of which are described by $n$ distinct codewords. With $p_i$ denoting the frequency of occurrence of a state, the Shannon entropy is then given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i). \tag{2.3}$$

The information-theoretical lower bound on the average length of a codeword describing a step of the random walk is then found through a weighted average of the entropy associated to the length of the codewords describing entering a new community and the entropies corresponding to the codewords describing steps in each community. This is captured in the map equation:

$$L(\mathsf{P}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{\alpha=1}^{c} p_{\circlearrowright}^{\alpha} H(\mathcal{P}^{\alpha}). \tag{2.4}$$

Here $H(\mathcal{Q})$ is the frequency-weighted average codelength corresponding to the codewords that signal entering a new community, while $H(\mathcal{P}^\alpha)$ is the frequency-weighted average codeword length describing steps within a community $\alpha$. These entropy terms are respectively weighted by the probability that a random walker exits a community, $q_{\curvearrowright}$, and the probability of using the codes corresponding to steps in community $\alpha$, denoted by $p^{\alpha}_{\circlearrowright}$. If $q_{\alpha\curvearrowright}$ denotes the probability of exiting community $\alpha$, the probability to leave any community is $q_{\curvearrowright} = \sum_{\alpha=1}^{c} q_{\alpha\curvearrowright}$. Then, the entropy associated to encoding switches between communities is

$$H(\mathcal{Q}) = -\sum_{\alpha=1}^{c} \frac{q_{\alpha\curvearrowright}}{q_{\curvearrowright}} \log_2\left(\frac{q_{\alpha\curvearrowright}}{q_{\curvearrowright}}\right). \tag{2.5}$$

An expression for the entropy related to encoding movements within a community, including a signal to exit the community, $q_{\alpha\curvearrowright}$, is

$$H(\mathcal{P}^\alpha) = -\sum_{i\in\alpha} \frac{\pi_i}{p^{i}_{\circlearrowright}} \log_2\left(\frac{\pi_i}{p^{\alpha}_{\circlearrowright}}\right) - \frac{q_{\alpha\curvearrowright}}{p^{\alpha}_{\circlearrowright}} \log_2\left(\frac{q_{\alpha\curvearrowright}}{p^{\alpha}_{\circlearrowright}}\right). \tag{2.6}$$

Here, $\pi_i$ is the ergodic frequency of a random walker visiting node $i$. The probability of using the codes corresponding to movements within community $\alpha$ are then $p^{\alpha}_{\circlearrowright} = q_{\alpha\curvearrowright} + \sum_{i\in\alpha} \pi_i$.

In the case of our directed network, the ergodic visiting frequency of a node cannot be readily determined from the adjacency matrix $\mathbf{P}(t_0, \tau)$, since the corresponding Markov chain is not necessarily irreducible. Put differently, the corresponding network is not necessarily strongly connected, meaning that from any given node, it may be impossible to reach all other nodes by following directed edges. In a Lagrangian flow network, this can be due to the fact that surface velocities are not divergence-free since the actual flow is three-dimensional. Having only knowledge about the surface, Lagrangian particles may be attracted to regions exhibiting convergence, corresponding to downwelling, while they are repelled from regions exhibiting divergence, corresponding to upwelling. Nodes corresponding to upwelling regions may thus have no incoming edges. Even when a divergence-free field is considered, disconnected regions in a basin or insufficient Lagrangian trajectories may also cause the flow network not to be strongly connected.

In order to still find a steady-state visiting frequency, a small probability $\sigma$ that the random walker will teleport to any other node at random is introduced into the Markov chain. Doing so, each node can now be reached from any other node, making the corresponding modified Markov chain irreducible and aperiodic. Then, according to the Perron-Frobenius theory, there exists one unique steady state for the visiting frequencies $\boldsymbol{\pi}$ [61], which we need for evaluating equation (2.6). We reduce the dependency of $\boldsymbol{\pi}$ on $\sigma$ by making the probability of teleporting to a node proportional to the total weight of the edges pointing to that node. The node visiting frequencies $\boldsymbol{\pi}$ can then be computed iteratively through

$$p_{i;k+1} = (1-\sigma)\sum_{j} \mathbf{P}(t_0,\tau)_{ij} p_{j;k} + \sigma \frac{\sum_j \mathbf{P}(t_0,\tau)_{ji}}{\sum_{i,j} \mathbf{P}(t_0,\tau)_{ji}}, \tag{2.7}$$

until $p$ converges to $\boldsymbol{\pi}$. The first term on the right-hand side of (2.7) corresponds to reaching a node by arriving to it from its neighbors, while the second term corresponds to visiting the node by teleportation. The procedure of calculating the steady state visiting frequencies is in fact equal to calculating the PageRank of each node [4].

Even though introducing teleportation is essential to be able to find the ergodic visiting frequency $\boldsymbol{\pi}$, it comes with a major drawback, namely that this approach introduces artificial links between nodes in different communities. To circumvent this, the iterative calculation of $\mathbf{p}$ can be adjusted by only recording the steps of random walkers along links, without recording teleportation steps. This is the *unrecorded teleportation* scheme [39], which comprises a system of three iterative equations [3]:

$$\mathbf{p}^*_{i;k+1} = (1-\sigma)\sum_j \mathbf{P}(t_0,\tau)_{ij}\mathbf{p}_{j;k} + \sigma\frac{\sum_j \mathbf{P}(t_0,\tau)_{ij}}{\sum_{i,j}\mathbf{P}(t_0,\tau)_{ji}}, \tag{2.8a}$$

$$\mathbf{q}_{(j,i);k+1} = \mathbf{p}^*_{j;k+1}\mathbf{P}(t_0,\tau)_{ji}, \tag{2.8b}$$

$$\mathbf{p}_{i;k+1} = \sum_j \mathbf{q}_{j,i;k+1}. \tag{2.8c}$$

Equation (2.8a) corresponds to equation (2.7), but the probability of teleportation to a node is now weighted proportional to its outgoing nodes. Then we use the node visiting frequencies (2.8a) in (2.8b) to compute the frequency of visiting an edge, denoted by $\mathbf{q}_{(j,i)}$. These edge visiting frequencies are consequently used again in (2.8c) to compute node visiting frequencies, by now just summing up the edge visiting probabilities of all incoming edges to a node $i$, thus not taking teleportation into account. This way, the visiting frequency of a node is only calculated iteratively through the visiting frequency of its neighbors.

With an expression for the ergodic visiting frequency of each node, we can compute the probability of exiting community $\alpha$ by using the elements of $\mathbf{P}(t_0,\tau)$:

$$q_{\alpha\curvearrowright} = \sigma\left(1 - \sum_{i\in\alpha}\frac{\sum_j \mathbf{P}(t_0,\tau)_{ij}}{\sum_{i,j}\mathbf{P}(t_0,\tau)_{ji}}\right)\sum_{i\in\alpha}\boldsymbol{\pi}_i + (1-\sigma)\sum_{i\in\alpha}\sum_{j\notin\alpha}\boldsymbol{\pi}_i\mathbf{P}(t_0,\tau)_{ij}. \tag{2.9}$$

The first term corresponds to the probability of teleporting to any node outside the current community, which is adjusted for the probability of teleporting into the current community. The second term corresponds to the probability of leaving the current community by following outgoing edges.

Through equations (2.5), (2.6) and (2.9) and using the ergodic visiting frequency of our nodes as found through equation (2.8), we can evaluate the map equation (2.4) without actually simulating the trajectories of any random walker. Instead, all we need are the steady state visiting frequencies and a transition matrix. The first can be calculated efficiently using a power iteration approach, while the second is provided. Therefore, $L(\mathsf{P})$ can be calculated efficiently for any given partition $\mathsf{P}$, and $L(\mathsf{P})$ can subsequently be minimized to find a good partition.

*Infomap* uses a stochastic and recursive heuristic algorithm to minimize the map equation. Its core algorithm roughly follows the following steps. Initially, each node is assigned its own community. Then, in random order, each node is moved to the neighboring community that would reduce $L$ the most, unless no move reduces $L$, in which case the node remains in its original community. It then applies this iteration recursively until no move results in a reduction of $L$. After that, *Infomap* is recursively applied on the resulting partition, now using the communities as nodes, until $L$ can be no longer reduced.

Multiple improvements have been added to this core algorithm of *Infomap* [61], including the detection of nested structures [63], overlapping communities [77], and notably, a method that introduces a tuning parameter to investigate community structures at different scales [67, 68, 37], referred to as the Markov-time. Remaining mindful of Lagrangian flow networks, we note that investigating the flow using nested community structures comes with a difficulty in interpretation, since it is not clear which nested communities should be expanded and which one should remain collapsed. Moreover, since we wish to use

*Infomap* to find well-defined barriers to transport, we will not consider the case of overlapping communities, since their boundaries are difficult to interpret in terms of barriers to transport. However, it is useful to have the possibility of tuning *Infomap* as to choose the spatial scale at which we investigate community structures in the network. This would allow us to choose a scale at which we can examine connectivity at a scale that is useful for investigating oceanographic structures.

The spatial scale of communities can be adjusted by changing the time it takes for a random walker to transition to another state. By default, a random walker changes states (or follows a self-loop to its current state) at each discrete timestep. Instead, the random walk can be considered through a continuous-time analogue, where the event that a random walker takes a step follows a Poisson distribution with the average time for transitioning denoted by $\mathsf{t_m}$ [37]. This Poisson process can be parameterized in discrete time by using an adjusted transition matrix

$$\tilde{\mathbf{P}}(t_0,\tau) = \begin{cases} (1-\mathsf{t_m})I + \mathsf{t_m}\mathbf{P}(t_0,\tau) & \mathsf{t_m} < 1 \\ \mathsf{t_m}\mathbf{P}(t_0,\tau) & \mathsf{t_m} \geq 1. \end{cases} \tag{2.10}$$

In essence, $\tilde{\mathbf{P}}(t_0,\tau)$ represents the lower probability of random walkers having not yet transitioned after a discrete timestep by adding self-edges. Conversely, a higher probability of taking a step is represented through higher transition probabilities. In the map equation (2.4), the Markov-time $\mathsf{t_m}$ does not influence the steady state node visit frequency, since the steady state visiting rates are independent of how often a state is sampled [37]. However, the Markov-time linearly scales the rate at which a random walker exits or enters a community, $q_{\alpha\curvearrowright}$, such that

$$q_{\alpha\curvearrowright}(\mathsf{t_m}) \equiv \mathsf{t_m} q_{\alpha\curvearrowright}. \tag{2.11}$$

Therefore, instead of actually using $\tilde{\mathbf{P}}(t_0,\tau)$, the Markov-time parameter can instead be introduced into the map equation by only considering equation (2.11) and the original transition matrix.

The effect of the Markov-time on community sizes can be interpreted as follows. When $\mathsf{t_m}$ is smaller than 1, random walkers are less likely to transition to a different node in one timestep. Going back to Huffman codes, this can be interpreted as a higher likelihood of the same node being encoded multiple times in the path description. Transitions between communities are therefore less likely, so the number of communities in the optimal encoding may be higher. Conversely, when the Markov-time parameter is larger than 1, a random walker may traverse multiple nodes before its position is encoded. If an optimal encoding should not include many transitions between communities, it should therefore allow for less communities in this case [37].

Throughout our description of *Infomap*, it is tempting to draw parallels between the fictional random walkers considered by *Infomap* and our 'physical' particles that traveled along the trajectories that gave rise to our transition matrix $\mathbf{P}(t_0,\tau)$). However, there are important differences [71]. For instance, the random walker keeps traversing nodes with probabilities that arise from the initial and final locations of our particles determined by the flow in our time window. While transition matrices have previously been used as a computationally inexpensive way to model the spread of tracer [36, 69], these studies use a succession of different transition matrices to capture the temporal variation of the flow field. Actual particles may follow different trajectories since the flow field is highly unsteady and transition probabilities are different when considering different values of $t_0$ and $\tau$. Using transition matrices to simulate flow also introduces artificial dispersion [44]. Furthermore, the connection to physical flow is further impaired by letting random walkers teleport.

With these remarks in mind, we have now given a description of how *Infomap* can be used to demarcate hydrodynamic provinces, corresponding to communities in the graph description of the flow. To efficiently find good solutions, *Infomap* works heuristically and stochastically: it yields a solution that has a short average code description length by using an optimization strategy in which the random order in which nodes are moved influences the topology of the resulting communities. This allows the algorithm to find different local minima of $L$. The use of a heuristic and stochastic approach has important implications. Due to the random moving of nodes in its core algorithm, different passes of *Infomap* may yield different locally optimal solutions. The algorithm may be run multiple times such that the partition $\mathsf{P}$ that yields the lowest value of $L(\mathsf{P})$ can be picked as a final solution. However, many degenerate solutions may exist, which all have similar values of $L$ while they may exhibit considerable topological differences [23, 6]. The transition matrices used by *Infomap* are by themselves already approximations of the real surface flow. This means that when one solution $\mathsf{P}_a$ has a slightly lower value of $L$ than another solution $\mathsf{P}_b$ while exhibiting a significantly different topology, there is no reason to assume that solution $\mathsf{P}_a$ carries more physical meaning than $\mathsf{P}_b$. Investigating the structure of only one solution might be misleading, especially when a community structure is weak [6]. While some communities are consistently found across different good solutions, others may not. Different solutions may be merged to find a consensus solution [74, 40], but in doing so, information on which community boundaries are weak may be lost. Instead, it is insightful to compare multiple solutions to see on which structures solutions agree and to figure out in which regions of the network the community structures are weaker [6].

### 2.2.2 Quality of Hydrodynamic Provinces

*Infomap*'s sole criterion for finding a good partition $\mathsf{P}$ is minimizing $L(\mathsf{P})$. The general criterion for community detection in network theory, namely finding groups that have few edges between each other while having many edges in the interior, can be translated into two criteria for hydrodynamic provinces. First, the ratio between Lagrangian particles leaving and staying in a hydrodynamic province within time $\tau$ should be low. This criterion interprets hydrodynamic provinces as almost-invariant areas of fluid, such that flow within a region $A$ is nearly mapped onto itself after time $\tau$: $\Phi_{t_0}^{\tau}(A) \approx A$ [71]. Second, hydrodynamic provinces should have strong internal mixing, making sure that different areas of each hydrodynamic province exchange fluid.

Ser-Giacomi et al. propose two quality parameters to assess the extent to which these criteria are met [71]. The first criterion is assessed through the *coherence ratio*, $\rho_{t_0}^{\tau}(\alpha)$, measuring the ratio between particles that leave and stay within a community $\alpha$ within time step $\tau$.

$$\rho_{t_0}^{\tau}(\alpha) = \frac{\sum_{i,j\in\alpha} m(B_i)\mathbf{P}(t_0,\tau)_{ij}}{\sum_{i\in\alpha} m(B_i)}. \tag{2.12}$$

For a partition $\mathsf{P}$ that divides the domain into $c$ communities $\alpha = 1, \dots, c$, the global coherence ratio is the average of the coherence ratio of each community:

$$\rho_{t_0}^{\tau}(\mathsf{P}) = \frac{1}{N_B}\sum_{\alpha=1}^{c} \#\{B_i | i \in \alpha\}\rho_{t_0}^{\tau}(\alpha). \tag{2.13}$$

Unlike in Ser-Giacomi et al. [71], here the global coherence ratio is weighted by the amount of bins in a community, such that we minimize the effect of small communities produced by noise in the data. The coherence ratio is determined only through $\mathbf{P}(t_0, \tau)_{ij}$, which is constructed from only the initial and final particle locations, meaning that particles may temporarily leave a community within the interval $[t_0, t_0 + \tau]$.

The second criterion is assessed using a measure of mixing proposed by Ser-Giacomi et al [71]. The *mixing parameter* indicates how strongly fluid within a community is mixed. To do so, only flow occurring within a community $\alpha$ is considered, which we can represent through a reduced transition matrix

$$\mathbf{R}(t_0, \tau|\alpha)_{ij} = \frac{\mathbf{P}(t_0, \tau)_{ij}}{\sum_{k\in\alpha} \mathbf{P}(t_0, \tau)_{ik}}, \ i, j \in \alpha. \tag{2.14}$$

The mixing parameter for a community $M_{t_0}^{\tau}(\alpha)$ is given by the normalized sum of the Shannon entropy associated to the transition probabilities between each pair of bins:

$$M_{t_0}^{\tau}(\alpha) = \frac{-\sum_{i,j\in\alpha} \mathbf{R}(t_0, \tau|\alpha)_{ij} \log \mathbf{R}(t_0, \tau|\alpha)_{ij}}{Q_\alpha \log Q_\alpha}, \tag{2.15}$$

with $Q_\alpha = \#\{B_i | i \in \alpha\}$. The mixing parameter reaches its maximum value of 1 when particles within a bin $B_i, i \in \alpha$ are dispersed uniformly to all other boxes in $\alpha$ ($\mathbf{R}_{ij} = \frac{1}{Q_\alpha}\ \forall i, j \in \alpha$). The global mixing parameter is then the weighted average of the mixing parameter

$$M_{t_0}^{\tau}(\mathsf{P}) = \frac{1}{N_B} \sum_{\alpha=1}^{c} Q_\alpha M_{t_0}^{\tau}(\alpha). \tag{2.16}$$

For practical applications such as investigating barriers to transport when planning MPAs, a third, qualitative, criterion may be added, namely that the communities found by *Infomap* take on spatial scales that are useful for identifying these barriers. The Markov-time parameter in *Infomap* allows us to change the spatial scale for investigation, and may thus be used to fulfill this criterion. This parameter is not considered by Ser-Giacomi et al. [71].

## 2.3 Oceanic structures in the Arctic domain

Lagrangian particles at the surface passively follow surface flow. In certain regions, this flow is dominated by persisting surface currents. Since the transport between two regions may be hindered by the presence of these currents, it is thus useful to provide a short description of the major currents in the Arctic ocean. Major seas in the Arctic are shown together with the bathymetry in figure S1, while an overview of Arctic surface currents is shown in figure S2.

On the side of Atlantic Ocean, inflow of warm and saline water is provided by the North Atlantic Current. The North Atlantic Current branches off into the Norwegian Current and the Irminger Current. The Norwegian Current, located west of Norway, closely follows the local topography due to the conservation of potential vorticity [66] and exhibits baroclinic instability [49]. It eventually splits into the West Spitsbergen Current and the North Cape Current, flowing into the Barents Sea. The Irminger Current flows along the western slope of Reykjanes Ridge separating the Irminger Basin from the Icelandic basin. It splits into two branches. The westward branch merges with the East Greenland Current, while the other branch forms the North Icelandic Irminger Current, flowing northward and later eastward around the coast of Iceland [41]. Inflow from the Pacific Ocean passes through the Bering Strait, which is both shallow ($\sim 50\,\text{m}$) and narrow ($\sim 85\,\text{km}$).

The main current through which water leaves the Arctic Ocean is through the East Greenland Current. The East Greenland Current is constrained to the Greenland Continental Margin due its low density water originating from sea ice and due to the conservation of potential vorticity [30]. Another region of outflow is the Canadian archipelago and the Davis Strait.

Surface velocities are also influenced by the presence of sea ice, which dampens the effect of wind stress exerted on the sea surface. When sea ice protrudes into the upper layer, it also provides a lateral barrier to flow. Within regions that contain sea ice, surface flow and sea ice drift patterns are dominated by the Beaufort Gyre, located in the Beaufort Sea, and the Transpolar Drift, which flows from the Laptev Sea and the East Siberian Sea to the Fram Strait. Sea ice drift and surface flow in the Beaufort Gyre exhibit anticyclonic motion, forced by the Beaufort Sea High atmospheric pressure system [58].

# 3 | Methods

Our methods for finding hydrodynamic provinces from a Lagrangian flow network closely follow the approach by Rossi et al. [60] and Ser-Giacomi et al. [71] as described in the previous chapter. This chapter describes the data, methods and configurations used in our experiments. First, we present the flow field data that is used, including an exact description of our domain. Then we illustrate how we discretize the domain and describe the implementation we used for determining Lagrangian particle trajectories. Finally, we report on our configuration of *Infomap*.

## 3.1 Hydrodynamical data

The connectivity between different geographical regions in the ocean is determined by the underlying hydrodynamics. So, in order to obtain a realistic description of connectivity, we need an accurate description of the hydrodynamics. The best available description is given by reanalysis data, in which fields of state variables are reconstructed through a synthesis of observations constrained by the physical governing equations [12].

Here, we use of the Global Ocean Physical Reanalysis product[1] [18], made available by the Copernicus Marine Environment Monitoring Service (CMEMS). This product provides reanalysis data for the global ocean at a resolution of 1/12°, corresponding to a latitudinal length of 9.3 km per grid cell. Data is provided for 50 vertical levels. The dataset contains daily mean fields over the period 1993-2018, for which ocean surface altimetry data and satellite sea ice data are available. The product includes fields describing fluid dynamics (horizontal velocities), thermodynamics (salinity and temperature) and sea ice features (concentration, thickness and horizontal velocities). These quantities are retrieved by assimilating model output from the NEMO 3.1 ocean model [50] and the LIM2 EVP sea ice model [25] with observational data. While values in NEMO an LIM2 are computed on a tripolar Arakawa C-grid, final fields are interpolated on a regular Arakawa A-grid. Atmospheric forcings are provided with 3-hourly and 24-hourly frequencies by the ERA-interim dataset [14] provided by the European Centre for Medium-Range Weather Forecasts.

For assessing ocean surface connectivity, we are specifically interested in surface velocities. To describe the ocean surface, we only use the uppermost layer of the dataset, which is situated at a depth of 0.49 m. The velocity field is not divergence-free due to vertical motions, but as discussed, since we use teleportation dynamics as in PageRank, this causes no issue for computing the steady-state visiting frequency of each node in the corresponding network. In order to investigate the effect of sea ice on the topology of hydrodynamic provinces, we also use the sea ice concentration fields that are provided in the dataset. The sea ice concentration in each grid cell is defined as the ratio of the cell's area that is covered in sea ice. Sea ice thickness and sea ice velocities are not considered, since

[1] `GLOBAL_REANALYSIS_PHY_001_030`

the presence of sea ice in our upper layer is implicitly incorporated in our velocity fields. The daily resolution and multidecadal timespan of the dataset enables us to investigate how seasonality and decadal trends influence the topology, coherence and mixing of hydrodynamic provinces. Moreover, the temporal resolution and extent allows us to investigate the persistence of features over time.

An extensive assessment of the quality of the dataset can be found in [16]. In summary, the main ocean currents and sea ice extent variability are reproduced well. However, winter sea ice extent maxima are overestimated. Spring and summer are marked by an excess of ice melting, while in winter, the spread of ice extent is reduced when compared to observations. Root mean square differences and biases compared to observations are stable over the entire period of dataset.

An intercomparison study of different reanalyses, including a coarser predecessor of our dataset (GLORYS2V4 at 1/4°), shows an inter-reanalysis agreement on sea ice concentration, which can be expected since different reanalysis experience the same constraints in surface temperature from atmospheric forcing, and from direct assimilation of sea ice concentration observations [8]. We therefore conclude that our dataset is the best available approximation for sea ice concentrations.

A limitation of the dataset is that while it is eddy-permitting, it is not necessarily eddy-resolving. A meridional resolution of 1/12° corresponds to an effective resolution of 9.3 km. However, the first baroclinic Rossby radius of deformation, which is the natural scale of baroclinic boundary currents, eddies and fronts takes values between 1 and 16 km in the Arctic ocean, sometimes even assuming values below 1 km in shallow seas like the Barents Sea [53]. Therefore, the reanalysis data does not resolve eddies, fronts and boundary currents of these scales in certain regions of the Arctic. Instead, it resolves larger-scale or aggregate structures larger than our grid size. Moreover, we should be careful with our interpretation of the zonal resolution increasing northward due to convergence of meridians at the poles. Since values calculated on a tripolar C-grid are interpolated onto an A-grid, a resolution increase on the interpolated A-grid does not correspond to smaller scale structures being resolved any better. In fact, closer to the pole, many grid cell values may be interpolated from just a few cells on the C-grid.

## 3.2 Spatial domain

Although the coverage of our hydrodynamical data is global, we limit ourselves to studying the Arctic domain. We define it as the area above 60°N and only load hydrodynamical data within this domain. The domain is indicated in figure 3.1. An inherent effect of having a domain with an open boundary is a loss of connectivity information. There is a possibility that at a timescale $\tau$, two geographic regions within our domain may exchange fluid parcels through currents or eddies that (partially) fall outside of the domain. The loss of information is dependent on the timescale $\tau$ at which the trajectories of a fluid parcels are investigated, and therefore on the distance from the boundary of our domain: the further a parcel is from the boundary, the less probable it is to reach the boundary within our integration time $\tau$, so information is lost less likely.

In particular, our domain choice causes the Denmark Strait and Davis Strait to be disconnected, since the southernmost tip of Greenland lies outside the domain, which can be seen in figure 3.1. The East Greenland Current flows around this tip at Cape Farewell where it transitions into the weaker West Greenland Current (see figure S2). This is a clear example of a location where information loss occurs: trajectories between the east and west of Greenland cannot be resolved, so although these areas may be connected by

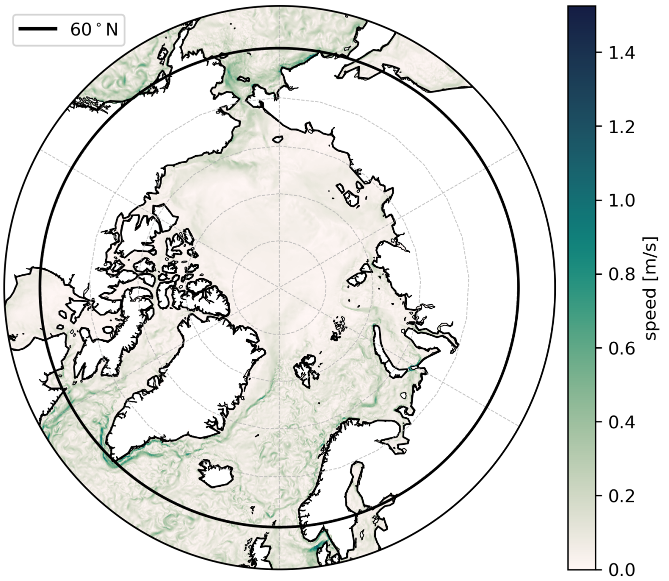


**Figure 3.1:** Map of our spatial domain, which lies north of 60°N (solid line). The map includes a snapshot of mean surface speeds at January 1st, 2018. In our experiments, velocity fields are only loaded above 60°N, while here they are also shown below 60°N. Currents and eddies may advect particles outside of the domain.

flow, this is not be visible in our Lagrangian flow network.

In section 4.3, we try to assess how the loss of information due to our open domain boundaries influences the communities that *Infomap* finds. This information loss can be prevented by including hydrodynamical data south of our domain, which we could easily do since our dataset has global coverage. However, this would come with an increase in computational cost in case we wish to study the ocean surface at the same resolution. Trivially, loading velocity fields of the global ocean takes much more time than loading just the velocity fields around the Arctic. One could suggest to load only a small portion of data outside our domain which may allow particles to return. However, how large this portion should be cannot be determined a priori. Generally, if we keep following Lagrangian particles outside of the domain, they may spend a lot of time there without ever returning. Computing trajectories outside of our domain may thus be computationally wasteful and we avoid doing so.

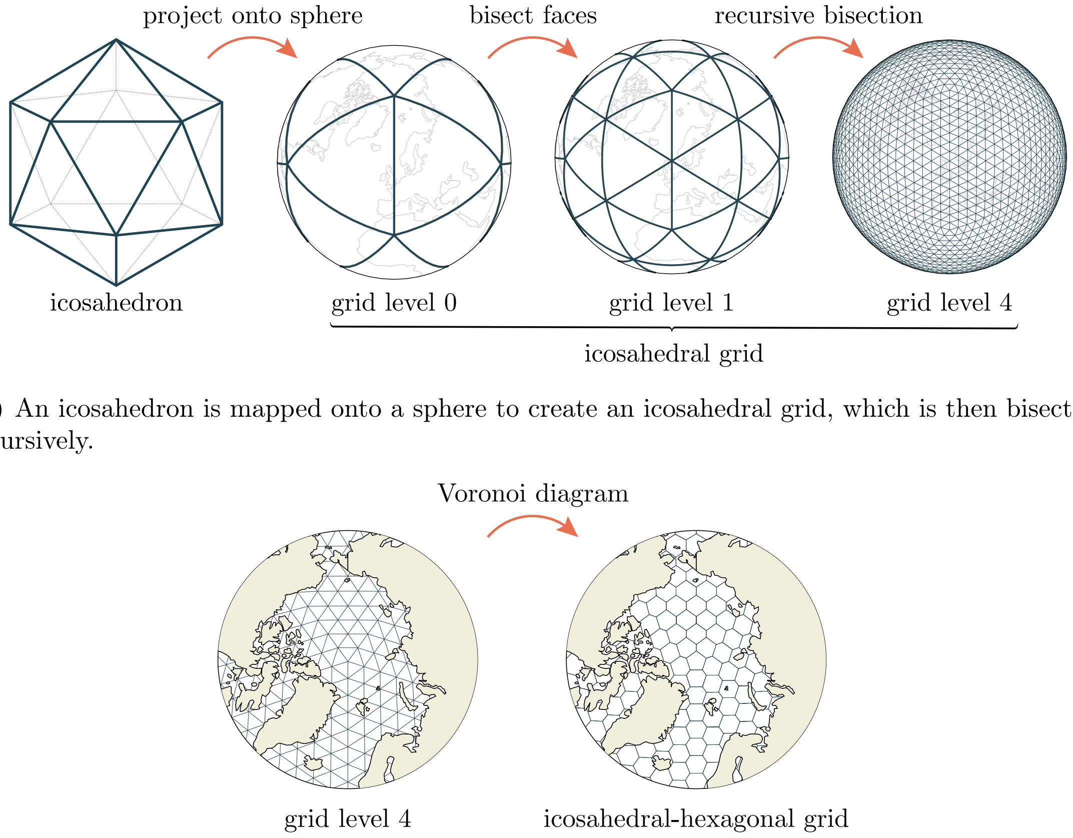
## 3.3 Domain discretization

The domain needs a discrete description in order for the hydrodynamics to be mapped onto a network. The domain is discretized by dividing it into bins, which will correspond to the network's nodes. In principle, a regular grid would provide a straightforward rectangular binning, but it comes with a problem. Due to the convergence of meridians at the poles, these bins would get smaller with increasing latitude. Especially in the polar regions, this would cause bins to have drastically varying areas. Ideally, our course-grained description of the flow considers exchange between regions of comparable size and shape. This way, transition probabilities between different areas can be compared straightforwardly, without accounting for bin sizes. If we would use regular grids, probabilities to reach individual bins would become smaller as bin areas become smaller closer to the pole.

To circumvent curvature-related issues, we make use of a *icosahedral-hexagonal grid.*

This class of grids is composed of a tessellation of hexagons and 12 pentagons, which are all of similar size and shape. This allows for easier comparison of particle exchange between boxes. This grid has been used in several geophysical fluid modeling studies due to its desirable isotropy [65, 78].

There exist multiple implementations of icosahedral-hexagonal grids that are different in construction. We use a widely-used construction method that goes as follows. First, we map the vertices of an icosahedron onto a sphere. Then, we bisect each triangular face into four new triangles. We do this recursively until a specific level of refinement has been reached. The number of recursive bisections is referred to as the *grid level*. This *icosahedral grid* with triangular faces is constructed using the `Stripy` Python package (version 1.0.2) [48]. Then, we make use of the Voronoi diagram of this grid. A Voronoi diagram partitions a metric space $X$ containing a collection of $k$ generating points $S$ into $k$ separate regions $R$, such that each region is defined as the set of points for which the distance of point $x$ to generating point $S_i$, $d(x, S_i)$, is not greater than the distance to any other generating point $S_{j \neq i}$. Expressed mathematically: $R_i = \{x \in X | d(x, S_i) \leq d(x, S_j)\ \forall j \neq i\}$. The Voronoi diagram of a tessellation of equilateral triangles corresponds to a tessellation of regular hexagons. The Voronoi diagram of an icosahedral grid corresponds to the icosahedral-hexagonal grid, with each Voronoi region providing us with hexagonal (or pentagonal) bins. This grid construction procedure is visualized in figure 3.2.



**(a)** An icosahedron is mapped onto a sphere to create an icosahedral grid, which is then bisected recursively.

**(b)** The Voronoi diagram of this triangular tessellation provides an icosahedral-hexagonal grid.

**Figure 3.2:** Procedure for constructing an icosahedral-hexagonal grid.

Although the icosahedron that we used to generate this grid consists of equilateral faces, it is impossible to bisect the projected faces on the sphere into equilateral spherical triangles [78]. As a consequence, our icosahedral-hexagonal grid is not made up of perfect spherical hexagons. However, the maximum ratio in distances between neighboring grid points is bounded, and so is the ratio of largest and smallest areas of hexagons in the Voronoi diagram [78]. For grid level 7, this ratio is 1.36. At this refinement level, the average area of a hexagonal bin in the Voronoi diagram is 3113 km$^2$, while the average distance between adjacent bin centers is 60.16 km (0.54°).

Using an icosahedral-hexagonal grid for our domain discretization has two advantages. Firstly, bins mostly have a comparable area. While we could also construct equal-area bins for example by using a sinusoidal projection [60], this causes bins to become stretched out as we approach the poles. Instead, our grid contains bins of similar shapes. Secondly, while rectangular bins in regular grids have diagonal neighbors that share no edge with each other, adjacent bins in the icosahedral-hexagonal grid always share an edge. Therefore, particles leaving a bin always spend some time in directly neighboring bins.

## 3.4 Particle simulation

Particle simulations are carried out using the *Parcels* Lagrangian framework (version 2.1.2) [15]. The *Parcels* framework provides an accessible Python interface to Lagrangian ocean analysis. It takes advantage of C-compiled code to efficiently integrate particle trajectories on user-provided velocity fields. *Parcels* includes field interpolation schemes to interpolate particle velocities in space and time. We set it to use a fourth-order Runge-Kutta integration scheme for determining particle trajectories. We note that *Parcels* interpolates the hydrodynamical data supplied on a regular A-grid. The icosahedral-hexagonal grid is not used for the particle simulations, but only for subsequent binning.

Even though we consider passive buoyant particles, *Parcels* allows us to specify particle behavior, which is useful for setting the boundary conditions at our open boundary and at the coast. At the open boundary, we freeze particles that reach latitudes below 60°N. Therefore, in our transition matrix, transport to regions outside the domain is represented through the bins that lie at the boundary. This is useful to determine the connectivity of the regions of exit with respect to the rest of the domain.

Particles may also get stuck as they get pushed towards cells where their speed becomes zero. This can be the case at land cells, where the meridional and zonal velocity fields become zero. Particles can reach these cells since the velocity fields do not have impermeable boundary conditions at the coast. Although we could specify a boundary condition where particles reaching the coast are sent back into the ocean domain, methods to do so are ambiguous. Rossi et al. [60] and Ser-Giacomi et al. [71] remove these stuck particles. Instead, we keep these stuck particles in $\mathbf{P}(t_0, \tau)$, and interpret this as the beaching of buoyant particles.

In representing flow through a transition matrix by using particle simulations, several factors need to be balanced. For instance, the domain needs to be discretized into bins with high enough resolution to resolve physical structures while providing a statistical description of the flow. To this end, we should also initialize a large enough number of particles per bin to capture the flow statistically. These factors both influence the total number of particles that needs to be simulated, which determines the simulation time. Two other factors that influence the simulation time are the total advection time $\tau$ and the advection timestep $\Delta t$.

We choose a domain discretization using an icosahedral-hexagonal grid at grid level 7 and initialize our particles on the vertices of the triangles of the icosahedral grid at grid level 11. Particles that lie on land are removed. Bins that contain no land therefore initially contain between 253 and 258 particles, the slight variation being due to irregularities in the grid. The number of initial particles in bins that contain land may be much lower. Figure 3.3 shows the initial distribution of particles per bin that is used throughout our experiments. In total, 1450665 particles are initialized.

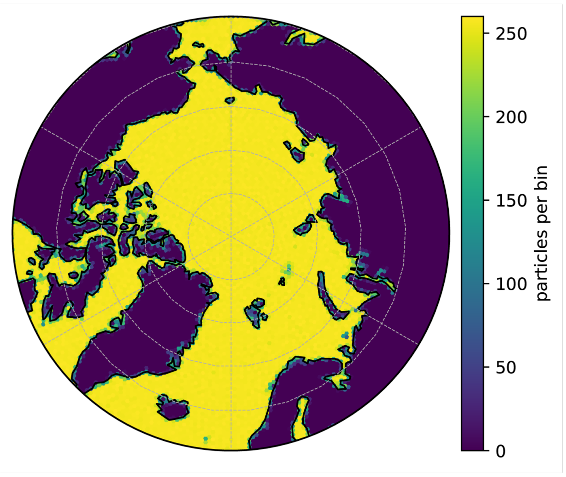


**Figure 3.3:** Initial distribution of particles when initializing particles on the vertices of the icosahedral grid at grid level 7, counted in bins on the icosahedral-hexagonal grid at grid level 7.

To investigate connectivity at different timescales, we simulate particle trajectories for an advection time of 90 days. Particle locations are stored daily, such that connectivity at intermediate timescales can also be assessed. Specifically, we look at $\tau = 30$ and 90 days.

We choose an advection timestep $\Delta t$ of 20 minutes. When comparing the locations of particles released at the same location, but advected with a timestep of 1 minute, the average euclidean distance after 30 days is of the order 3 km. Therefore, we assume that using this advection timestep, we are able to resolve trajectories to a high degree of accuracy.

Particle simulations are carried out on a Dell PowerEdge R730 machine, equipped with 2 16-core Intel Xeon CPU E5-2683 v4 processors running at 2.10 GHz with hyper-threading enabled to support 64 threads. Each simulation has access to 23.242 GB of virtual memory. The machine runs Scientific Linux V7.3. With this configuration, typical particle simulation (wall clock) times for 1450665 particles advected for $\tau = 90$ days with a timestep of $\Delta t = 20$ minutes are 3.5 to 5.5 hours.

We release particles at March 1st and September 1st, such that seasonal conditions correspond to high and low sea ice extent respectively. We carry out these simulations for each year between the period 1993 and 2018, which allows us to find trends and patterns of connectivity that persist over years. For the year 2017, we carry out simulations at the start of each month, in order to investigate seasonal effects.

## 3.5 Matrix and graph construction

The initial and final locations of the simulated particles are used to construct transition matrices $\mathbf{P}(t_0, t_0 + \tau)$ and their corresponding network description, as described in section 2.1. While the bin sizes and number of initialized particles in our domain discretization vary, these variations are normalized when constructing the transition matrix.

Using the icosahedral-hexagonal grid for a discretization into bins, the domain contains 6614 bins that (partially) contain fluid, such that $\mathbf{P}(t_0, t_0 + \tau)$ is a square matrix with dimensions $6614 \times 6614$.

We can efficiently determine the initial and final bin of a particle trajectory by considering which generating point $S_i$ in the icosahedral grid is nearest. By definition this point corresponds to the containing Voronoi bin $R_i$. We efficiently determine the nearest generating point by using a k-d tree lookup. A k-d tree is a data structure that for a given coordinate allows us to efficiently look up the nearest point in a predefined set of points. We construct a k-d tree using the points generating the Voronoi tessellation. For this, we use *SciPy*'s [33] `spatial.cKDTree` implementation of the algorithm described by Maneewongvatana and Mount [42]. This way, we can compute the containing bins of all particles in the order of a few seconds. This then allows us to determine $\mathbf{P}(t_0, \tau)$ using equation (2.2).

## 3.6 Community detection using Infomap

Finally, we obtain a division of our network into clusters by using *Infomap* (version `1.0.0-beta.51`). We configure *Infomap* to take into account the characteristics of our flow network.

To start with, we specify that the network should be interpreted as a directed network. In order for the steady-state visiting frequencies to be determined, we use the standard value for the teleportation probability of $\sigma = 0.15$. By making use of the *unrecorded teleportation* scheme, solutions are robust in the regime $\sigma \in (0.05, 0.95)$ [39]. For lower values, the steady state visiting frequency $\boldsymbol{\pi}$ becomes unstable, while for higher values, the steady state approaches the weights of each link.

In addition, we make sure that self-edges, which point from a node $i$ to itself, are included. In fact, $\mathbf{P}(t_0, \tau)$ often has values on the diagonal, meaning that particles stay within a bin after timestep $\tau$, making self-edges an indispensable part of our flow description.

Furthermore, we only consider a two-level community description, meaning that we do not consider nested communities. As discussed, in a nested description, it can be difficult to assess which communities should be expanded, making it hard to compare structures. Instead, we let *Infomap* only return one layer of communities, which are partitioned as to minimize the map equation.

Lastly, different experiments are carried out to determine a Markov-time parameter $\mathsf{t_m}$ that produces communities of a convenient spatial scale. This scale should be large enough to allow for comparison of solutions between different seasons and such that we may attempt retrieve oceanographically relevant structures.

In all experiments, we run *Infomap* 20 times, after which the partition $\mathsf{P}$ with minimum $L(\mathsf{P})$ is saved. This ensures that partitions are at a high quality, while keeping computation times reasonable. Contrary to what Ser-Giacomi et al. [71] report for the Mediterranean, solutions do not converge by running Infomap more often. Instead, the value for $L$ convergences only in the coarse- and fine-tuning steps that *Infomap* executes in one run. The corresponding solution does depend on a random order in which nodes are moved by the algorithm, such that different runs do not produce the same result. A higher quality partition may be found by running *Infomap* more often, which may sometimes yield a better value of $L(\mathsf{P})$, but we later show that further improvements in $L(\mathsf{P})$ have only small influences on the global coherence ratio and global mixing parameter.

*Infomap* is run on a MacBook Pro (late 2013), equipped with one Dual-Core Intel Core i5 processor running at 2.4 GHz, with hyper-threading enabled to support 4 threads. The machine has a memory of 8 GB and runs MacOS 10.15. With this configuration, one outer loop of *Infomap*, which is run 20 times, takes approximately 1.5 s. In all experiments, the same set of random seeds is used (using option `-s 314159`), except when we explicitly aim to compare differences in solutions due to *Infomap*'s stochastic nature (varying the seed between 1 and 100).

In general, solutions are compared from multiple perspectives. First, the topology is assessed with respect to oceanographic features as well as the persistence of boundaries among different solutions corresponding to different time intervals. Furthermore, we assess whether our criteria for coherence and mixing are met, as defined in section 2.2.2. Lastly, we compare values of $L(\mathsf{P})$ through the map equation (2.4). This way, solutions are assessed from an oceanographic and information theoretical perspective, and the connection between these two perspectives is evaluated.

# 4 | Experiments & Results

Due to the large temporal extent of our dataset and the various possible configurations possible with *Infomap*, there are many dimensions through which we can study connectivity in the Arctic through the methods explained in chapter 3. For example, to investigate the system at different temporal and spatial scales, we can carry out analyses for a range of values of $\tau$ and Markov-times. Also, since 26 years of data are available, in principle we can run *Infomap* over many different transition matrices. However, while aiming to provide a comprehensive overview of the different features that govern the topology of hydrodynamic provinces, we wish to avoid a lengthy and excessive treatment of each variable that could potentially be at play. This motivates the following structure for this chapter: first, we compare solutions for one transition matrix obtained with different Markov-times and we choose one value to carry out all other analyses with. Then, we aim to assess to what extent solution degeneracy introduces variations in community topology and our quality parameters among different solutions for four transition matrices, corresponding to March and September 2018 and different time scales. Subsequently, we assess the effect of having an open boundary in our domain. Having assessed these effects, we investigate the persistence of community boundaries over time. Lastly, we examine connections between community structures, sea surface velocities and sea ice and investigate temporal trends and seasonal cycles. Throughout each section, topologies are discussed in the context of physical structures.

Throughout this chapter, maps with hydrodynamic provinces returned by *Infomap* are colored with arbitrary colors in such a way that two neighboring communities never share the same color. However, since *Infomap* does not have any information on how the network is embedded in space, communities may exhibit enclaves, meaning that different parts of the same community may not be connected in space. Due to limitations in visualization, these enclaves are not explicitly indicated in the figures in this chapter.

## 4.1 Varying Markov-time

As discussed, we can investigate hydrodynamic provinces at different spatial scales by tuning the Markov-time parameter $\mathsf{t_m}$. However, it is a priori not clear what spatial scale should be used. For the sake of consistency, we wish to continue further analyses with just one value of $\mathsf{t_m}$. This value should yield solutions with communities at a spatial scale that is convenient for analysis for different values of $\tau$. A specific definition of a good spatial scale depends on the specific application. We only set two broad criteria. On the one hand, solutions should contain communities that are not too small or thin, such that we can later easily assess the persistence of bins that border on other communities. If communities are only a few bins wide, many bins will be marked as boundary bins, and it will be difficult to assess the persistence of community boundaries. On the other hand, communities should not be too large such that they span tens of latitudes or longitudes, containing many features that are known to function as physical barriers to transport, since this would be at odds with our aim of finding communities with boundaries that correspond to barriers to transport themselves. Both criteria should hold for the range of time intervals $\tau = 30$ to 90 days.

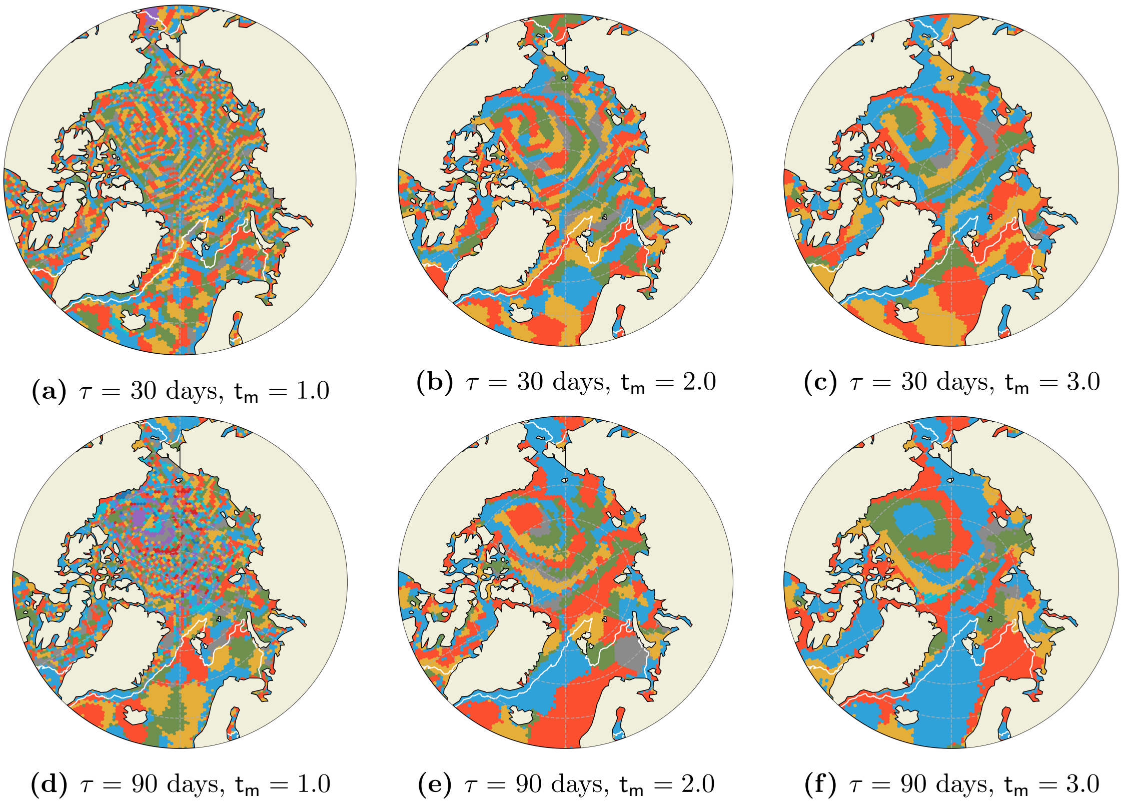

(a) $\tau = 30$ days, $\mathsf{t_m} = 1.0$
(b) $\tau = 30$ days, $\mathsf{t_m} = 2.0$
(c) $\tau = 30$ days, $\mathsf{t_m} = 3.0$
(d) $\tau = 90$ days, $\mathsf{t_m} = 1.0$
(e) $\tau = 90$ days, $\mathsf{t_m} = 2.0$
(f) $\tau = 90$ days, $\mathsf{t_m} = 3.0$

**Figure 4.1:** Comparison of solutions returned by Infomap for different values $\mathsf{t_m}$ for $\tau = 30$, 90 days. Particles are initialized on $t_0$ = 2018-03-01. White contours indicate average the sea ice extent in March 2018, defined as the contour line corresponding to a sea ice concentration of 15%.

Figure 4.1 shows a comparison of solutions returned by *Infomap* for different values of $\mathsf{t_m}$ and $\tau$. All solutions exhibit a circular configuration of hydrodynamic provinces around the Beaufort Gyre. For $\tau = 30$ days, solutions contain filamental structures around the East Greenland Current.

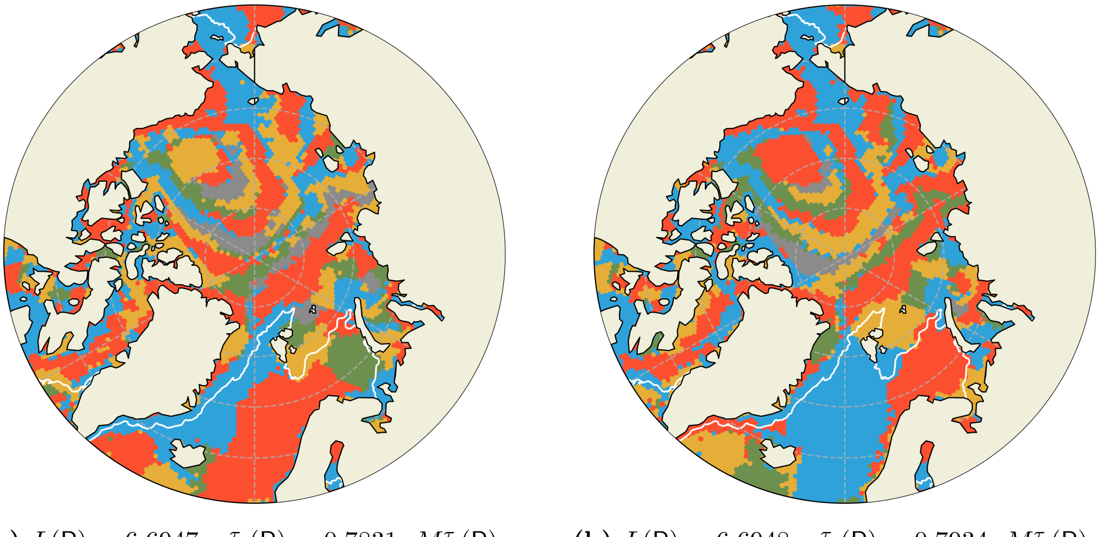
Community sizes increase with an increase in $\tau$ or $\mathsf{t_m}$. This is to be expected, since an increase in $\tau$ allows Lagrangian particles to travel farther, thus connecting more bins. Moreover, due to the chaotic nature of ocean flow, particles originating from the same bin may over time follow different currents and eddies, such that when longer time spans are considered, the spread in particle distributions becomes larger. From a network perspective, this decreases the average distance between nodes and a random walker on the network can therefore traverse larger distances, connecting bins that are separated by larger distances too. For increasing $\mathsf{t_m}$, random walkers may also traverse more nodes before having their position recorded. Therefore, it makes sense that *Infomap* draws community boundaries at larger distances as either $\tau$ or $\mathsf{t_m}$ increases.

For $\tau = 30$ days and $\mathsf{t_m} = 1.0$, the solution consists of many small communities that have sizes that are too small to assess boundary persistence. For this $\mathsf{t_m}$ and both values of $\tau$, communities are especially small in the presence of sea ice. This makes sense, as the surface velocities in these areas are drastically lower. In these areas, some communities also consist of many spatially disconnected bins. Communities are largest for $\tau = 90$ days and $\mathsf{t}_m = 3.0$. For this solution, one hydrodynamic province spans from the edge of the domain at 60° between Iceland and Norway all the way to the sea ice boundary and protrudes far into the sea ice. We deem the value $\mathsf{t_m} = 3.0$ too high, since surface velocities at the sea ice boundary drop drastically and it should therefore provide a natural boundary in the system. We choose to continue with $\mathsf{t_m} = 2.0$, since the solutions for both values of $\tau$ exhibit a community scale that fits both of our criteria.

## 4.2 Individual solutions and solution degeneracy

As discussed in chapter 2, *Infomap* is a stochastic and heuristic algorithm suffering from solution degeneracy, meaning that *Infomap* returns locally optimal solutions, which each might exhibit different topologies. In order to use *Infomap* to study oceanographic structures, the persistence of boundaries, and temporal trends, we must first evaluate the role that solution degeneracy plays. We do this by running *Infomap* on the same transition matrix with the exact same parameters, only varying the random seed. We compare results for 100 different seeds in terms of codelength, global coherence, global mixing and boundary persistence. Differences in results must then be due to the degeneracy of *Infomap*.

For 100 solutions obtained for $\mathbf{P}(t_0 = \text{2018-03-01}, \tau = 90 \text{ days})$, the average codelength is $L(\mathsf{P}) = 6.694$, while the associated standard deviation is 0.011 (from now on reported in parentheses). The average global coherence ratio is $\rho^{\tau}_{t_0}(\mathsf{P}) = 0.7843$ ($\pm 0.0067$), while the average global mixing parameter is $M^{\tau}_{t_0}(\mathsf{P}) = 0.3392$ ($\pm 0.0023$).



**(a)** $L(\mathsf{P}) = 6.6947$, $\rho^{\tau}_{t_0}(\mathsf{P}) = 0.7831$, $M^{\tau}_{t_0}(\mathsf{P}) = 0.3401$.

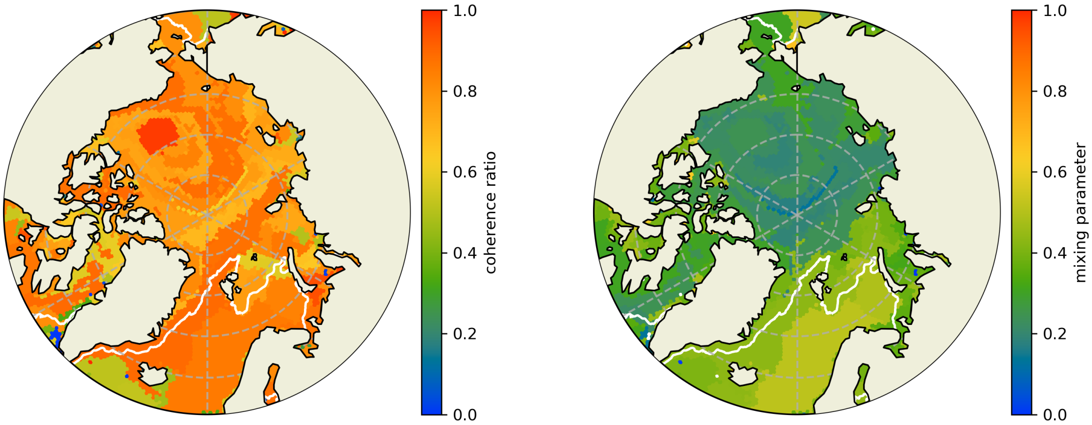
**(b)** $L(\mathsf{P}) = 6.6948$, $\rho^{\tau}_{t_0}(\mathsf{P}) = 0.7934$, $M^{\tau}_{t_0}(\mathsf{P}) = 0.3346$.

**Figure 4.2:** Two solutions found for $\mathbf{P}(t_0 = \text{2018-03-01}, \tau = 90 \text{ days})$. White contours indicate average sea ice extent in March 2018.

Figure 4.2 shows the two solutions that have their codelengths closest to the average value. Both solutions are good solutions since they have been obtained by running *Infomap* for a different seed 20 times and picking the partition with lowest codelength. The codelengths of the two solutions differ by 0.0001. Both solutions exhibit similar topologies in the Davis Strait, the Beaufort Gyre, and the Chukchi Sea. However, for certain areas, topologies are very different. This can clearly be seen in the Norwegian Sea. The solution in figure 4.2a separates the Norwegian Sea from the Greenland Sea, while the solution in 4.2b clusters these seas together.

Coherence ratios and mixing parameters exhibit spatial patterns. This can be seen in figure 4.3, which shows the coherence ratio and mixing parameter associated to each community of the partition depicted in figure 4.2a. Communities that lie close to the boundary of the domain can exhibit low coherence ratio since particles may exit the domain from these communities. Bins where particles exit the domain are often clustered as single communities. In contrast, the community at the center of the Beaufort Gyre shows a coherence ratio close to 1, meaning this community retains almost all particles that were released there.

Mixing parameters are generally higher for communities in ice-free regions, likely due to higher velocities and the presence of eddies that can stir Lagrangian particles across a community. Values are especially high in the Norwegian sea, which contains the Norwegian Current which exhibits baroclinic instability [49]. This area is characterized by high mesoscale activity [29]. The solution depicted in figure 4.2b exhibits similar patterns in coherence and mixing. This can be seen in figure S3.



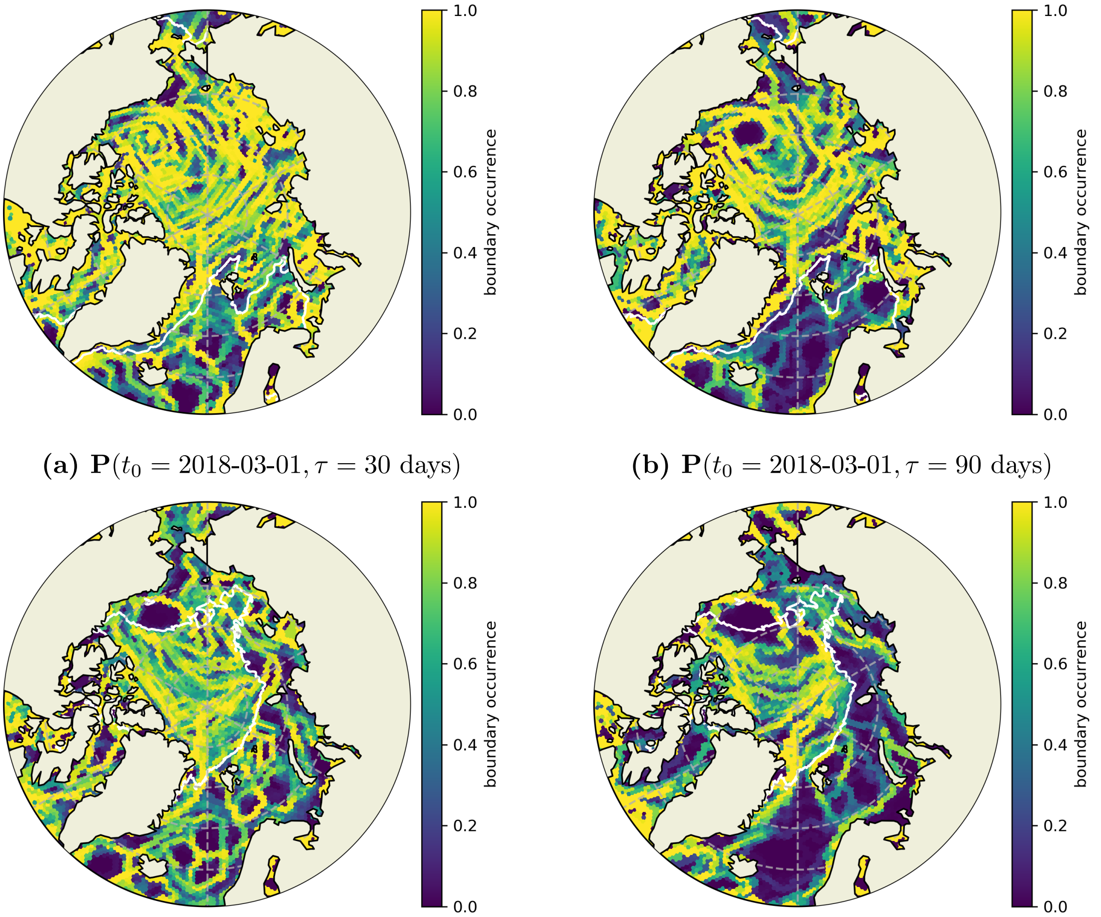
**(a)** Coherence ratio

**(b)** Mixing parameter

**Figure 4.3:** The coherence ratio and mixing parameter associated to each community in the partition depicted in 4.2a, for $\mathbf{P}(t_0 = \text{2018-03-01}, \tau = 90 \text{ days})$.

Ser-Giacomi et al. investigate the persistence of community boundaries across years [71]. We follow a similar procedure to assess the persistence at which *Infomap* draws boundary locations. We flag the bins that lie at the interface between two communities as a boundary bin and assess the frequency at which each bin is marked as such among our 100 solutions. This is shown in figure 4.4 for March and September 2018, with $\tau = 30$ and 90 days.

In certain regions, *Infomap* is able to draw boundaries persistently. This is mainly the case in areas containing sea ice. The circular structure around the anticyclonic Beaufort Gyre can clearly be seen. Boundaries are also often found separating the Irminger Basin and Iceland Basin from each other and the rest of the domain. These basins are physically separated by the Irminger Current, coinciding with the Reykjanes Ridge. A boundary also persists at the edge of the continental shelf east of Greenland, where the East Greenland Current is located. For $\tau = 30$ days, small communities persist in the Norwegian sea, but locations differ between March and September. Especially for $t_0 =$ September 1st, boundary-free regions can clearly be seen. For $\tau = 30$ days, the Norwegian sea contains ring-like boundaries. For $\tau = 90$ days *Infomap* is less persistent in drawing boundaries in



**(a)** $\mathbf{P}(t_0 = \text{2018-03-01}, \tau = 30 \text{ days})$ **(b)** $\mathbf{P}(t_0 = \text{2018-03-01}, \tau = 90 \text{ days})$

**(c)** $\mathbf{P}(t_0 = \text{2018-09-01}, \tau = 30 \text{ days})$ **(d)** $\mathbf{P}(t_0 = \text{2018-09-01}, \tau = 90 \text{ days})$

**Figure 4.4:** Persistence of community boundaries in a set of 100 solutions found for different transition matrices. White contours indicate average sea ice extent in March and September 2018 respectively.

and between the Norwegian Sea and Greenland Sea. For this $\tau$, a frontier is visible running from the coast of the Scandinavian peninsula to Novaya Zemlya, isolating the White Sea and its outflow.

Plotting the persistence of boundaries yields information on where *Infomap* is consistent in assigning neighboring nodes to different communities. Instead of investigating a single solution to identify barriers to transport, an ensemble of solutions can effectively extract these barriers from the given flow field.

Flagging boundaries does not necessarily show in which regions *Infomap* is inconsistent in assigning nodes to the same communities. For a given region, low values for the occurrence of a boundary in a specific bin can be due to no boundary being drawn at all in this region across different solutions, but it may also be due to boundaries being drawn at different locations in each partition. In the latter case, this means that two points in a region may be disconnected across solutions, but we would not be able to readily infer this from looking at the boundary persistence in figure 4.4.

Multiple methods exist to quantify the similarity between two community partitions [13, 45, 34]. These methods generally compare the global similarity of solutions, taking into account the topology of *all* communities. Instead, to assess how persistent the topology of

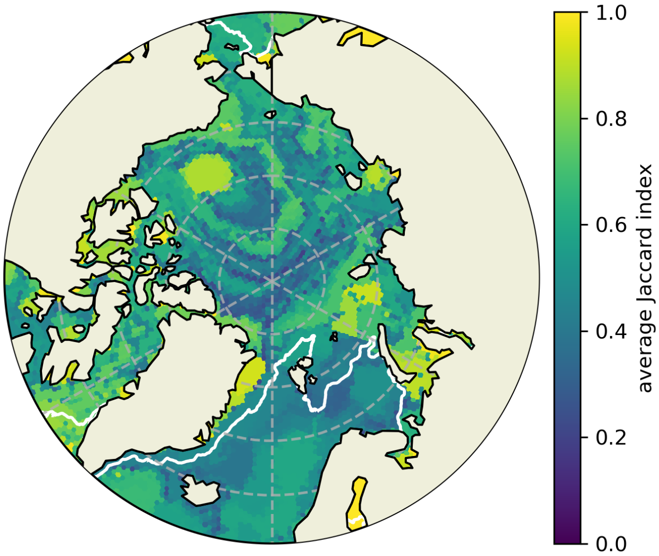


**Figure 4.5:** Average Jaccard index among solution pairs per bin as given by equation (4.2). Solutions correspond to $\mathbf{P}(t_0 = 2018\text{-}03\text{-}01, \tau = 90 \text{ days})$. White contours indicate average the sea ice extent in March 2018.

a single community is across solutions, we can investigate local differences by assessing for a given bin how often it is clustered together with the same group of bins. One method to compare similarity in node assignment is by considering the *Jaccard index*, also referred to as *Jaccard distance*. The Jaccard index of two sets $A$ and $B$ is the size of their intersection over the size of their union:

$$J(A, B) = \frac{\#\{A \cap B\}}{\#\{A \cup B\}}. \tag{4.1}$$

When investigating degeneracy, Calatayud et al. [6] advocate to investigate the solution landscape by grouping similar solutions. Their algorithm to do so is based on finding minimum Jaccard distances between different communities. Since individual solutions may each differ from one another in a different area, we refrain from clustering complete partitions in terms of similarity. However, we can still make use of the Jaccard index to assess the persistence in node assignment. For each node, we can for each of the $\binom{100}{2}$ pairs of solutions determine the Jaccard distance between the communities that a node is clustered in. Let $\kappa$ denote a set of $K$ different of partitions, $\kappa = \{\mathsf{P}_a | a = 1, \ldots, K\}$. Let $\alpha_i^a$ indicate the community that bin $i$ falls under in solution $\mathsf{P}_a$. Then for each node, the persistence of being assigned to similar communities in an ensemble of solutions is the average Jaccard index among solution pairs, $A_i$, as given by

$$A_i(\kappa) = \frac{2}{K(K-1)} \sum_{a=1}^{K-1} \sum_{b=a+1}^{K} \frac{\#\{B_j | j \in \alpha_i^a \wedge j \in \alpha_i^b\}}{\#\{B_j | j \in \alpha_i^a \vee j \in \alpha_i^b\}}. \tag{4.2}$$

In regions where $A_i$ is high, *Infomap* is consistent in the topology of communities, while in regions where $A_i$ is low, *Infomap* is inconsistent in assigning a node to the same community.

Figure 4.5 shows the average Jaccard distance for $\mathbf{P}(t_0 = 2018\text{-}03\text{-}01, \tau = 90 \text{ days})$ as defined by $A_i$ in equation (4.2) for our set of 100 solutions. Here we see that community assignment is particularly persistent on the northeast of Greenland, the Kara Sea, the Baltic Sea and the center of the Beaufort Gyre. In other regions, assignments are less consistent, such that it is important to always look at more than one solution when assessing connectivity in those regions.

## 4.3 Sensitivity to domain boundary

As discussed in section 3.2, having a domain with an open boundary may entail losing information about connectivity between bins. Since particle trajectories are frozen as soon as a particle exits the domain, the choice of the domain boundary is of influence on the transition probability of a particle. We attempt to assess the extent to which our community boundaries are influenced by the latitude at which we define our domain. The codelength of a partition yielded by *Infomap* depends on the transition probabilities between bin pairs. Changes in transition probabilities due to shifting the domain boundary should be localized to bins that lie close to the boundary. Therefore, we expect the influence of our domain choice on the locations of boundaries locations to be localized to the domain boundary.

We investigate the influence of the location of the domain boundary by comparing the boundary persistence for different domains, but with simulations obtained for the same values of $t_0$ and $\tau$. We choose $t_0$ = 2018-09-01 and $\tau$ = 90 days, such that communities are large enough for the boundary persistence not to be too noisy and to reduce the effect of sea ice. Ideally, we would compare to the case where we do not have an open domain at all, which can be achieved by expanding the domain to the global ocean. However, as discussed in section 3.2, this would come with increased computational costs, partly due to loading significantly more hydrodynamical data and also due to the extra computation of trajectories of particles below 60°N. Instead, we compare our normal domain bounded at 60°N, for which the boundary persistence is found in figure 4.4d, to a smaller domain bounded at 70°N. Since particle trajectories are obtained deterministically by using equation (2.1), trajectories of particles that do not reach latitudes lower than 70°N are the same as when considering a domain bounded by 60°N. Large parts of the resulting transition matrix should thus be equal to that of 60°N. We apply *Infomap* 100 times on the transition matrix obtained for the simulation in the modified domain, which for 67% of originating bins (columns in the transition matrix) is completely equal to the transition matrix in the original domain. The persistence of boundaries in these 100 new solutions is shown in figure 4.6.

*Infomap* may either draw community boundaries in the two domains at different locations due to the underlying transition matrices being different, or it may do so due to its heuristic and stochastic nature. Figure 4.7 shows the difference in persistence of solutions obtained from the 60°N and 70°N domains. Here, bins are left white if this difference is lower than the standard deviation associated to flagging a bin as a community boundary across the 100 solutions obtained using the 70°N domain. This allows us to see where differences in community boundaries are likely due to solution degeneracy and where they may be due to the different domain choice. Note that the differences in boundaries are significant mostly near the domain boundary. A notable exception is in the Greenland Sea, where this difference comprises a large portion of bins. However, differences in the persistence of community boundaries between the two domains become smaller closer to the interior of the domain. We theorize that when comparing communities in the 60°N domain to communities obtained from transition matrices for the global ocean, differences in boundary persistence should similarly be localized to our current domain boundary at 60°N. We thus assume that boundary persistences as found in figure 4.4 are mostly the same as when instead we would have considered the global ocean.

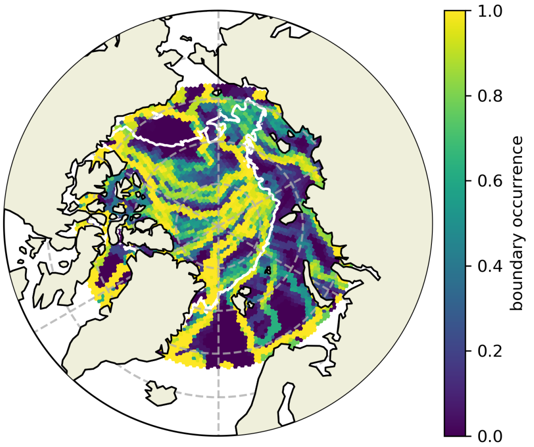


**Figure 4.6:** Persistence of community boundaries for 100 solutions obtained for $\mathbf{P}(t_0 = \text{2018-09-01}, \tau = 90 \text{ days})$, with the redefined domain being bounded by 70°N. White contours indicate average sea ice extent in September 2018.

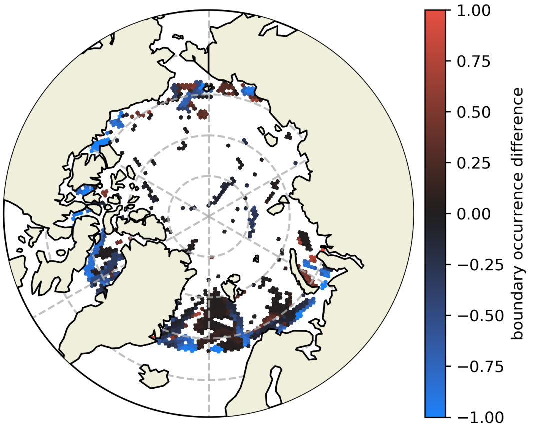


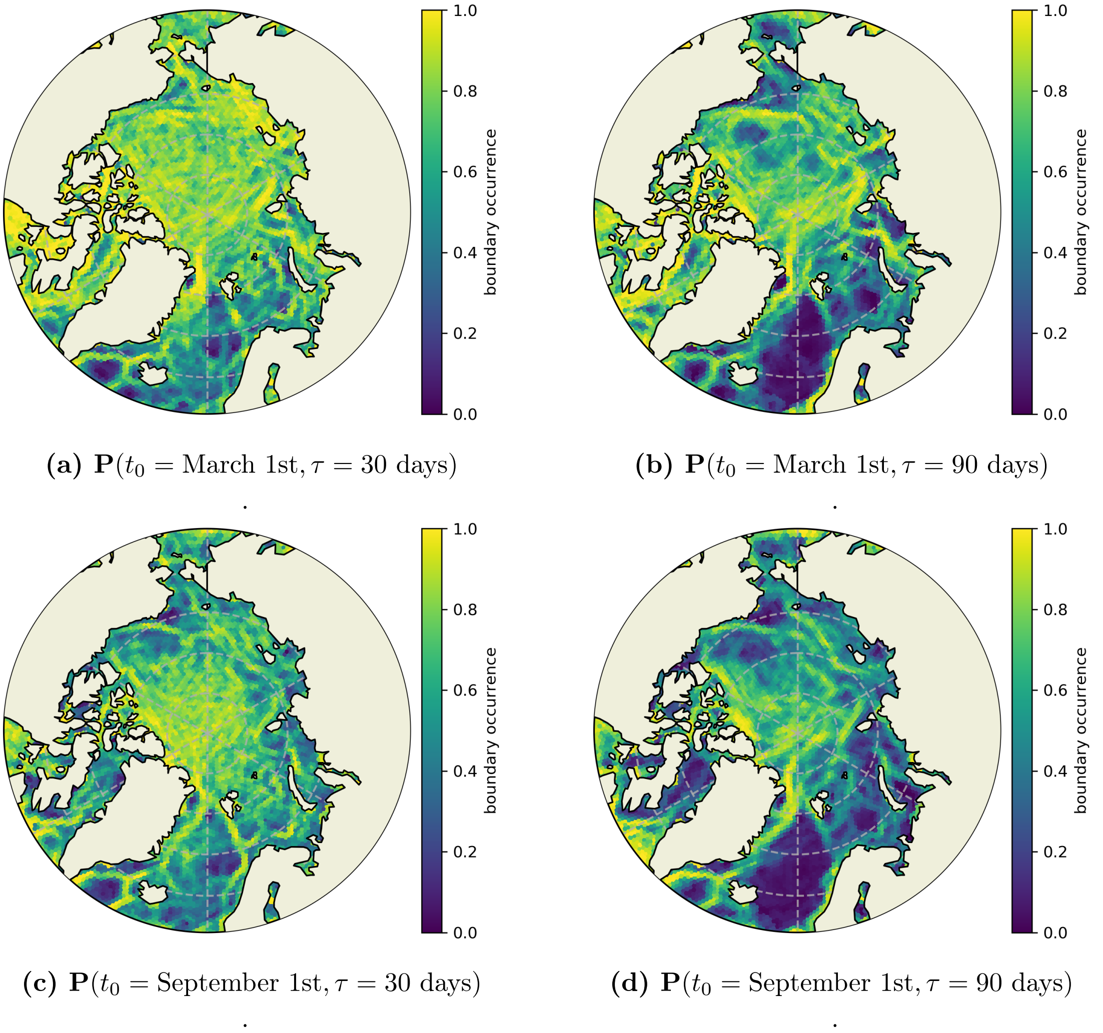
**Figure 4.7:** Difference in community boundary persistence, obtained by subtracting the persistence in the 70°N domain in figure 4.6 from the persistence in the 60°N domain in figure 4.4d. Differences are only shown when they exceed the standard deviation associated to flagging a bin as a boundary across the 100 solutions obtained using the 70°N domain (calculated for each bin).

## 4.4 Barriers to transport

Having assessed the solution degeneracy and the effect of using an open boundary, it is insightful to assess the persistence of boundaries between hydrodynamic provinces over different years. Like Rossi et al. [60] and Ser-Giacomi et al. [71], we take the average of the community boundaries of solutions obtained from transition matrices corresponding to different years and seasons. One difference is that here we take solution degeneracy into account by including 100 solutions for each transition matrix.

Figure 4.8 shows the persistence of boundaries averaged over the years 2009-2018 for March and September 2018, with $\tau = 30$ and 90 days. Each subfigure is thus composed using 10 transition matrices, corresponding to each year, and for each transition matrix, 100 solutions are obtained using *Infomap*. This way, boundaries that are due to degeneracy or natural variability are filtered out.

Across all solutions, we again observe a circular structure around the Beaufort Gyre. For particles released in September, the East Greenland Current is also persistently visible. The North Atlantic Current, including the Norwegian Current and West Spitsbergen Current are prominently visible for particles initialized in September with $\tau = 30$ days, and to a lesser extent for the solutions with $\tau = 90$ days. The Irminger Current persists across solutions. For $\tau = 90$ days, the solutions for particles initialized in March show a boundary at the north-eastern coast of Iceland, while such a boundary is not visible for particles initialized in September. This may be due to the seasonality in the strength of the North Icelandic Irminger Current [41]. For $\tau = 30$ days, boundaries occur more often in the Norwegian Sea and Greenland sea than for $\tau = 90$ days, illustrating that different physical structures provide boundaries to transport at different time scales.



**(a)** $\mathbf{P}(t_0 = \text{March 1st}, \tau = 30 \text{ days})$
.

**(b)** $\mathbf{P}(t_0 = \text{March 1st}, \tau = 90 \text{ days})$
.

**(c)** $\mathbf{P}(t_0 = \text{September 1st}, \tau = 30 \text{ days})$
.

**(d)** $\mathbf{P}(t_0 = \text{September 1st}, \tau = 90 \text{ days})$
.

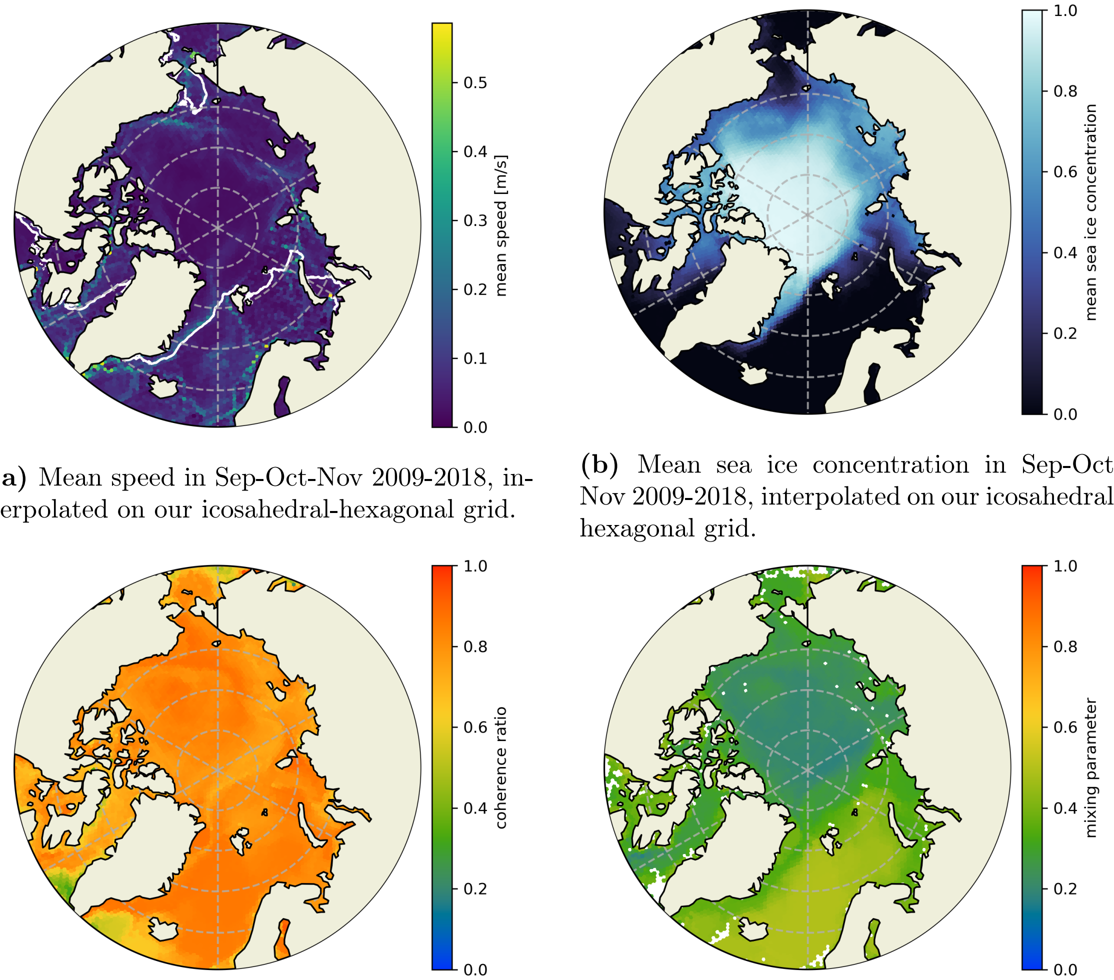
**Figure 4.8:** Persistence of community boundaries between 2009 and 2018. For each transition matrix, 100 different solutions are obtained using *Infomap*.

## 4.5 Connections to sea surface velocities and sea ice concentrations

To better understand how *Infomap* is affected by the physics that give rise to our transition matrices, we investigate and attempt to explain correlations between codelengths, community boundaries, coherence, mixing, surface velocities and sea ice concentrations. Of these quantities, codelengths and community boundaries pertain to the graph description of our physical system, while surface velocities and sea ice concentrations are inherent to the physical fields that govern the dynamics of Lagrangian particles. Coherence and mixing are dependent on both the community division found by *Infomap*, as well as the physical trajectories of Lagrangian particles, thus bridging the physical- and graph descriptions of our system. We note that sea ice concentrations are not explicitly used to determine particle trajectories, but the presence of sea ice does influence the surface velocity field and therefore implicitly affects the system.

To investigate the correlation between codelengths, global coherence ratios and global mixing parameters, we use a set of 1000 solutions, which comprises of 100 solutions for

10 transition matrices obtained through simulations between 2009 and 2018 with $t_0 =$ September 1st and $\tau = 90$ days. For these global quantities, we find that the codelength and global mixing parameter have a negative Pearson correlation coefficient of $r = -0.093$, with an associated $p$-value of $p = 3.1 \times 10^{-3}$. Simultaneously, we find a positive correlation between codelength and global coherence, with $r = 0.18$ and an associated $p$-value of $p = 1.4 \times 10^{-8}$. This means that as *Infomap* finds better partitions with smaller codelengths, this will on average slightly increase the global mixing, while it will on average slightly decrease the global coherence.



**(a)** Mean speed in Sep-Oct-Nov 2009-2018, interpolated on our icosahedral-hexagonal grid.

**(b)** Mean sea ice concentration in Sep-Oct-Nov 2009-2018, interpolated on our icosahedral-hexagonal grid.

**(c)** Average coherence ratio per bin.

**(d)** Average mixing parameter per bin.

**Figure 4.9:** Average speed, sea ice concentration, coherence ratio and mixing parameter for each bin. Coherence ratio and mixing parameter correspond to average values of the communities a bin is partitioned with in each of the 1000 solutions for $\mathbf{P}(t_0 = \text{September 1st}, \tau = 30 \text{ days})$ between 2009-2018.

To locally assess correlations between the velocity, sea ice, coherence, mixing, and boundary persistence of each bin, we make use of the same 1000 solutions for the transition matrices $\mathbf{P}(t_0 = \text{September 1st}, \tau = 90 \text{ days})$ between 2009-2018. We use transition matrices with particles released in September, since a larger portion of the domain is ice-free. The corresponding boundary persistence can be found in figure 4.8, while the other quantities can be found in figure 4.9.

Meridional and zonal velocities are averaged per grid cell over September, October and

November 2018, interpolated onto the icosahedral-hexagonal grid, and converted into mean speed. Sea ice concentrations are similarly averaged and interpolated.

We find a positive correlation between mean speed and boundary persistence, with a Pearson correlation coefficient of $r = 0.38$ and $p = 0$ (below machine precision) in regions where the sea ice concentration is less than 0.15. If we include all bins, this statistically significant correlation vanishes. This indicates that in ice-free regions, currents correlate to community boundaries and thus provide barriers to transport. This correlation disappears in the presence of sea ice, meaning that in the sea ice regime, other factors govern the existence of boundaries.

When comparing correlations with sea ice, we find a positive correlation between coherence ratio and sea ice concentration of $r = 0.21$ and $p = 6.3 \times 10^{-69}$. On visual inspection of the average coherence ratio in figure 4.9c, it is difficult to see a direct relation to sea ice. The correlation between sea ice and coherence ratio may be biased due to the low coherence ratio of some communities at the edges of the domain, where particles may escape to communities containing only a few bins.

The mixing parameter and sea ice concentration are negatively correlated with $r = -0.24$ and $p = 2.6 \times 10^{-88}$. From figure 4.9d we can observe that mixing is generally higher in ice-free regions. However, the mixing parameter is also low around the East Greenland Current. This makes sense, since this current flows south, such that within communities in this region, it is only possible for particles to spread to bins that lie south. In contrast, mixing is strong in the Norwegian Sea and Barents Sea. The high mesoscale activity in the Norwegian Sea may provide relatively efficient mixing in the communities located there.

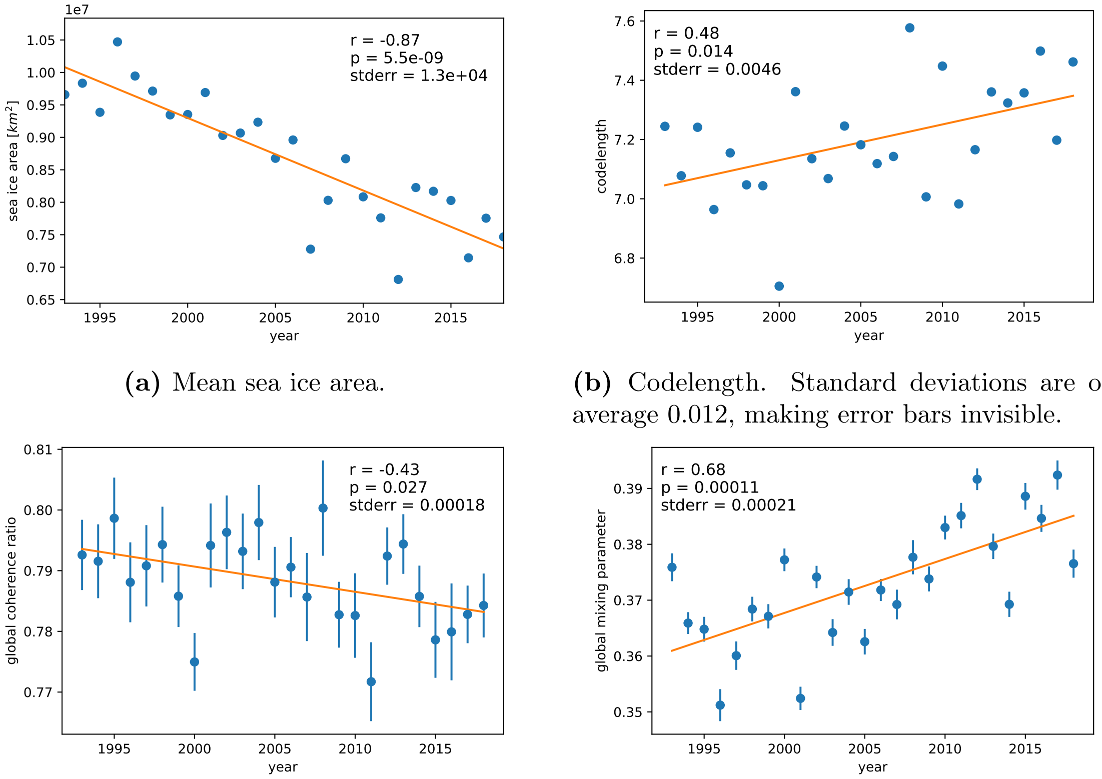
## 4.6 Trends and Seasonality

Since the sea ice extent in the Arctic experiences a strong seasonal variation with implications for ocean surface flow, we expect the quality of solutions to be affected by this. Furthermore, the decrease in summer sea ice extent that has been observed in the past few decades is also of influence on surface flow, thus it also affects solutions found by *Infomap*. It is insightful to assess these effects by looking at the monthly and yearly temporal evolution of solution quality.

Figure 4.10 shows the evolution of sea ice, codelength, global coherence ratio and global mixing parameter calculated from 100 degenerate solutions obtained for $\tau = 90$ days and $t_0 =$ September 1st for each year between 1993 and 2018. Linear trend regressions are included. For codelengths, the standard deviation associated to the solution degeneracy is much smaller than the differences in average codelengths between different years, and is thus hardly visible. This indicates that most differences among solutions cannot be attributed to solution degeneracy, but are instead due to differences in the underlying flow, mirrored in the transition matrices. The global mixing parameter also exhibits standard deviations that are smaller than the variation of mean values between years. In contrast, for the global coherence ratio, the standard deviations due to solution degeneracy are of the order of the variation of mean values between years.

We find a negative correlation between sea ice area and codelength, with $r = -0.44$ and $p = 0.023$. The correlation between sea ice and the global mixing parameter is also negative, with $r = -0.75$ and $p = 9.4 \times 10^{-6}$. We do not find a significant correlation between sea ice and the global coherence ratio.

While the sea ice area exhibits a clear downward trend, the codelength and global mixing parameter show positive trends. The coherence ratio shows a slight downward



**(a)** Mean sea ice area.

**(b)** Codelength. Standard deviations are on average 0.012, making error bars invisible.

**(c)** Global coherence ratio.

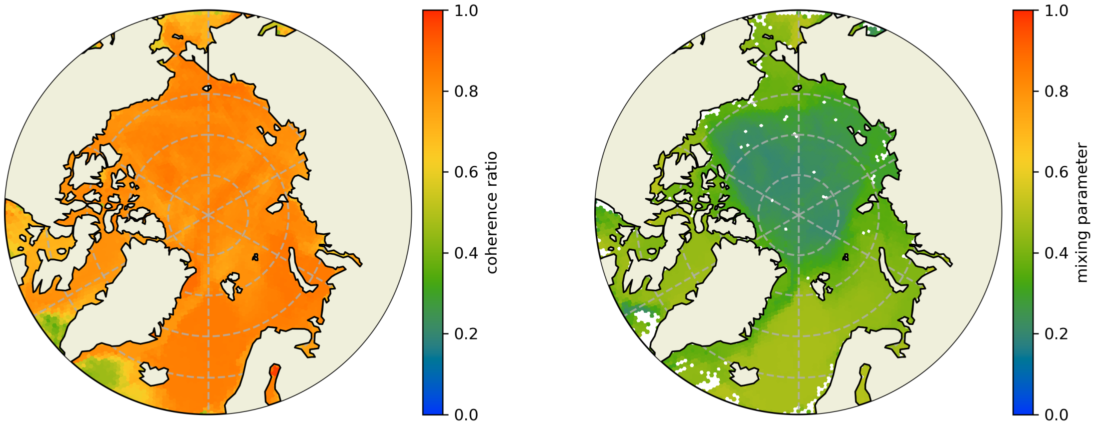
**(d)** Global mixing parameter.

**Figure 4.10:** Evolution of sea ice, codelength, global coherence ratio and global mixing parameter in September between 1993 and 2018. 100 solutions have been obtained for each year by initializing particles on the first day of September 1st. $\tau = 90$ days. Trend regression is indicated in orange, including associated correlation coefficient $r$, $p$-value and standard error. Standard deviations related to solution degeneracy are indicated using error bars, except for sea ice area.

trend, although with a higher $p$-value than the codelength, sea ice area and mixing.

To supplement the inspection of solution quality in different years, figure 4.11 shows average values of the coherence ratio and mixing parameter for 100 solutions in 1993-2002 for particles initialized at $t_0 =$ September 1st in the respective years, with $\tau = 90$ days. We can compare this to the average coherence ratio and mixing parameter between 2009-2018, as shown in figures 4.9c and 4.9d. These time spans correspond to the first and last ten years of our dataset. No clear topological changes can be seen for the average coherence ratios. In contrast, we see that for the mixing parameter, the region in the middle of the domain that exhibits low values is shrinking. We theorize this to be due to the decreasing summer sea ice extent, such that the ocean surface may exhibit larger velocities and provide more efficient mixing.

Seasonal development of the sea ice area, codelength, global coherence ratio and global mixing parameter are assessed by comparing 100 solutions for 12 transition matrices, for which $t_0$ equals the first day of each month in 2017, while $\tau = 90$ days. Figure 4.12 shows the monthly evolution of these parameters. For the codelength and mixing parameter, a clear seasonal cycle can be observed, with maxima in summer and minima in winter, which coincides with the seasonal cycle in sea ice area. Indeed, we find a negative correlation between sea ice area and codelength of $r = -0.85$ with $p = 4.0 \times 10^{-4}$ and a negative correlation between sea ice area and the global mixing parameter of $r = -0.85$ and $p = 5.4 \times 10^{-4}$. A seasonal cycle is absent for the global coherence ratio and we do not find a



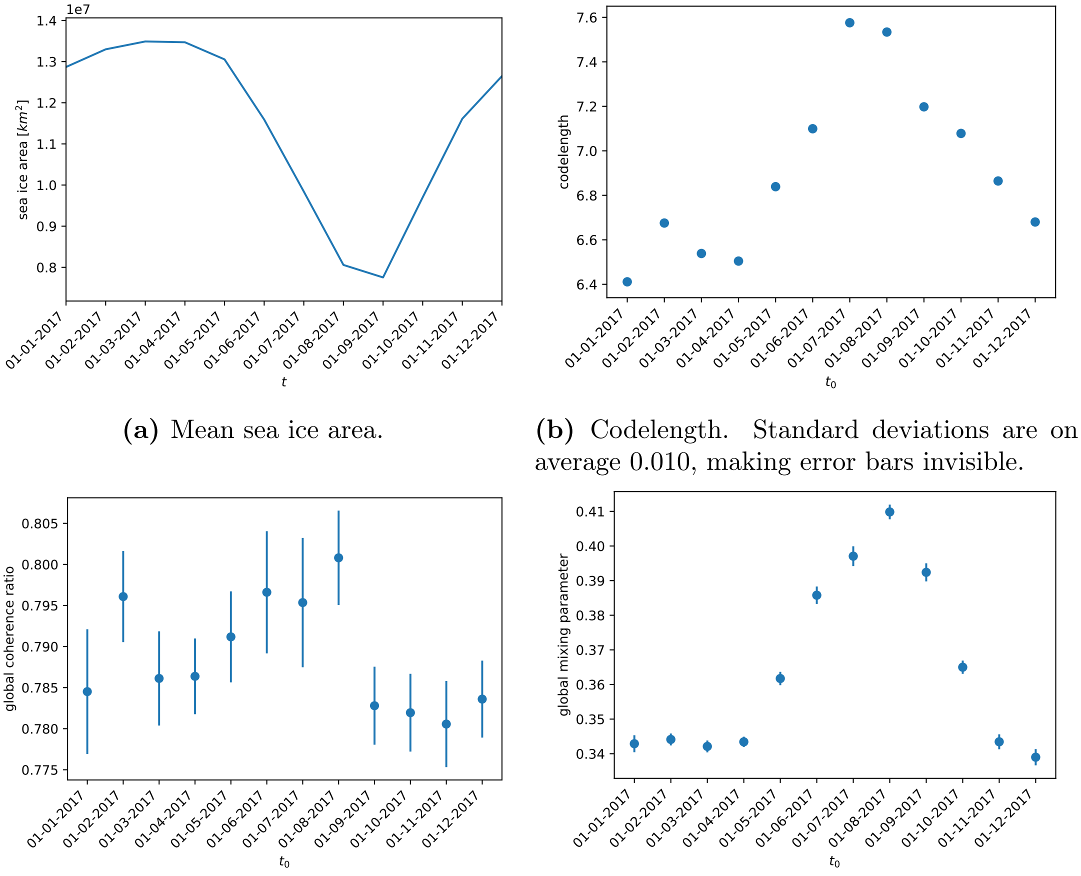
**(a)** Coherence ratio for 1993-2002.

**(b)** Mixing parameter for 1993-2002.

**Figure 4.11:** Average coherence ratio and mixing parameter for each bin from 1000 solutions for $\mathbf{P}(t_0 = \text{September 1st}, \tau = 30 \text{ days})$ between 1993-2002, similar to figure 4.9c and 4.9d.

correlation between sea ice and coherence ratio here.

Since sea ice coverage is minimal in summer, the seasonal cycle of the codelength agrees with the yearly trend of increasing codelength as sea ice declines. This also corresponds to what can be observed for the mixing parameter, which is lower in areas with sea ice and which globally increases over time, as sea ice cover decreases. This can be interpreted as follows: as the sea ice extent declines, surface velocities increase. Therefore, particles can travel larger distances, which increases the connectivity between bins. This in turn reduces the distance between edges in the network. This allows a random walker to traverse the network more easily. Nodes that were previously visited infrequently, increase in steady-state visiting frequency, making the distribution of $\boldsymbol{\pi}$ more balanced. This causes the average codelength to increase, as infrequently visited nodes that are assigned larger codewords are visited relatively more frequently. Simultaneously, in areas covered by ice, mixing is lower than average. As sea ice disappears and velocities increase, these regions become more mixed.



**(a)** Mean sea ice area.

**(b)** Codelength. Standard deviations are on average 0.010, making error bars invisible.

**(c)** Global coherence ratio.

**(d)** Global mixing parameter.

**Figure 4.12:** Monthly evolution of the sea ice area, codelength, coherence ratio and mixing in 2017. 100 solutions have been obtained for each month by initializing particles on the first day of each month. $\tau = 90$ days. Error bars indicate the standard deviation due to solution degeneracy.

# 5 | Discussion

In this chapter, we discuss to what extent *Infomap* is successful at identifying clear boundaries to transport. We also present several limitations to our methods and experiments, related to the data and simulations, as well as limitations that are inherent to *Infomap*'s community detection strategy. Additionally, we raise different caveats that should be heeded when applying *Infomap* to Lagrangian flow networks in practical settings, such as planning Marine Protected Areas. Lastly, we point out differences of this study with respect to earlier work and provide suggestions for future work.

## 5.1 Resolving physical structures from Lagrangian flow networks

In the context of hydrodynamic provinces, each individual community partition returned by *Infomap* corresponds to a local minimum of the map equation. The standard deviation of the codelength due to solution degeneracy is low compared to the differences in codelength due to seasonal and yearly variations in flow. Therefore, each individual solution is good in the sense that it corresponds to a low average codelength and the corresponding community partition should have boundaries to transport such that the transitions between different communities are locally minimized. When investigating individual degenerate solutions, such as in figure 4.2, different communities in the sea ice free domain can be seen to correspond to different seas. Boundaries have been shown to correlate with velocities, meaning that currents provide effective barriers to cross-community particle exchange. However, different partitions may each resolve different physically relevant boundaries. To obtain a more complete picture, it is thus useful to consider an ensemble of solutions, such as is done in figure 4.4. We note that these solutions only arise from a single transition matrix. Seasonal pictures can be obtained by considering ensembles over different years [71].

The boundaries in regions with sea ice seem to arise from the flow slowly moving in a anticyclonic fashion around the Beaufort Gyre, coinciding with the flow of sea ice. Particles thus move in concert, and hydrodynamic provinces in this area do not satisfy the criterion of high internal mixing well. Sea ice cover is shown to be anticorrelated with codelength and global mixing, both seasonally and yearly.

For individual partitions, we note that it is always important to consider the mixing parameter and coherence ratio of the corresponding communities. By only looking at the topology of a community, it may be tempting to think that any two bins that fall under the same community exchange particles with one another. In contrast, in certain communities, the underlying flow may have one clear direction, such that the corresponding nodes in the network are not strongly connected. This is the case for example in communities that coincide with strong currents, such as the East Greenland Current. In these communities, particles generally only travel south, following the flow. This manifests itself in a relatively low mixing, making the inspection of the mixing parameter an important step of the analysis.

*Infomap* may provide a useful lens to investigate connectivity and barriers to transport in a large domain. Nonetheless, different mechanisms may give rise to each community and their barriers. Different experiments may be necessary to better understand the nature of each community and community barrier.

## 5.2 Limitations and caveats

The representation of true flow in terms of Lagrangian flow networks is limited by several factors. First, the hydrodynamic data and domain discretization suffer from a finite resolution. This has the consequence that mesoscale structures are not fully resolved in the flow field in certain regions of the domain. Additionally, where mesoscale structures are resolved in the flow, they are only represented statistically in the graph description of the flow. The representation can thus be improved by using a higher resolution flow field that is eddy resolving and by increasing the resolution of the domain discretization.

Additionally, since there is a limited amount of trajectories originating from each bin, representation can be improved by increasing the number of Lagrangian particles that is simulated. However, this bears extra computational costs, especially when the grid resolution is increased.

Furthermore, the representation of the flow is influenced by having an open boundary. When choosing an open domain, it is important to assess to which extent this influences community topologies. For our domain, this effect seems to be localized to the domain boundary, and *Infomap* is still able to find barriers to transport that carry physical significance. Nevertheless, this limitation can only be fully overcome by considering flow in the global ocean.

Our results are also sensitive to multiple parameters. Naturally, the communities returned by *Infomap* are dependent on $t_0$ and $\tau$, since these parameters govern the time and timescale at which the flow is recorded into a transition matrix. In addition, results are sensitive to the choice of the Markov-time parameter. Here we choose one specific Markov-time parameter to tune *Infomap* in such a way that it provides a convenient spatial scale for our analyses. For more specific applications, it may be difficult to assess which Markov-time should be considered, since it is impossible to know a priori which Markov-time corresponds to each specific spatial scale at which the flow can be considered.

When using community detection in practical contexts, such as planning Marine Protected Areas, it is important to take the aforementioned limitations in mind. Here we considered the community detection algorithm *Infomap* due to successful previous applications and since its underlying algorithm optimizes a balance between high internal connectivity and good coherence, while emphasizing the flow description of a network. However, the way in which *Infomap* balances coherence and mixing cannot be set explicitly.

Furthermore, the degeneracy of solutions makes the interpretation of a single solutions misleading. If the optimal solution with the minimum average codelength could be found, it should in principle yield a partition which optimizes our criteria for strong internal mixing and good community coherence. However, this solution would have been obtained through cumulative approximations of the flow, for example arising from uncertainties in the observed flow fields, limited spatial and temporal resolutions, a limited amount of modeled Lagrangian trajectories, and a freedom in parameter choice. Therefore, there is no reason to assume that the optimal solution corresponds to a division in communities that carries the most physical meaning. This gives further motivation to always consider an ensemble of solutions. It is useful to assess the uncertainty in node assignment by using equation (4.2). Even when an ensemble of solutions is considered, community detection can be useful, but should be interpreted with caution due to the propagation of uncertainties, approximations and errors.

Specifically in the context of planning Marine Protected Areas, it would be erroneous to solely base MPA locations on the basis of single solutions, and it is particularly useful to consider ensembles of solutions. Hydrodynamic provinces can then provide a useful supplement in the framework of parameters that are taken into account when planning MPAs. However, in an ensemble of solutions, it may still be difficult to assess which locations are often connected to one another. When an assessment of connectivity between existing MPAs or specific potential MPA sites is desired, it may therefore be useful to also directly consider the exchange of particles between specific regions, as is done in Coleman et al. [9]. Moreover, when clustering existing MPAs based on their connectivity, it is useful to also consider possible degeneracy.

## 5.3 Comparison to existing literature and implications for future research

In many respects, this thesis extends on the method for identifying hydrodynamic provinces through community detection in Lagrangian flow networks as proposed by Rossi et al. [60] and Ser-Giacomi et al. [71]. Foremost, we assessed the importance that solution degeneracy plays in yielding different partitions and stress the importance of considering ensemble solutions. Other key differences are the consideration of the Markov-time parameter, the application to a larger, open domain, an assessment of the evolution of global quality metrics in a seasonal and yearly context, and the establishment of correlations between hydrodynamic province boundaries, flow speed and sea ice.

Still, many improvements can be made to the accuracy of hydrodynamic provinces. First, improvements can be made to the modeling of particle trajectories. Accuracy can be improved by considering larger particle ensembles. Moreover, the artificial diffusion introduced by using transition matrix can be assessed and may be supplemented. This can be parameterized in Lagrangian modeling by adding a stochastic diffusion term related to the local eddy diffusivity [59]. Additionally, wave-driven Stokes drift influences particle trajectories at the surface and may be included in simulations [54].

Besides, a logical extension of this research would be the consideration of a larger, or perhaps global, domain. This goes hand in hand with longer computation times, but computational efficiency may be further improved. For example, a larger advection timestep $\Delta t$ may be used, although this comes with a decrease in accuracy of the particle trajectories.

Moreover, the information-theoretical approach used by *Infomap* only indirectly optimizes for the coherence of and mixing in communities. Here, the general criteria for community detection in graph theory are translated into specific criteria that are relevant for oceanography. The construction of an algorithm that directly optimizes these criteria would prove useful for our applications.

Furthermore, community divisions may be further fine-tuned to increase the relevance for marine spatial planning. For example, when connectivity of specific marine species is considered, individual particle behavior may be modified to mimic species behavior [32, 7], instead of considering passive buoyant particles, to yield a community description of the domain relevant to an individual species.

Lastly, since the Arctic domain is largely subject to climatic changes, it is interesting to investigate future connectivity of the Arctic ocean by using velocity field output from coupled global climate models. This can reveal how changes in the climate may affect barriers to transport. In turn, this may help policymakers design Marine Protected Areas such that their future connectivity is more resilient to climate change.

# 6 | Conclusion

In this thesis, we have applied the *Infomap* algorithm to detect hydrodynamic provinces in the Arctic ocean surface by using a network description of the flow.

We have given a comprehensive account of how *Infomap* finds community partitions and how the underlying algorithm relates to our Lagrangian flow networks. Similarities and differences between the Lagrangian particles that give rise to the transition matrix and the random walkers considered by *Infomap* have been discussed. The Markov-time parameter is a useful way to tune the community sizes and allows connectivity to be assessed at different spatial scales.

Since *Infomap* yields degenerate solutions, care must be taken with the interpretation of single partitions. Instead, ensembles of solutions may be investigated to determine the persistence of community borders. We also present a method to assess in which regions node assignment is consistent, supplementing the identification of persistent border nodes. It is useful to assess the standard deviation in global quality parameters associated to solution degeneracy, in order to assess the significance of temporal trends.

We have shown that Lagrangian flow networks may be used for assessing connectivity in open domains. The effect of using an open domain remains largely limited to the vicinity of the domain boundary.

Although in certain regions the resolution of our data and domain discretization is too coarse to resolve structures at the Rossby radius of deformation or smaller, *Infomap* is still able to resolve important oceanographic structures in the Arctic ocean, such as different currents, seas and the Beaufort Gyre. We find correlations between codelength, the global mixing parameter and the global coherence ratio, verifying that *Infomap* indirectly optimizes for coherence and mixing. We also find a correlation between the presence of boundary nodes and mean surface speed in ice-free regions, indicating that currents provide barriers to transport. In regions with sea ice, coherence is large, but mixing is small, since particles move slowly and in concert. We also find that the decrease in Arctic summer sea ice cover is mirrored in increasing codelength and mixing. The seasonal cycle, dominated by changes in sea ice cover, is also mirrored in codelength and global mixing.

# Bibliography

[1] M. Andrello et al. 'Low Connectivity between Mediterranean Marine Protected Areas: A Biophysical Modeling Approach for the Dusky Grouper Epinephelus marginatus'. In: *PLOS ONE* 8.7 (2013). Ed. by J. G. Hiddink, e68564. DOI: 10.1371/journal.pone.0068564 (cit. on p. 1).

[2] Arctic Monitoring and Assessment Programme (AMAP). *AMAP assessment report: Arctic pollution issues.* Ed. by A. M. a. A. Programme. Oslo, Norway: Arctic Monitoring and Assessment Programme, 1998. ISBN: 978-82-7655-061-0 (cit. on p. 49).

[3] L. Bohlin, D. Edler, A. Lancichinetti and M. Rosvall. 'Community detection and visualization of networks with the map equation framework'. In: *Measuring Scholarly Impact.* Ed. by Y. Ding, R. Rousseau and D. Wolfram. Springer Cham, 2014, pp. 3–34. ISBN: 978-3-319-10376-1. DOI: 10.1007/978-3-319-10377-8_1 (cit. on p. 10).

[4] S. Brin and L. Page. 'The anatomy of a large-scale hypertextual web search engine'. In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117. DOI: 10.1016/S0169-7552(98)00110-X (cit. on p. 9).

[5] N. Broekhuizen. 'Simulating motile algae using a mixed Eulerian-Lagrangian approach'. In: *Journal of Plankton Research* 21.7 (1999), pp. 1191–1216. DOI: 10.1093/plankt/21.7.1191 (cit. on p. 1).

[6] J. Calatayud et al. 'Exploring the solution landscape enables more reliable network community detection'. In: *Physical Review E* 100 (5 2019), p. 052308. DOI: 10.1103/PhysRevE.100.052308 (cit. on pp. 2, 12, 28).

[7] P. Cetina-Heredia et al. 'Strengthened currents override the effect of warming on lobster larval dispersal and survival'. In: *Global Change Biology* 21.12 (2015), pp. 4377–4386. DOI: 10.1111/gcb.13063 (cit. on pp. 1, 39).

[8] M. Chevallier et al. 'Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project'. In: *Climate Dynamics* 49.3 (2017), pp. 1107–1136. DOI: 10.1007/s00382-016-2985-y (cit. on p. 16).

[9] M. A. Coleman et al. 'Anticipating changes to future connectivity within a network of marine protected areas'. In: *Global Change Biology* 23.9 (2017), pp. 3533–3542. DOI: 10.1111/gcb.13634 (cit. on pp. 1, 39).

[10] J. C. Comiso. 'Large Decadal Decline of the Arctic Multiyear Ice Cover'. In: *Journal of Climate* 25.4 (2012), pp. 1176–1193. DOI: 10.1175/JCLI-D-11-00113.1 (cit. on p. 2).

[11] R. K. Cowen and S. Sponaugle. 'Larval Dispersal and Marine Population Connectivity'. In: *Annual Review of Marine Science* 1.1 (2009), pp. 443–466. DOI: 10.1146/annurev.marine.010908.163757 (cit. on p. 1).

[12] B. Cushman-Roisin and J.-M. Beckers. *Introduction to geophysical fluid dynamics: physical and numerical aspects*. Vol. 101. Academic press, 2011. ISBN: 9780120887590 (cit. on p. 15).

[13] L. Danon, A. Díaz-Guilera, J. Duch and A. Arenas. 'Comparing community structure identification'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (2005), P09008–P09008. DOI: `10.1088/1742-5468/2005/09/P09008` (cit. on pp. 4, 27).

[14] D. P. Dee et al. 'The ERA-Interim reanalysis: configuration and performance of the data assimilation system'. In: *Quarterly Journal of the Royal Meteorological Society* 137.656 (2011), pp. 553–597. DOI: `10.1002/qj.828`. eprint: `https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.828` (cit. on p. 15).

[15] P. Delandmeter and E. van Sebille. 'The Parcels v2.0 Lagrangian framework: new field interpolation schemes'. In: *Geoscientific Model Development Discussions* (2019), pp. 1–24. DOI: `10.5194/gmd-2018-339` (cit. on p. 19).

[16] M. Drévillon et al. *Quality Information Document for the Global Ocean Physical Reanalysis Products GLOBAL_REANALYSIS_ PHY_001_030* (cit. on p. 16).

[17] D. Edler, L. Bohlin and a. Rosvall. 'Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap'. In: *Algorithms* 10.4 (2017), p. 112. DOI: `10.3390/a10040112` (cit. on p. 7).

[18] E. Fernandez and J. Lellouche. *Product User Manual for the Global Ocean Physical Reanalysis product GLOBAL_REANALYSIS_ PHY_001_030* (cit. on p. 15).

[19] S. Fortunato and M. Barthelemy. 'Resolution limit in community detection'. In: *Proceedings of the National Academy of Sciences* 104.1 (2007), pp. 36–41. DOI: `10.1073/pnas.0605965104` (cit. on p. 6).

[20] S. Fortunato. 'Community detection in graphs'. In: *Physics Reports* 486.3-5 (2010), pp. 75–174. DOI: `10.1016/j.physrep.2009.11.002` (cit. on pp. 4, 5).

[21] G. Froyland, K. Padberg, M. H. England and A. M. Treguier. 'Detection of Coherent Oceanic Structures via Transfer Operators'. In: *Physical Review Letters* 98.22 (2007), p. 224503. DOI: `10.1103/PhysRevLett.98.224503` (cit. on p. 6).

[22] G. Froyland, R. M. Stuart and E. van Sebille. 'How well-connected is the surface of the global ocean?' In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24.3 (2014), p. 033126. DOI: `10.1063/1.4892530` (cit. on pp. 1, 5).

[23] B. H. Good, Y.-A. De Montjoye and A. Clauset. 'Performance of modularity maximization in practical contexts'. In: *Physical Review E* 81.4 (2010), p. 046106. DOI: `10.1103/PhysRevE.81.046106` (cit. on pp. 2, 12).

[24] H. Goosse and T. Fichefet. 'Importance of ice-ocean interactions for the global ocean circulation: A model study'. In: *Journal of Geophysical Research: Oceans* 104.C10 (1999), pp. 23337–23355. DOI: `10.1029/1999JC900215` (cit. on p. 2).

[25] H. Goosse and T. Fichefet. 'Importance of ice-ocean interactions for the global ocean circulation: A model study'. In: *Journal of Geophysical Research: Oceans* 104.C10 (1999), pp. 23337–23355. DOI: `10.1029/1999JC900215` (cit. on p. 15).

[26] R. Guimera, S. Mossa, A. Turtschi and L. A. N. Amaral. 'The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles'. In: *Proceedings of the National Academy of Sciences* 102.22 (2005), pp. 7794–7799. DOI: `10.1073/pnas.0407994102` (cit. on p. 4).

[27] A. Hadjighasem et al. ‘A critical comparison of Lagrangian methods for coherent structure detection’. In: *Chaos* 27.5 (2017), pp. 1–25. DOI: 10.1063/1.4982720 (cit. on p. 2).

[28] G. Haller and G. Yuan. ‘Lagrangian coherent structures and mixing in two-dimensional turbulence’. In: *Physica D: Nonlinear Phenomena* 147.3-4 (2000), pp. 352–370. DOI: 10.1016/S0167-2789(00)00142-1 (cit. on p. 2).

[29] C. Hansen, E. Kvaleberg and A. Samuelsen. ‘Anticyclonic eddies in the Norwegian Sea; their generation, evolution and impact on primary production’. In: *Deep Sea Research Part I: Oceanographic Research Papers* 57.9 (2010), pp. 1079–1091. DOI: 10.1016/j.dsr.2010.05.013 (cit. on p. 26).

[30] T. S. Hopkins. ‘The GIN Sea—A synthesis of its physical oceanography and literature review 1972–1985’. In: *Earth-Science Reviews* 30.3-4 (1991), pp. 175–318. DOI: 10.1016/0012-8252(91)90001-V (cit. on p. 14).

[31] D. A. Huffman. ‘A method for the construction of minimum-redundancy codes’. In: *Proceedings of the IRE* 40.9 (1952), pp. 1098–1101 (cit. on p. 7).

[32] M. N. Jacobi, C. André, K. Döös and P. R. Jonsson. ‘Identification of subpopulations from connectivity matrices’. In: *Ecography* 35.11 (2012), pp. 1004–1016. DOI: 10.1111/j.1600-0587.2012.07281.x (cit. on pp. 1, 39).

[33] E. Jones, T. Oliphant, P. Peterson et al. *SciPy: Open source scientific tools for Python.* 2001 (cit. on p. 21).

[34] B. Karrer, E. Levina and M. E. J. Newman. ‘Robustness of community structure in networks’. In: *Physical Review E* 77.4 (2008). arXiv: 0709.2108, p. 046119. DOI: 10.1103/PhysRevE.77.046119 (cit. on p. 27).

[35] T. Kawamoto and M. Rosvall. ‘Estimating the resolution limit of the map equation in community detection’. In: *Physical Review E* 91.1 (2015), p. 012809. DOI: 10.1103/PhysRevE.91.012809 (cit. on p. 7).

[36] S. Khatiwala, M. Visbeck and M. A. Cane. ‘Accelerated simulation of passive tracers in ocean circulation models’. In: *Ocean Modelling* 9.1 (2005), pp. 51–69. DOI: 10.1016/j.ocemod.2004.04.002 (cit. on p. 11).

[37] M. Kheirkhahzadeh, A. Lancichinetti and M. Rosvall. ‘Efficient community detection of network flows for varying Markov times and bipartite networks’. In: *Physical Review E* 93.3 (2016), p. 032309. DOI: 10.1103/PhysRevE.93.032309 (cit. on pp. 6, 7, 10, 11).

[38] Y. Kim, S.-W. Son and H. Jeong. ‘Finding communities in directed networks’. In: *Physical Review E* 81.1 (2010), p. 016103. DOI: 10.1103/PhysRevE.81.016103 (cit. on p. 6).

[39] R. Lambiotte and M. Rosvall. ‘Ranking and clustering of nodes in networks with smart teleportation’. In: *Physical Review E* 85.5 (2012), p. 056107. DOI: 10.1103/PhysRevE.85.056107 (cit. on pp. 7, 10, 21).

[40] A. Lancichinetti and S. Fortunato. ‘Consensus clustering in complex networks’. In: *Scientific reports* 2 (2012), p. 336. DOI: 10.1038/srep00336 (cit. on p. 12).

[41] K. Logemann and I. Harms. ‘High resolution modelling of the North Icelandic Irminger Current (NIIC)’. In: *Ocean Science* 2.2 (2006), pp. 291–304. DOI: 10.5194/os-2-291-2006 (cit. on pp. 13, 30).

[42] S. Maneewongvatana and D. M. Mount. 'It's okay to be skinny, if your friends are fat'. In: *4th Annual CGC Workshop on Computational Geometry.* 1999, p. 8 (cit. on p. 21).

[43] N. Maximenko, J. Hafner and P. Niiler. 'Pathways of marine debris derived from trajectories of Lagrangian drifters'. In: *Marine Pollution Bulletin* 65.1-3 (2012), pp. 51–62. DOI: `10.1016/j.marpolbul.2011.04.016` (cit. on p. 1).

[44] R. McAdam and E. van Sebille. 'Surface Connectivity and Interocean Exchanges From Drifter-Based Transition Matrices'. In: *Journal of Geophysical Research: Oceans* 123.1 (2018), pp. 514–532. DOI: `10.1002/2017JC013363` (cit. on p. 11).

[45] M. Meilă. 'Comparing clusterings—an information based distance'. In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. DOI: `10.1016/j.jmva.2006.11.013` (cit. on p. 27).

[46] N. Molkenthin, K. Rehfeld, N. Marwan and J. Kurths. 'Networks from Flows - From Dynamics to Topology'. In: *Scientific Reports* 4.1 (2015), p. 4119. DOI: `10.1038/srep04119` (cit. on p. 4).

[47] A. Montanari and A. Saberi. 'The spread of innovations in social networks'. In: *Proceedings of the National Academy of Sciences* 107.47 (2010), pp. 20196–20201. DOI: `10.1073/pnas.1004098107` (cit. on p. 4).

[48] L. Moresi and B. Mather. 'Stripy: A Python module for (constrained) triangulation in Cartesian coordinates and on a sphere.' In: *Journal of Open Source Software* 4.38 (2019), p. 1410. DOI: `10.21105/joss.01410` (cit. on p. 18).

[49] L. A. Mysak and F. Schott. 'Evidence for baroclinic instability of the Norwegian Current'. In: *Journal of Geophysical Research* 82.15 (1977), pp. 2087–2095. DOI: `10.1029/JC082i015p02087` (cit. on pp. 13, 26).

[50] NEMO System Team. *NEMO ocean engine.* 27. Version 3.1. Zenodo. DOI: `10.5281/zenodo.1464816` (cit. on p. 15).

[51] M. Newman and M. Girvan. 'Finding and evaluating community structure in networks'. In: *Physical Review E* 69.2 (2004), p. 026113. DOI: `10.1103/PhysRevE.69.026113` (cit. on pp. 4, 6).

[52] M. Newman. *Networks.* Oxford university press, 2018. ISBN: 9780198805090 (cit. on pp. 4–6).

[53] A. J. G. Nurser and S. Bacon. 'The Rossby radius in the Arctic Ocean'. In: *Ocean Science* 10.6 (2014), pp. 967–975. DOI: `10.5194/os-10-967-2014` (cit. on p. 16).

[54] V. Onink, D. Wichmann, P. Delandmeter and E. Sebille. 'The Role of Ekman Currents, Geostrophy, and Stokes Drift in the Accumulation of Floating Microplastic'. In: *Journal of Geophysical Research: Oceans* 124.3 (2019), pp. 1474–1490. DOI: `10.1029/2018JC014547` (cit. on p. 39).

[55] S. R. Palumbi. 'Population Genetics, Demographic Connectivity, and the Design of Marine Reserves'. In: *Ecological Applications* 13.sp1 (2003), pp. 146–158. DOI: `10.1890/1051-0761(2003)013[0146:PGDCAT]2.0.CO;2` (cit. on p. 1).

[56] I. V. Polyakov et al. 'Recent oceanic changes in the Arctic in the context of long-term observations'. In: *Ecological Applications* 23.8 (2013), pp. 1745–1764. DOI: `10.1890/11-0902.1` (cit. on p. 2).

[57] Protection of the Arctic Marine Environment (PAME) Working Group. *Framework for a Pan-Arctic Network of Marine Protected Areas*. Tech. rep. Arctic Council, 2015 (cit. on p. 2).

[58] H. C. Regan, C. Lique and T. W. K. Armitage. 'The Beaufort Gyre Extent, Shape, and Location Between 2003 and 2014 From Satellite Observations'. In: *Journal of Geophysical Research: Oceans* 124.2 (2019), pp. 844–862. DOI: 10.1029/2018JC014379 (cit. on p. 14).

[59] O. N. Ross and J. Sharples. 'Recipe for 1-D Lagrangian particle tracking models in space-varying diffusivity'. In: *Limnology and Oceanography: Methods* 2.9 (2004), pp. 289–302. DOI: 10.4319/lom.2004.2.289 (cit. on p. 39).

[60] V. Rossi, E. Ser-Giacomi, C. Lõpez and E. Hernández-García. 'Hydrodynamic provinces and oceanic connectivity from a transport network help designing marine reserves'. In: *Geophysical Research Letters* 41.8 (2014), pp. 2883–2891. DOI: 10.1002/2014GL059540 (cit. on pp. 1–6, 15, 19, 30, 39).

[61] M. Rosvall, D. Axelsson and C. T. Bergstrom. 'The map equation'. In: *The European Physical Journal Special Topics* 178.1 (2009), pp. 13–23. DOI: 10.1140/epjst/e2010-01179-1 (cit. on pp. 2, 6, 7, 9, 10).

[62] M. Rosvall and C. T. Bergstrom. 'Maps of random walks on complex networks reveal community structure'. In: *PNAS* 105.4 (2007), pp. 1118–1123. DOI: 10.1073/pnas.0706851105 (cit. on pp. 6, 8).

[63] M. Rosvall and C. T. Bergstrom. 'Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems'. In: *PLOS ONE* 6.4 (2011). Ed. by F. Rapallo, e18209. DOI: 10.1371/journal.pone.0018209 (cit. on p. 10).

[64] M. Rosvall, J.-C. Delvenne, M. T. Schaub and R. Lambiotte. 'Different approaches to community detection'. In: *arXiv:1712.06468 [physics]* (2017). arXiv: 1712.06468 (cit. on p. 6).

[65] R. Sadourny, A. Arakawa and Y. Mintz. 'Integration of the nondivergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere'. In: *Monthly Weather Review* 96.6 (1968), p. 6 (cit. on p. 18).

[66] R. Sætre. 'Features of the central Norwegian shelf circulation'. In: *Continental Shelf Research* 19.14 (1999), pp. 1809–1831. DOI: 10.1016/S0278-4343(99)00041-2 (cit. on p. 13).

[67] M. T. Schaub, J.-C. Delvenne, S. N. Yaliraki and M. Barahona. 'Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit'. In: *PLOS ONE* 7.2 (2012). Ed. by O. Sporns, e32210. DOI: 10.1371/journal.pone.0032210 (cit. on p. 10).

[68] M. T. Schaub, R. Lambiotte and M. Barahona. 'Encoding dynamics for multiscale community detection: Markov time sweeping for the map equation'. In: *Physical Review E* 86.2 (2012), p. 026112. DOI: 10.1103/PhysRevE.86.026112 (cit. on p. 10).

[69] E. van Sebille. 'Adrift.org.au — A free, quick and easy tool to quantitatively study planktonic surface drift in the global ocean'. In: *Journal of Experimental Marine Biology and Ecology* 461 (2014), pp. 317–322. DOI: 10.1016/j.jembe.2014.09.002 (cit. on p. 11).

[70] E. van Sebille et al. 'Lagrangian ocean analysis: Fundamentals and practices'. In: *Ocean Modelling* 121 (2018), pp. 49–75. DOI: `https://doi.org/10.1016/j.ocemod.2017.11.008` (cit. on pp. 1, 5).

[71] E. Ser-Giacomi, V. Rossi, C. López and E. Hernández-García. 'Flow networks: A characterization of geophysical fluid transport'. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.3 (2015), p. 036404. DOI: `10.1063/1.4908231` (cit. on pp. 2–6, 11–13, 15, 19, 21, 26, 30, 37, 39).

[72] C. E. Shannon. 'A mathematical theory of communication'. In: *Bell system technical journal* 27.3 (1948), pp. 379–423. DOI: `10.1002/j.1538-7305.1948.tb01338.x` (cit. on p. 8).

[73] K. Soramäki et al. 'The topology of interbank payment flows'. In: *Physica A: Statistical Mechanics and its Applications* 379.1 (2007), pp. 317–333. DOI: `10.1016/j.physa.2006.11.093` (cit. on p. 4).

[74] A. Strehl and J. Ghosh. 'Cluster ensembles—a knowledge reuse framework for combining multiple partitions'. In: *Journal of machine learning research* 3 (3 2003), pp. 583–617. DOI: `10.1162/153244303321897735` (cit. on p. 12).

[75] C. J. Thomas et al. 'Numerical modelling and graph theory tools to study ecological connectivity in the Great Barrier Reef'. In: *Ecological Modelling* 272 (2014), pp. 160–174. DOI: `10.1016/j.ecolmodel.2013.10.002` (cit. on p. 1).

[76] D. Vaughan et al. 'Observations: Cryosphere'. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013. Chap. 4, pp. 317–382. ISBN: ISBN 978-1-107-66182-0. DOI: `10.1017/CBO9781107415324.012` (cit. on p. 2).

[77] A. Viamontes Esquivel and M. Rosvall. 'Compression of Flow Can Reveal Overlapping-Module Organization in Networks'. In: *Physical Review X* 1.2 (2011), p. 021025. DOI: `10.1103/PhysRevX.1.021025` (cit. on p. 10).

[78] N. Wang and J.-L. Lee. 'Geometric Properties of the Icosahedral-Hexagonal Grid on the Two-Sphere'. In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2536–2559. DOI: `10.1137/090761355` (cit. on pp. 18, 19).

# A | Supplementary Figures

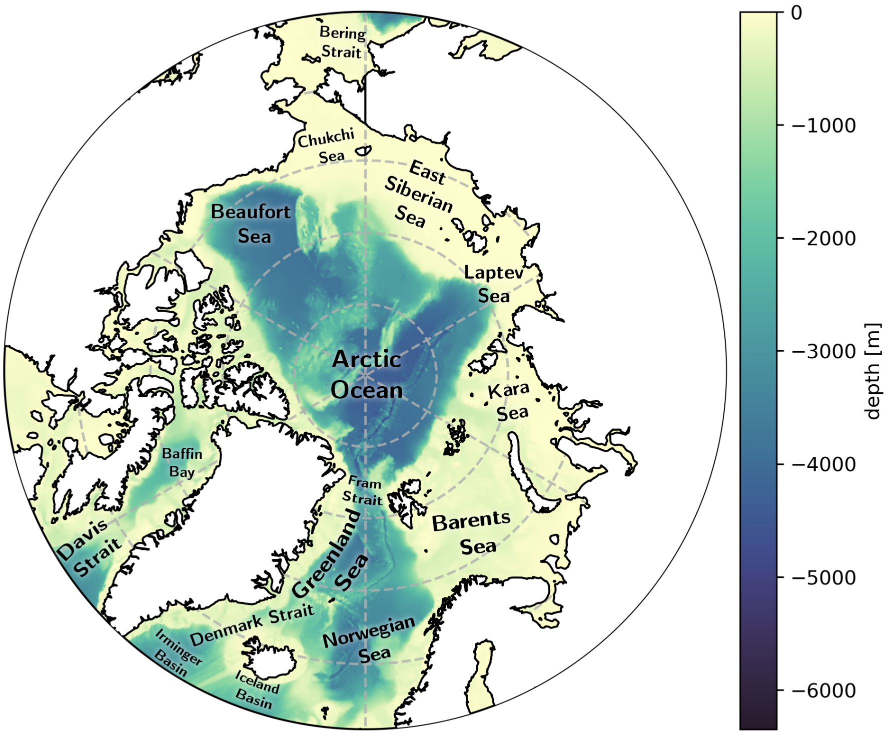


**Figure S1:** Seas and straits in the Arctic. Colors indicate bathymetry.

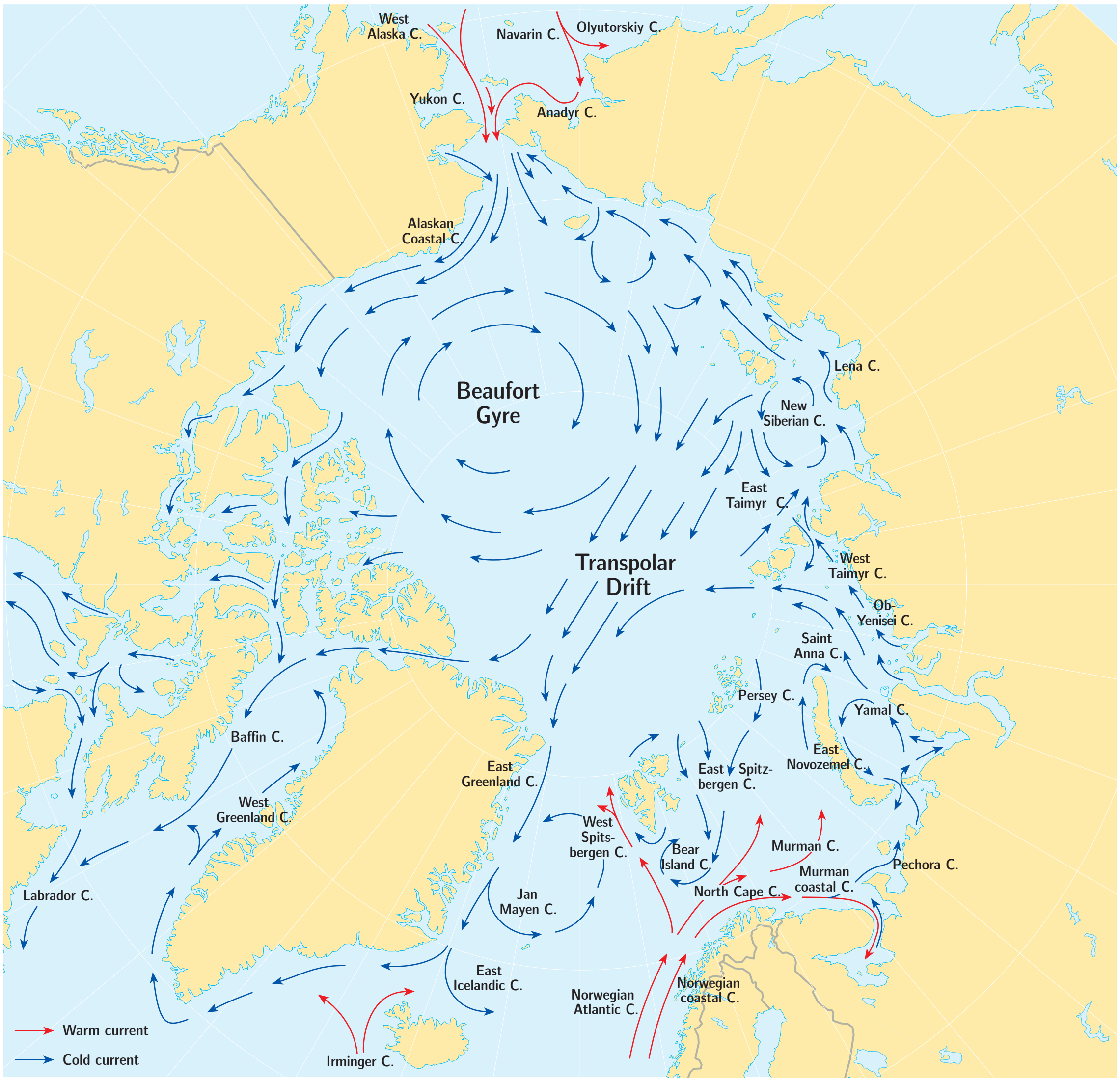
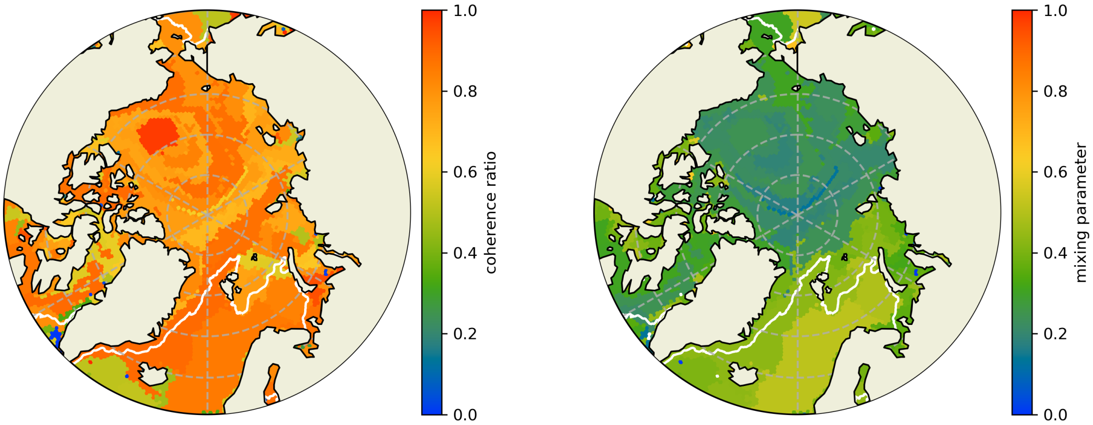


**Figure S2:** Surface currents in the Arctic. Reproduced from the Arctic Monitoring and Assessment Programme [2].



**(a)** Coherence ratio

**(b)** mixing parameter

**Figure S3:** The coherence ratio and mixing parameter associated to each community in the partition depicted in 4.2b, for $\mathbf{P}(t_0 = \text{March 1st, } 2018, \tau = 90 \text{ days})$.