Geospatial Trajectory Clustering to Identify Local Arctic Sea Ice Drift Patterns Related to Climate Forcing and Melting Sea Ice

by

Annabel Flatland

Professor Alice Bradley, Advisor

A thesis submitted in partial fulfillment

of the requirements for the

Degree of Bachelor of Arts with Honors

in Geosciences

WILLIAMS COLLEGE

Williamstown, Massachusetts

July 21, 2024

Abstract

This study compares local sea ice drift patterns in four Arctic locations from 1980 to 2022. Using the Polar Pathfinder Daily Ice Motion dataset, we track yearly drift trajectories for ice originating in 250 x 250 km^2 regions throughout the Arctic Ocean. For each year, this results in a characteristic drift pattern. Outlier drift years are identified using a density-based clustering algorithm (DBSCAN), and the remaining trajectories are sorted into three drift regimes with k-means clustering. We find a strong temporal relationship in drift patterns for ice originating in the marginal ice zone which has transitioned to a primarily first-year ice environment. We find weaker time-dependency in the central Arctic and none in persistent multi-year ice off Greenland's coast. This indicates that local drift patterns are strongly influenced by the transition to thinner and less persistent sea ice coverage. Local drift patterns appear mostly unrelated to large-scale climate indices like the Arctic Oscillation or the North Atlantic Oscillation, and years with similar drift patterns are mostly inconsistent across different locations. Our findings suggest that local climate forces drive local drift more than large-scale forces, although causal studies are needed to draw stronger conclusions. This study introduces new techniques in geospatial machine learning and presents key differences between small-scale and large-scale drift patterns.

Acknowledgments

This thesis has been a surprisingly massive undertaking (as everyone told me it would be), and my name feels strangely alone on the title page when there are so many people who have supported me along the way and made this project possible.

First I would like to thank my advisor, Professor Alice Bradley, for finding a way to combine all my interests into one project, and being my mentor through four years of research at Williams. Going to Bremerhaven to present an early version of my thesis work was a highlight of my college career, and honestly expanded my view of science to include a whole community of researchers and engineers. I will also never forget my the trip to New Hampshire and my stay in Granny's Attic, or the independent study for which only a few duckies were sacrificed. Thank you for your guidance throughout all my research projects—for pushing me to be more detail oriented, dig deeper into my results, and encouraging my interest in computer science.

I would also like to thank Professor Ronadh Cox for being my second reader. Thank you for your feedback on my second draft, and for being so flexible with my lateness. Your comments have made my final version stronger and more complete. I still think about the writing skills we learned in Sedimentology when I write anything important, including this thesis.

Thank you as well to Professor Rohit Bhattacharya in the Computer Science department for discussing machine learning and causal inference applications with me before I was ever

officially in your class. Thank you for always having time for my questions, and you class has had a big influence in the way I think about data analysis.

Also what would I do without all the people who made the geoscience department a wonderful place to work. Thank you to Dawn for keeping the kitchen stocked with endless snacks and for financing my tea addiction through many late nights (and some early mornings) in the GEOS lounge. Also thank you to Kennedy Lange, for making the Ice Lab an amazingly fun place since freshman year (go read her thesis!). And thank you to Berit Olsson: even though you're an astro major, I don't know what I would have done without our frolics through the woods when we both had so much writing to do.

I am also so lucky to have my family, who listened to all my problems and reminded me to take everything one step at a time. Thank you for being my captive audience through some pretty rough practice presentations and supporting me in every possible way.

Finally, thank you to all my fellow GEOS thesis-ers, for their emotional support and camaraderie, creating an incredible community I never expected to find. This thesis would not have been nearly as meaningful or complete without all of the incredible people and resources that have been a part of this journey.

Contents

List of Figures

Chapter 1

Introduction

In 1893, Fritjof Nansen froze his ship into Arctic sea ice, attracting international attention. He was not the first captain to do so: a decade earlier, the Jeanette was abandoned in an ice pack near Siberia, only to be discovered off the coast of Greenland 3 years later. However, Nansen was the first to freeze his ship intentionally. Nansen believed his ship would be transported poleward from Siberia, like the Jeanette, taking a crew over the North Pole for the first time before sending them to Greenland [Weeks, 2010]. Despite being told by experienced explorers like Aldophus Greely that such a feat was "an illogical scheme of self-destruction," Nansen secured a grant from the Norwegian Parliament to undertake the Fram expedition [Evans, 2012].

For three years, there was little word from Nansen or the Fram. Then, in the summer of 1896, as if popping out of a worm hole, the Fram appeared on the opposite side of the Arctic near Tromsø, Norway [1] [Evans, 2012]. Nansen and his crew were received enthusiastically by the King, the press, trained acrobats, and crowds of thousands. One of the most enthusiastic to see Nansen must have been Professor Henrik Mohn, who had theorized transpolar sea ice

[1] The ship arrived without Nansen, who got lost after unboarding near the North Pole. His plan was to ski to the pole, although after days of difficult travel he could only reach 86'N (170 miles farther north than any previous record) [Nansen, 1897].

drift, but had no observational proof [Nansen, 1897]. The data collected during the Fram expedition is the first record of large scale sea ice movements.

From Nansen's detailed observations and subsequent sea ice drift studies, we know that sea ice motion is largely dependent on winds and ocean currents, internal ice stresses, the Coriolis effect, and Ekman transport [Weeks, 2010]. These elements work together to move ice around the Arctic Ocean, forming two major ice motion features: the Transpolar Drift Stream (TDS) and the Beaufort Gyre (BG). While general patterns of sea ice drift are by now well established, the center and speed of these drift patterns varies every year. Previous studies have connected variations in drift patterns to large scale climate forcing and thinning sea ice, but there is little knowledge about whether these repeated variations are basin-wide, or local to smaller regions. Understanding variations in sea ice drift is important for understanding the redistribution of sea ice, with implications for ice growth, navigation, climate modeling, subsistence hunting, and more.

Figure 1.1: The Fram frozen into Arctic sea ice in the summer of 1894. The windmill on the deck provided electricity for the crew. Reproduced from the Norwegian Library of Congress.

1.1 Sea Ice Drift Patterns

Arctic sea ice is a thin, 1 to 3 meter thick layer over the sea surface. In the winter, this shell covers the central Arctic Ocean, or about 15 million square kilometers. Sea ice is mobile, moving mainly in response to winds and ocean currents. In fact, Arctic sea ice movement is defined by two major drift patterns: the anticyclonic Beaufort Gyre (BG) in the Canadian Arctic, and the Transpolar Drift Stream (TDS) running from Russia across the pole to

Greenland (fig. 1.2).

1.1.1 Short and Long Term Ice Motion

Away from the coasts, sea ice motion can be largely attributed to winds and ocean currents. However, their proportional influence is dependent on the time frame. In the short term (weeks), winds explain 70% of sea ice motion, while the rest is attributed to ocean circulation. These short term paths are meandering, often reversing direction for weeks at a time [Leppäranta, 2011]. In the long term (several months), winds explain 50% of ice motion, while the remainder is explained by ocean currents. The role of winds and currents are not as significant within ∼10-400 km of the coast, where sea ice frozen to shore or grounded to the shallow sea floor is less easily moved [Thorndike and Colony, 1982].

1.1.2 Beaufort Gyre (BG)

The BG is a region of clockwise circulation centered over the Beaufort Sea [Polyak et al., 2010, Serreze and Meier, 2019]. Sea Ice in the BG drifts at speeds of 0.85 - 2.6 km / day [Polyak et al., 2010]. Since this ice may recirculate, the BG contains some of the oldest sea ice in the Arctic [Leppäranta, 2011, Rigor et al., 2002]. The oldest ice is at the center of the BG, while the youngest clings to the edges and can take years to reach the center. When there is strong BG circulation, this causes the compaction of sea ice and an increased retention of multiyear ice. A weaker BG is associated with the dispersal of sea ice and a decreased retention of multiyear ice.

In recent decades, the BG has intensified and then stabilized [Lin et al., 2023]. The early 2000s were characterized by a period of intensification, partially as a result of sea ice loss: since 1979, the September sea ice extent has lost 2.3 million km^2 of sea ice, with the BG region seeing the largest losses in ice extent and thickness [Timmermans and Toole, 2023].

Figure 1.2: Map of the Arctic. The Beaufort Gyre is located in the Canadian Arctic, while the Transpolar Drift Stream stretches from Russia, across the pole to Greenland. The exact location and intensity of these sea ice drift patterns vary from year to year. Modified from `https://en.wikipedia.org/wiki/File:Arctic_Ocean_circulation_map.svg`.

Sea ice plays an important role in rotating the gyre, since ice exerts a larger drag force on water than wind on water alone. At least half of sea ice movement is related to the wind [Thorndike and Colony, 1982]. However, thick sea ice full of internal ice stresses resists the wind, slowing the movement of the gyre. The gyre is therefore sped up when sea ice is thin and responds more strongly to wind [NSIDC, 2024]. As ice becomes weaker, there is also an active eddy field under the ice that is likely to play a larger role in gyre motion [Timmermans and Toole, 2023]. Since 2008 the BG has stabilized, shifting south towards the Canadian Basin [Lin et al., 2023]. While the BG was centered near this location as recently as 2003, the BG is significantly stronger than in past decades.

1.1.3 Transpolar Drift Stream (TDS)

Motion along the TDS stretches from the Eurasian coast, across the pole, leaving the Arctic through the Fram Strait [Polyak et al., 2010, Serreze and Meier, 2019]. Sea ice in the TDS moves faster than ice in the BG, reaching speeds as high as 8.5 - 25 km / day [Martin and Gerdes, 2007]. Drift speeds increase closer to the Fram Strait, where there are fewer internal stresses and ice moves linearly instead of meandering [Weeks, 2010, NSIDC, 2024]. The Fram Strait is responsible for exporting nearly all the ice that leaves the Arctic [Polyak et al., 2010]: this is an estimated 3154 km^3 of ice per year [Weeks, 2010].

A strong TDS was previously connected to an increased long range transport of ice along the TDS: as ice is advected away from the coasts to the TDS, leads open, which then freeze over and produce more ice [Preußer et al., 2016]. However, this process may be interrupted as water along the coasts warm in response to climate change. Warm waters cannot replace ice that has been advected away. Ice that is produced is weak, and less likely to survive long range transport. The effects of weakening ice production is already noticeable: in 1990, up to 50% of first year ice survived to enter the TDS, while in 2005 only 20% of that ice survived

[Krumpen et al., 2019]. While a strong TDS does not necessarily increase long range ice transport, a weak TDS is associated with decreased sea ice transport and the retention of multiyear ice.

Like the BG, the TDS has intensified in the past decade. This is likewise related to sea ice loss: lower sea ice coverage weakens TDS. While the BG tightens with intensity, the TDS becomes wider [Kwok et al., 2013a]. The TDS is also shifting toward the Canadian basin in response to a growing BG [Kwok et al., 2013a, Yin et al., 2021], and the surface area of ice exported through the Fram Strait is increasing [Kwok et al., 2013a].

1.1.4 Other Patterns of Ice Motion

While the BG and TDS are the main features associated with ice motion, there are several smaller ice motion patterns that occur. This includes drift parallel to the Alaskan coast and near the pole perpendicular to the Canadian archipelago. In general, the highest drift rates occur near the ice edge, where ice floes are less likely to bump into one another [Weeks, 2010].

1.2 Variations in Drift Patterns

Long term patterns of ice motion like the BG and the TPD vary in size, strength and location. Previous studies have connected these variations to season, as well as large scale climate patterns. Recently, the trend toward thinner and less abundant sea ice has also been linked to significant changes in sea ice motion.

1.2.1 Seasonal Ice Motion

During the winter, ice is locked up and less likely to move. This results in a weaker and smaller BG and a weaker TPD. Conversely, during the warmer months, ice moves more easily in response to drivers like atmospheric pressure gradients. This results in faster BG circulation, a stronger TPD, and more outflow through the Fram Strait [Polyak et al., 2010].

1.2.2 Climate Modes

Previous studies have linked major differences in year long sea ice drift patterns with large scale climate patterns. These patterns are defined by long term fluctuations in sea level pressure, winds, and ocean currents, which themselves affect sea ice drift.

1.2.2.1 North Atlantic Oscillation (NAO)

The North Atlantic Oscillation (NAO) is defined by the SLP difference created from the northern hemisphere Ferrel cell: the subtropical high is centered over the Azores (Portugal, in the Atlantic Ocean), while the subpolar low is centered over Iceland. The positive phase of the NAO is associated with a stronger pressure difference, with an intensified low over Iceland [nce]. The strong Icelandic low weakens the anticyclonic Beaufort Gyre, slowing sea ice motion across the Beaufort Sea. The Icelandic low also weakens the Transpolar Drift, resulting in less ice exported through the Fram Strait [Kwok, 2000]. The negative phase of the NAO is associated with a smaller pressure difference between the subtropical high and subpolar low [nce]. This creates a strong, well defined BG and TPD, with higher drift speeds across the Arctic [Kwok, 2000].

The NAO was the first major climate mode used to describe sea ice motion, but only accounts for 22% of variation in Arctic SLP over winter and 3% over summer [Rigor et al., 2002]. This is because the NAO captures SLP around the Icelandic low, but fails to explain

other Arctic Ocean SLP variations, as well as the resulting ice movement [Rigor et al., 2002, Thompson and Wallace, 1998].

1.2.2.2 Arctic Oscillation (AO)

Since the 1950s, researchers have suspected that the NAO is part of a more general mode of variability affecting the northern hemisphere. This was formally identified as the Arctic Oscillation (AO) by Thompson and Wallace in the late 1998s. The AO is very similar to the NAO, in that there is a center of action over Icleand and in the Northern Atlantic near Portugal. The AO has an additional center of action in the Northern Pacific. Changes in the AO index are thought to be the surface manifestation of fluctuations in the intensity of the stratospheric polar vortex. In recent decades, the AO has become more positive. This is thought to be a result of stratospheric cooling cause by increased greenhouse gases, which increases the speed of the polar stratospheric vortex [Serreze and Meier, 2019].

In the positive phase of the AO, SLP is lower around Iceland, and higher in the Northern Pacific and Atlantic [Lindsey]. The lower pressure in the Arctic shrinks and slows the BG [Rigor et al., 2002]. The BG continues to exist except for rare episodes of extremely high AO. TPD compensates for the weakened BG, drifting towards the Canadian Basin to cut across the central Arctic [Wilson et al., 2021]. The transport of ice in the TPD increases, and more ice leaves the Arctic through the Fram Strait. This ice is mostly younger, thinner ice, which likely formed as the intensified TPD advects ice away from the coast, opening up leads in the East Siberian and Laptev seas [Rigor et al., 2002].

The negative phase of the AO is associated with higher SLP around Iceland, and lower SLP in the Northern Pacific and Atlantic [Lindsey, Rigor et al., 2002]. The higher pressure in the Arctic intensifies the BG, resulting in faster drift speeds. Ice is less likely to escape the gyre, and instead recirculates or forms ridges, becoming thick, multiyear ice. Meanwhile, the TPD is pushed toward the East Siberian Arctic [Wilson et al., 2021]. There is less advection

of ice away from the coasts, decreasing ice production [Rigor et al., 2002].

1.2.2.3 Arctic Dipole (AD)

The Arctic Dipole (AD) anomaly was identified because the AO and the NAO do not completely explain all sea ice motion anomalies [Holland, 2003, Rigor et al., 2002]. This caused many to conclude that there is likely a second, meridional, climate mode affecting sea ice motion. There are two centers of action for the AD: one over the Kara and Laptev Seas, and one stretching over the Canadian Archipelago to Greenland and Nordic seas [Wu et al., 2006].

A positive AD is characterized by positive SLP anomalies over the Canadian and Greenland marginal seas, and negative SLP anomalies centered over the Laptev Sea. Similar to the positive phase of the AO, this results in a weakened BG, increased export through the Fram Strait,and increased export from the Laptev and East Siberian Seas [Lei et al., 2021, Watanabe et al., 2006]. There is additionally cyclonic sea ice motion around the Laptev Sea negative SLP anomaly, and anticyclonic sea ice motion around the positive SLP anomaly [Wu et al., 2006].

A negative AD is characterized by negative SLP anomalies over the Canadian and Greenland marginal seas, and positive SLP anomalies over the Laptev Sea. Similar to the negative phase of the AO, the negative AD is associated with a strengthened BG, which retains sea ice in the central arctic and decreases ice export from the Arctic Basin [Wu et al., 2006, Watanabe et al., 2006, Zhang et al., 2022].

1.2.2.4 Central Arctic Index (CAI)

A more recently defined measure of meridional atmospheric forcing is the Central Arctic Index (CAI). Unlike the AD, whose centers can move, the CAI's centers are fixed: the CAI is simply the pressure gradient measured perpendicular to TPD across the Fram Strait

[Vihma et al., 2012]. More precisely, this was calculated by Zhang et. al. as the difference in SLP between 90°W, 84°N, and 90°E, 84°N [Zhang et al., 2022].

A positive CAI is associated with a stronger pressure gradient across the Fram Strait, resulting in stronger TPD, and a negative CAI is associated with a weaker pressure gradient across the Strait, resulting in weaker TPD. While the CAI is less studied than the AO or the NAO, Zhang et. al found that changes in sea ice speed throughout the Arctic were more closely correlated to the CAI than the NAO, AO, or AD, suggesting that changes in TPD have effects on the entire Arctic [Zhang et al., 2022].

1.2.3 Recent Trends in Sea Ice Properties

Sea ice is melting throughout the entire Arctic, decreasing in volume, concentration, age and thickness (fig. 1.3). In the marginal seas ice is melting completely, while in the central Arctic ice is thinning [Zhang et al., 2012]. Increased melting in the 21st century has decreased multiyear ice, or ice that has survived at least one summer season. This ice is the thickest in the Arctic, but is now only found off the coast of Greenland and the Canadian Arctic Archipelago [Kwok, 2018].

As a result of unprecedented melting, sea ice is weaker with fewer internal stresses. Where ice is abundant, weaker ice can contribute to thickening through the formation of ridges: these are thick sea ice features formed when 2 ice floes combine. In recent decades ridging has increased in the central Arctic, but decreased in the marginal seas, where sea ice is now sparse. Atypically thin ice from the marginal seas is then advected to the central Arctic, thinning the central Arctic as well [Zhang et al., 2012].

These drastic changes in sea ice qualities effect how sea ice moves. While Nansen observed that sea ice moves at 2% of surface wind speeds, this rule of thumb is changing because thinner ice is more responsive to winds and currents, drifting faster in response to the same

forces [Kaur et al., 2018, Meier, 2019]. Sea ice drift speeds are increasing throughout the Arctic, especially for areas with thin, first year ice [Zhang et al., 2022]. Areas that retain thick multiyear ice are not experiencing significant increases in drift speeds [Spreen et al., 2011, Kaur et al., 2018].

1.3 Arctic Remote Sensing

The Arctic is cold, treacherous, hard to get to, and generally not a pleasant place to make direct measurements of large scale sea ice movements. Luckily we do not (typically[2]) need to freeze ourselves into the sea ice like Nansen to collect the data we need to study sea ice drift. Since the start of the satellite era and the Arctic Ocean Buoy Program, there have been enough remote measurements to derive drift trajectories for most locations in the Arctic [Tschudi et al., 2016].

The satellite era began in the mid-1970s, when the first passive microwave sensors began observing the Arctic. Unlike visible or infrared light, passive microwave radiation is emitted even when there is no sun, and is often unaffected by cloud cover. These sensors observe the Arctic at least once per day, with the exception of a circular area over the pole known as the "pole hole." Unlike previous sea ice observations, the passive microwave data record is continuous for over 40 years. However, the data record is low resolution ($\sim$ 25 km), and less accurate where ice is melting, thin, or near the ice edge [Meier, 2022].

While the satellite era was just beginning, the International Arctic Buoy Program (IABP) was developing as well. This program started in 1979 by the U.S. National Academy of Science, becoming an international effort in 1991. Buoys provide meteorological and oceano-

[2]In 2019 the Fram Expedition was repeated by the MOSAiC expedition, where researchers from over 20 countries froze the Polarstern into sea ice near Severnaya Zemyla. Researchers took detailed measurements that aren't possible with remote sensing datasets, and this data is still being analyzed. Unfortunately, the coronavirus pandemic broke out a few months into the expedition, stranding scientists on the ship for longer than intended.

!h

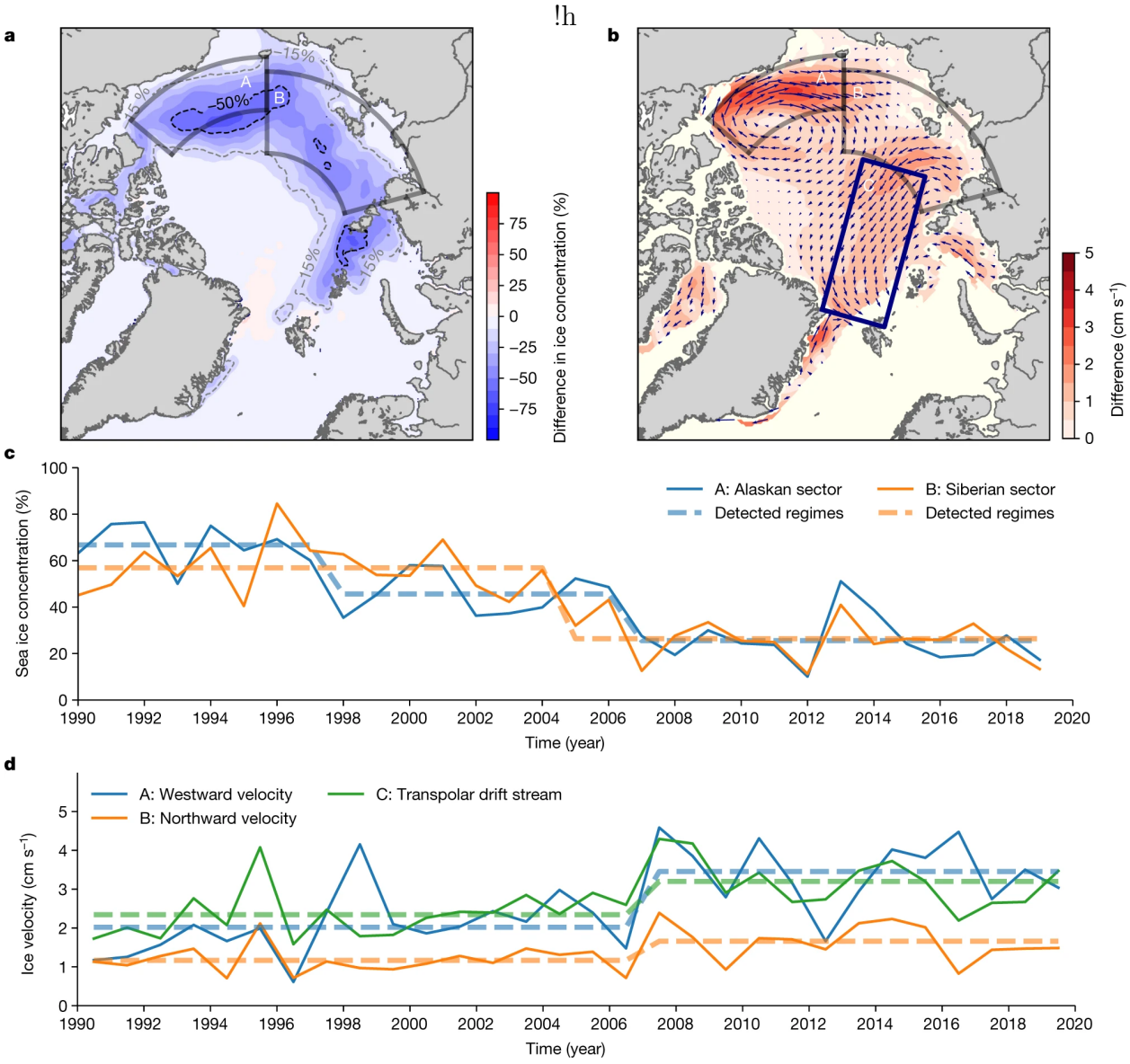

Figure 1.3: Difference in ice concentration and velocity between 1990 - 2006 and 2007-2019. (a) September sea ice concentration. The negative values indicate a loss in sea ice concentration in the later period. (b) Drift speeds calculated from December to May ice drift vectors. The transpolar drift stream is boxed in blue. (c) September sea ice concentration for the Alaskan (A) and Siberian (B) sectors shown on the maps. (d) Annual mean of ice velocity for sectors A, B, and the boxed transpolar drift stream area. Reproduced from Sumata et al. [2023].

graphic data for research purposes, typically taking measurements (such as GPS location) twice a day [Rigor and Ortmeyer, 2004]. Currently there are about 200 buoy maintained by the IABP, each with an average lifespan of about 18 months [Kiest, 2024].

1.4 Objectives

This thesis project has two goals. The first is to understand how variations in sea ice drift are connected throughout the Arctic Ocean, as well as the relationship to large scale climate forces and the transition to an Arctic covered by thinner, predominantly first-year ice. Previous studies have used climate indices to explain large-scale drift patterns, but few studies have considered whether these associations apply to small-scale drift patterns. Furthermore, most studies of the affects of large scale climate forces on sea ice are from the late 20th century and early 2000s, and haven't considered whether changes in ice properties in recent decades are having a larger effect on drift trajectories than climate variables.

The second goal is to develop methods for applying machine learning algorithms to geospatial data. Just as advances in ship-building technology enabled Nansen to study sea ice by becoming the first person to observe long-term drift, recent advances in machine learning algorithms may facilitate new discoveries using geospatial data. However, papers applying machine learning to geospatial data are still sparse, and so new methods need to be explored before many of these algorithms can be fully utilized [Rolf et al., 2024]. The methods developed in this project may be useful for future geospatial analyses and sea ice studies.

Using a data-driven approach, we tracked annual ice drift paths in four locations throughout the Arctic, and then determined years of similar drift. Two approaches were taken to find years of similar drift. In one, I used a subset of the path information, distilled to the most important components with principle component analysis (PCA). In the other, I used

the full path information, compressed with deep auto-encoders. In both cases, we used a clustering algorithm to sort annual drift paths by similarity. Once we determined years of similar drift for each location, we compared the groups to each other and to two major climate indices (Arctic Oscillation and North Atlantic Oscillation). We also analyzed the groups for similarities in basin-wide sea level pressures and wind velocities. Generally the drift patterns are not associated with each other or with climate indices, indicating that years of similar drift appear to be local rather than basin-wide. However, the clustered years of drift have a strong temporal association in regions where there have been the most dramatic changes in sea ice properties, indicating sea ice properties have a large influence on local ice trajectories.

Chapter 2

Methods

This chapter describes how we derived annual sea ice drift patterns for four locations through the Arctic. Sea ice motion is measured using a satellite-derived ice motion dataset (section 2.1.1), and developed an algorithm to compute ice trajectories from the ice motion vectors (section 2.2). We used this algorithm to calculate trajectories over the course of one year for ice floes originating in a particular source region ($\sim$250 x 250 km^2). Then I converted these trajectories to a reduced matrix form before applying clustering algorithms to the different years of ice drift. Our clustering analysis determines which sets of years the ice drift trajectories from a particular source region are the most similar (section 2.4). Then we compare the sets of years with similar ice drift to climate data (section 2.1.2).

I developed two different approaches for clustering. The first passes only the ice's start and end locations to the clustering algorithm, meaning the clustering algorithm has no knowledge of the complete ice drift trajectories (section 2.4.1). The second passes the ice's full drift path to the clustering algorithm (section 2.4.2). In both cases, we reduce the path data in size before clustering, compressing to the most important pieces of data. This is to overcome the "curse of dimensionality," or the diminishing performance of clustering algorithms as data points contain more information [Beyer et al., 1999]. Since the second

approach involves passing the complete trajectory information to the clustering algorithm, we compress the path data more aggressively. In both approaches, we find 2-3 clusters of similar drift patterns depending on the optimal "k."

2.1 Data

The data used in this project is sourced from a combination of satellite remote sensing and Arctic buoys. The Polar Pathfinder dataset uses multiple data sources to provide sea ice drift velocities, while additional meteorological and climate data is provided by NOAA.

2.1.1 Polar Pathfinder Dataset

Ice motion data is provided by the Polar Pathfinder Daily 25 km EASE-grid Sea Ice Motion Vectors, Version 4 [Tschudi et al., 2016]. Daily ice motion vectors cover the entire Arctic from September 15th, 1980 to December 31st, 2022. The product is defined using an azimuthal equal area EASE-Grid projection, which divides the Arctic into a 361 x 361 grid, centered on the north pole, where each grid cell is $\sim$25 x 25 km^2. The U and V ice velocity components are defined relative to this grid (fig. 2.1): In this map, $+U$ is motion from left to right, while $+V$ is motion from bottom to top [Tschudi et al., 2016]. This is different from the convention in which $+U$ is east and $+V$ is north.

2.1.1.1 Ice Velocity Data Sources

Ice velocity is computed from repeated remote sensing measurements by a combination of sensors, including satellite, buoy, wind, and passive microwave. Ice velocity is only calculated where sea ice concentration is greater that 15%, and where there are large swaths of ocean. There are no velocity estimates for regions without open ocean like the Canadian Archipelago [Tschudi et al., 2016].

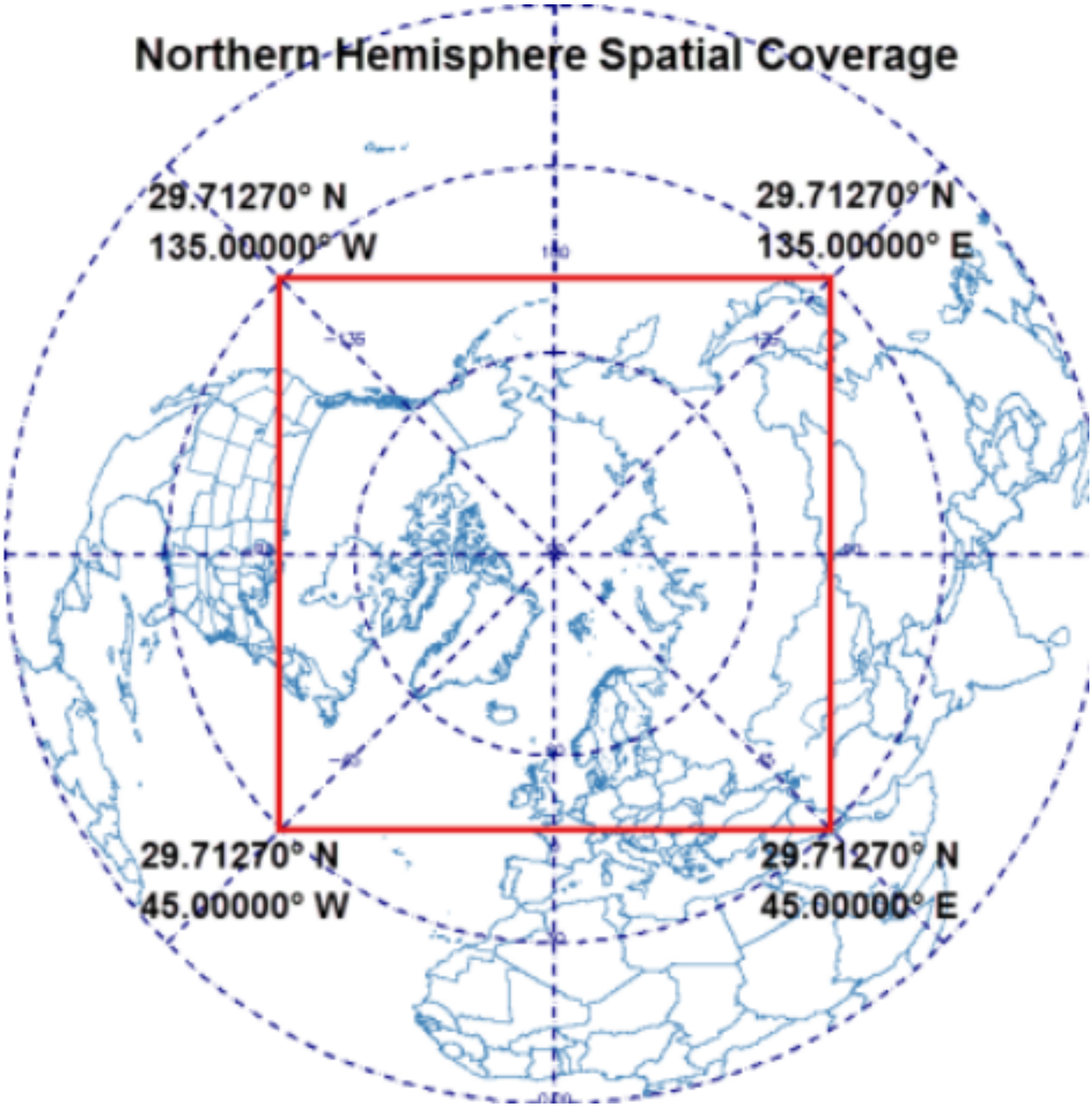

Figure 2.1: Polar Pathfinder 25 km Ease-Grid Sea Ice Motion Vectors spatial coverage. The red boxed area shows the area the dataset covers. U and V velocities are defined relative to the EASE-grid projection: In this map, +U is motion from left to right, while +V is motion from bottom to top. This is different from the convention in which +U is east and +V is north. Reproduced from Tschudi et al. [2016].

The AVHRR Global Area Coverage visible and infrared imagery is used to identify notable features such as ice floes, and ice velocity vectors are determined from the pixel displacement of these features in following images. Visible imagery is sun-dependent, available during late spring, summer, and early fall, but not during winter. Infrared imagery is not useful during melt periods, when sea ice and water are the same temperatures. Both types of AVHRR imagery are limited by cloud cover [Tschudi et al., 2019].

Passive microwave satellite data is available from four different instruments: SMMR, SSM/I, SSMIS, and AMSR-E. Brightness temperature data from these sensors is used to calculate ice motion. Passive microwave sensing is less sensitive to clouds compared to visible-infrared measurements, but this method is less accurate when ice is melting [Tschudi et al., 2019].

The most accurate ice motion estimates are from buoys. Buoy data is provided by The International Arctic Buoy Program (IABP), a network of satellite tracked buoys that typically record their locations every 12 hours with high accuracy. This method is limited by the number and location of buoys, which are expensive and logistically difficult to deploy, and there are no buoys off the coast of Russia. Buoys are mostly installed on multi-year ice [Tschudi et al., 2019].

The least accurate ice motion estimates in the Pathfinder dataset are calculated from wind velocity data. The wind data used by the pathfinder dataset is derived from NCEP/NCAR U-wind at 10 m. To convert wind estimates to ice drift velocities, Thorndike and Colony's (1982) observations are used: summer ice moves at 1% geostrophic wind speed, at angle of 20 degrees from the wind direction [Tschudi et al., 2019]. This estimation creates inaccuracies for several reasons. First, geostrophic winds affect sea ice less than surface winds, which are used in the calculation. Second, the dataset doesn't consider how winds affect ice motion less in the winter compared to the summer. Finally, recent studies have observed sea ice moving faster in response to the same wind forcing as sea ice thins and the Arctic becomes less

locked up [Kaur et al., 2018, Meier, 2019]. Therefore we should expect wind-based estimates of ice motion to generally underestimate ice velocities, especially in recent years. This is supported by comparisons of the Pathfinder derived drift trajectories to buoy drift paths by Gui et. al (2020).

Ice motion estimates from all sources are computed individually. The final ice motion estimate is a combination of every individual component, weighted by the expected accuracy of the data source. For example, buoys are weighted higher than wind derived estimates. The final estimates are mapped to the $\sim$25 x 25 km^2 grid cells [Tschudi et al., 2016].

2.1.1.2 Pathfinder-Derived Drift Trajectory Uncertainties and Validation

The ice motion vectors are less reliable where sea ice is melting, because surface melt makes the sea ice surface difficult to identify. The pathfinder data product does not include ice motion data for regions with a concentration of less than 15% ice concentration, meaning the computed ice trajectories may appear shorter than the actual ice trajectories. Ice motion calculations are missing over the pole and in regions where there is not enough open ocean, such as near the Canadian Archipelago.

Gui et. al (2020) validated the Polar Pathfinder data set for the years 2014 and 2016. Ice trajectories for ice originating in one of the grid cells were computed with the Polar Pathfinder dataset, using similar methods to those outlined in 2.2. These were compared to buoy drift paths which were not yet incorporated into ice motion products (fig. 2.2). The Polar Pathfinder derived drift trajectories were most accurate for meridional drift with less meander, and less accurate over zonal drift and areas with lower ice concentrations. Errors tended to accumulate over time, and generally sea ice speeds were underestimated [Gui et al., 2020].

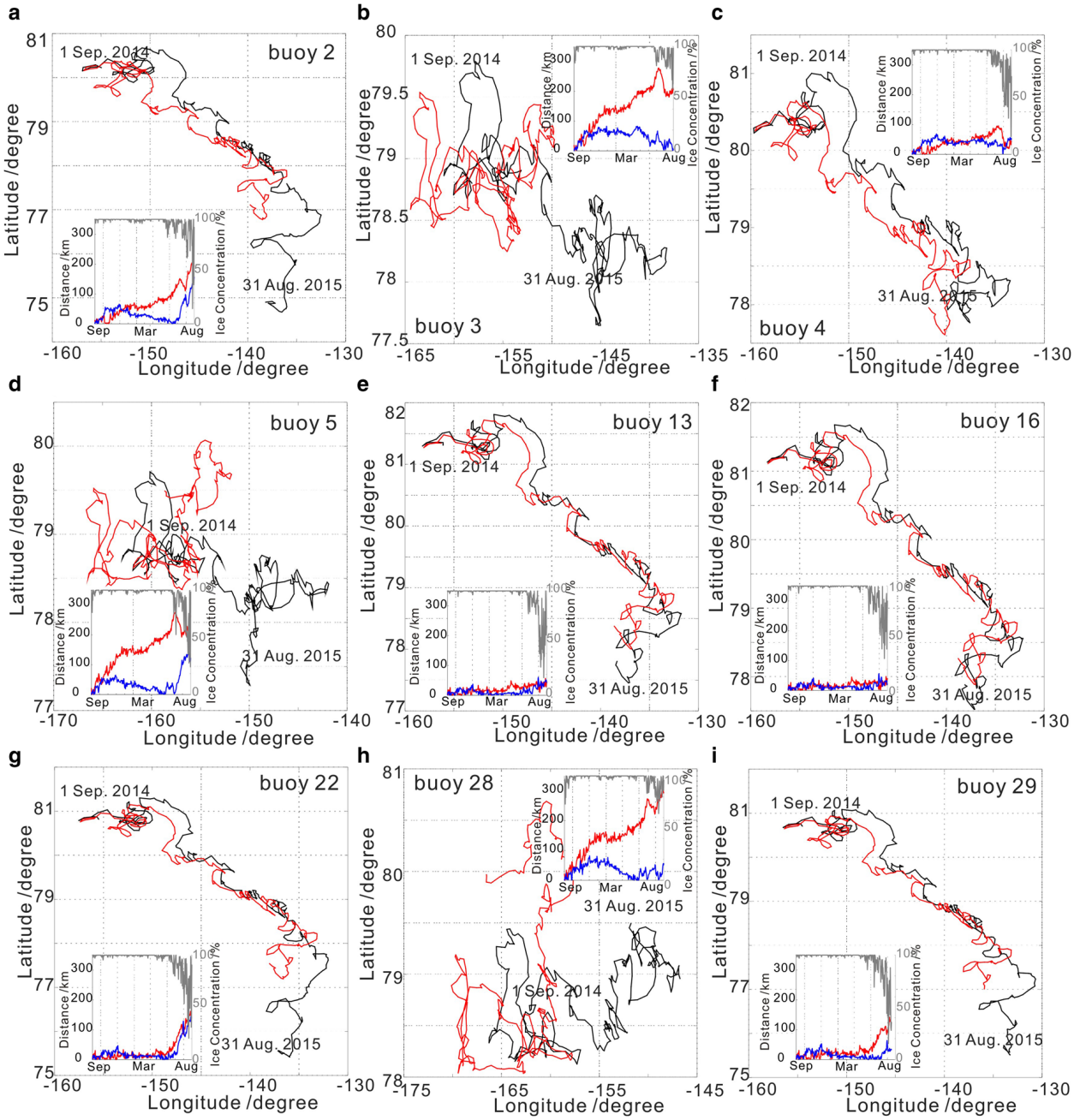

Figure 2.2: Polar Pathfinder derived sea ice drift trajectories (red line) compared to buoy drift trajectories (black line), during 2014/2015. The subplots show the Euclidean distance between the calculated drift paths and the observed buoy drift. Reproduced from Gui et al. [2020].

2.1.2 Meteorological Data

Meteorological Data is provided by the NOAA-CIRES-DOE 20th Century V3 reanalysis, available from January 1836 to December 2015, in 1 degree latitude by 1 degree longitude resolution. The data is plotted with the NOAA 20th Century Reanalysis Monthly Composites tool (https://psl.noaa.gov/cgi-bin/data/composites/plot20thc.v2.pl). Analysis in section 3.2 and section 3.1 plot meridional wind at 10 m, zonal wind at 10 m, and sea level pressure. Each image is the seasonal average of the chosen variable for the selected years. The NOAA Composites tool can also be used to find the difference between seasonal means for sets of years. The tool limits the number of years in a composite to 20, but this limit can be overcome if the form data is sent through the URL. A copy of this code is attached in appendix A.2.

2.1.3 Climate Indices

NAO and AO monthly indices are calculated by the NOAA Climate Prediction Center [Center, b,a]. The index calculated for each year is the average of the monthly indices for September to October of the next year. Positive indices are greater than one standard deviation above the mean, while negative indices are more than one standard deviation below the mean. All other indices are neutral.

2.2 Ice Tracking Algorithm

We used the Polar Pathfinder Dataset to track pixel-sized (25 km^2) parcels of sea ice at daily time steps, over the course of one year. To keep track of the ice's location, we used three location vectors:

1. C_m, the ice parcel's precise location in meters. The coordinate system's origin is the

north pole, and the axis extend up to 4512.2 km in any direction (covering the same area as the polar pathfinder EASE-grid).

2. C_r, the projection of the EASE-grid in meters. The coordinate system's origin is the north pole, with the axis extending up to 4512.2 km in any direction. The axis is not continuous, and there are only discrete values for each pixel in the EASE-grid.

3. C_p, the ice parcel's location relative to the 361 x 361 EASE-grid (fig. 2.1). This is the EASE-grid pixel containing the majority of the ice parcel, whose true location is recorded by C_m.

The pixel containing the majority of the parcel, C_p, is used to determine the velocity of the ice for that day. The majority pixel is calculated as:

$$C_{p_x} = min(C_{m_x} - C_{r_x})^2$$

$$C_{p_y} = min(C_{m_y} - C_{r_y})^2$$

The velocity belonging to C_p in the pathfinder dataset determines C_m for the next time step, or day. The U and V velocity components are used. We compute ΔC_m as:

$$\Delta C_{m_x} = u\Delta d$$

$$\Delta C_{m_y} = v\Delta d$$

Where d is the number of days that have passed. We calculate ice motion for every day, beginning of September 15th of one year and ending 270 days later on June 25th of the next.

If there is no ice in C_p, the velocity component in the pathfinder dataset is "NaN". This happens when the ice concentration in the pixel is $< 15\%$, or the pixel does not contain enough open water for sea ice motion calculations (e.g., in the Canadian Archipelago)

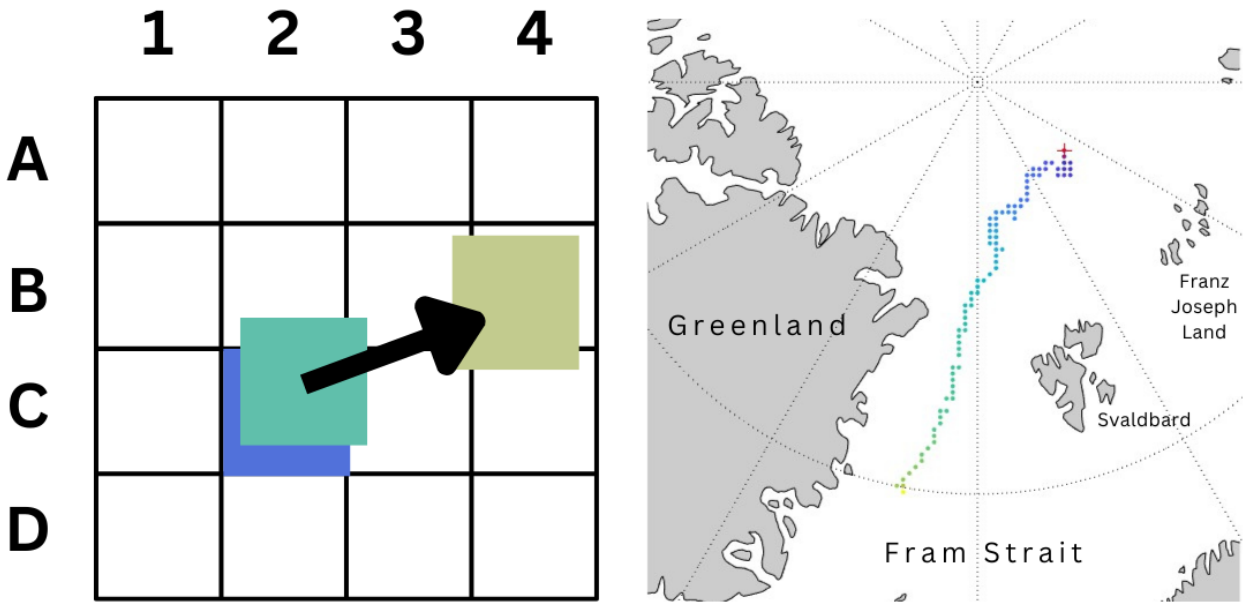

Figure 2.3: Illustrative examples of the ice tracking algorithm. (Left) The ice tracking algorithm shown on a small grid. The colored squares are the true ice parcel locations, recorded by C_m. Ice motion vectors are determined by C_p, the pixel containing the majority of the ice parcel. (Right) Drift trajectory for ice originating in the pixel at the red "+," over the course of one year. Bluer pixels show C_p at the beginning of the tracking period, while yellower pixels show C_p at the end of the period.

[Tschudi et al., 2016]. When this occurs, the current implementation of the algorithm sets the velocity of the ice to 0 and waits for the concentration of the pixel to increase enough for an ice motion measurement, or for the tracking period to end. This means that the original parcel of ice could melt completely, new ice could drift into the pixel containing the parcel's last location, and tracking would resume as before. This is not the intended behavior of the algorithm, which is meant to only track one ice parcel throughout a season. Simple changes to the ice tracking algorithm (booleans to check whether the parcel's current pixel turns "NaN" after tracking begins) would fix this issue. For more details on the implementation, see the ice tracking code attached in appendix A.3.

2.2.1 Example Ice Trajectories

Figure 2.3 illustrates a simplified example of the tracking algorithm. The ice originates in grid cell $C2$, shown by the dark blue square, where C_p and C_m refer to the same location on the grid. The ice motion vectors in $C2$ determine the parcel's motion for the day. After one day, the parcel drifts the the location shown by the turquoise square. The location of the turquoise square is recorded by C_m, but the C_p is still $C2$ since $C2$ contains the majority of the pixel. Therefore, we use the ice motion vector at $C2$ to determine the parcel's ice motion for the day. The next day, the parcel drifts to the location shown by the dark yellow square. Again, the pixel's precise location is recorded by C_m, but C_p is $B4$ since this pixel contains the majority of the parcel.

A real example of the ice tracking algorithm is shown on the right panel of figure 2.3. We track ice originating in the pixel marked by the red "+" over the course of one year. Bluer pixels show the C_p at the beginning of the tracking period, while yellower pixels show C_p at the end of the tracking period.

2.2.2 Validation

Drift trajectories are validated by comparison to "browse" images of polar pathfinder ice motion vectors. A more extensive validation of the pathfinder dataset for deriving drift trajectories of individual pixel sized ice parcels is done by Gui et al. [2020]. A similar analysis could be done for this ice tracking algorithm by comparing the calculated ice motion paths with known buoy tracks.

2.3 Machine Learning Algorithms Background

Clustering is a machine learning algorithm which sorts objects into groups. Similarity is defined by an empirical distance metric, such as the Euclidean distance formula, which is applied iteratively until the groups are sorted. The first clustering algorithm were published in the late 1960s [MacQueen et al., 1967], and today there are many different types of clustering algorithms, each optimized for different sorting tasks.

Clustering workflows are similar regardless of the algorithm chosen. First data is represented in matrix form, where each row in the matrix represents an object or "sample" to be sorted. The columns in the row are "features," or information about each particular sample. In this thesis, samples represents a year of ice drift originating in a particular location, while the features are the EASE-grid pixel locations of sea ice in its sample's year (e.g. 2.8). While the clustering algorithm is running, the chosen distance metric measures similarity between the vectors representing each sample. The evaluation of clustering algorithms depends on the desired outcome, but generally an effective clustering approach is reproducible (the algorithm would produce the same clusters if run again), with distinct differences between clusters and maximized similarity between items in the same cluster. While clustering algorithms are a powerful sorting tool, they often under perform when sorting high dimensional vectors, or matrices with many more features than samples [Beyer et al., 1999]. Dimensionality reduction algorithms are often needed to compress features into the most important components. The following sections give an overview of the three clustering algorithms and two dimensionality reduction techniques that are used to cluster drift trajectories in this thesis.

2.3.1 K-Means Clustering

K-means clustering is an algorithm first published in 1967 by James MacQueen, and is now one of the most widely used clustering algorithms [MacQueen et al., 1967]. The goal of k-means clustering is to group samples into K clusters. Similarity is measured using a pre-defined metric that measures the distance between feature vectors and the "centroid," or mean value of each proposed cluster. The squared euclidean distance is typically used as the distance function:

$$d(x, c) = (x - c)(x - c)'$$

In this formula, x is the input data vector and c is the vector representing the centroid of each cluster. The goal of K-means clustering is to limit the distance between the centroid of each cluster and all elements belonging to that cluster [Sharma, 2024, Divam, 2019]. The algorithm is outlined below:

1. Centroids are initialized randomly.
2. The distance between each data point and each centroid is calculated. Data points are assigned to the cluster with the nearest centroid.
3. The centroids are repositioned: they become the mean value of all the data points now assigned to that cluster.
4. Steps 2 and 3 are repeated until a stable clustering is created (data points are no longer repositioned because they are already assigned to the closest centroid) or for a set number of iterations.

The algorithm is also illustrated by figure 2.4, which shows the data before and after being sorted into clusters or "labelled." The data points in this figure, however, only have

two dimensions (features), plotted on the X and Y axis. This is much simpler than the ice drift data, where each sample has 200 or more features, and would be impossible to visualize.

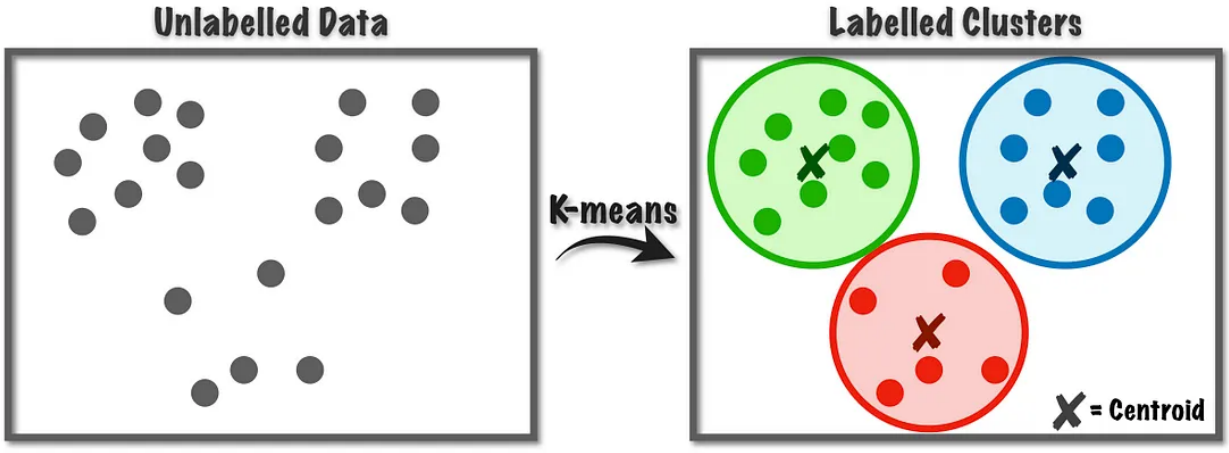

Figure 2.4: The K-means clustering algorithm illustrated for a two dimensional dataset. The centroids (shown with the black "x") are iteratively repositioned and data points are assigned to the closest centroid until a stable clustering is reached. Reproduced from https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

One drawback of K-Means clustering is that the number of clusters, K, must be set before the algorithm runs. To determine the optimal K, K-means clustering is run multiple times with various K. For each run, a metric is computed that measures the quality of the clustering. This is often a measure of variance like the within-cluster sum of squares (WCSS), or the sum of the euclidean distance formula applied to every point in the clustering. The number of clusters (K) is plotted against the variance metric, creating an "elbow" plot. As the number of groups increases, the variance decreases because points are closer to their centroids and in smaller clusters. However, the rate of decrease in variance slows after K becomes sufficiently large, creating the "elbow" in the plot. The optimal number of K is where this elbow is [Bholowalia and Kumar, 2014].

K-means clustering forms circular groupings, where each item is within some radius

distance from the centroid (fig. 2.5). K-means clustering does not perform well when there are outliers, which pull around the centroids and prevent optimal groupings from forming. K-means clustering also performs poorly on data with more features than samples, which is a common weakness among algorithms that rely on distance functions, and is further discussed in section 2.3.4. K-means clustering was chosen to sort years of similar sea ice drift since drift in each year should be similar to the mean sea ice drift in the group (section 2.4.1). However, k-means clustering is only effective at clustering sea ice drift data once the outliers are removed and the matrix dimensions are reduced (see section 2.4).

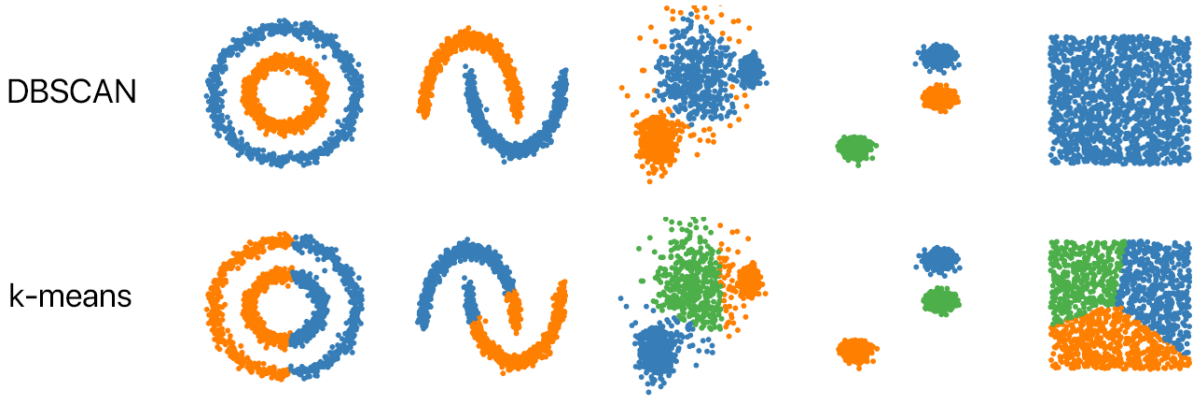

Figure 2.5: Comparison of DBSCAN and K-Means clustering. DBSCAN can capture clusters of arbitrary shapes, while the k-means algorithm forms circular clusters. Reproduced from https://github.com/NSHipster/DBSCAN.

2.3.2 K-Mediods Clustering

K-medioids clustering is like k-means clustering's cousin, similar except for the way in which the center of each cluster is determined. During initialization, data points are randomly chosen to be the initial centers, called medioids. The data points are assigned to the medioids with the distance function (just as they are to the centroids in k-means clustering), and the

updated medioids are the median data point in each cluster (instead of the mean). Similar to k-means clustering, samples are iteratively assigned to medioids and medioids are adjusted, until there are no changes in medioids, or until a predefined number of max iterations is reached.

K-medioids is more robust to outliers than K-means clustering, since outliers are far from other data points and less likely to be chosen as medioids. This makes k-medioids optimal for small data sets where medians are more meaningful than means [Arora et al., 2016]. This algorithm is used in section 2.4.2 since there are relatively few samples (years of ice drift) compared to features. K-medioids created more stable, repeatable clusters compared to k-means. However, k-means may have been more appropriate since there is significant variation among similar ice drift trajectories, meaning a cluster's average drift trajectory is likely more meaningful than a median.

2.3.3 DBScan Clustering

DBScan is another popular clustering algorithm introduced in 1996 by Ester et al. [1996]. Unlike K-Means or K-Medioids clustering, DBScan finds clusters of arbitrary shapes and is not affected by outliers. There are several key parameters used by the DBScan algorithm, including:

1. **Epsilon (ϵ):** the distance threshold at which points are considered neighbors.
2. **Minimum points (MinPts):** the minimum number of points required to form a cluster.
3. **Core points**: Points with at least MinPts neighbors. These are at the center of the clusters.

4. **Border points**: Points within ϵ distance of a core point, but with fewer than MinPts neighbors.

5. **Outliers**: Points that are neither core nor border points.

The parameters MinPts and ϵ are initially passed to DBScan, and these are carefully determined in the context of each data set and the desired clustering. Once these parameters are selected, the algorithm calculates the distances between all points and identifies core, border, and outlier points based on ϵ and MinPts. Clusters are formed by connecting core points that are within ϵ distance from one another. Border points join their core point's cluster, and outliers are not included in any cluster.

DBScan is often used over other clustering techniques because the number of clusters are automatically determined based on density. Unlike k-means clustering, density based clusters can take on any shape in order to encapsulate pockets of density (fig. 2.5). DBScan is also unaffected by outliers, since outlier points are kept separate from the clustered points. However, DBScan is very sensitive to the choice of ϵ and MinPts. Additionally, border points on the edge of multiple clusters can be within ϵ of core points in multiple clusters, meaning their assignment to one of the clusters is arbitrary.

DBScan is used in this project to identify unusual ice drift patterns before ice drift data is given to k-means or k-medioids clustering. This is done by choosing a small ϵ and MinPts, so that only the years with the most unique drift trajectories are identified as outliers. These outliers are removed from the data set before clustering begins.

2.3.4 "Curse of Dimensionality" and PCA

One concern with almost any machine learning algorithm is the "Curse of Dimensionality." This is the observation that high dimensional spaces have properties that do not occur in lower dimensional spaces. In particular, traditional distance measures (like Euclidean

distance) become less meaningful in high dimensional spaces, because the distance between the farthest points and the nearest points approach the same value. If the vectors that are being clustered are too high in dimension, the distance function will not meaningfully differentiate between points that are far away and points that are close [Beyer et al., 1999]. Since distance functions can not tell the difference between samples, meaningful clusters cannot be created. Therefore, lower dimensional datasets are preferred for most clustering algorithms background.

There are several techniques for reducing the dimension of a dataset. For example, feature selection is using only the most informative features in the clustering analysis, identified either with statistical techniques or domain knowledge. This is employed in section 2.4.1 when choosing to cluster only the last locations of the ice parcels on their ice trajectories. The assumption is that the last locations of the ice parcels reflect the ice's trajectory for each year. Another technique is feature aggregation, which is combining multiple related features into a single feature. This idea is explored when clustering the complete ice trajectory paths in section 2.4.2: a 1-dimensional convolution averages together nearby coordinates on the ice path, reducing the length of the ice trajectory by a factor of 3.

The size of the data set can also be reduced with more complex mathematical techniques. Principle Component Analysis (PCA) is often used to reduce a dataset to the principle components while capturing a significant portion of the dataset's variance (usually 95% or more). PCA works by creating new variables that are linear functions of the original data, expressing the principle components of the data in fewer dimensions. However, this assumes that the underlying structure of the data is linear [Abdi and Williams, 2010]. Even with linear data, PCA should be used with caution: the 5% or less of variation which is not captured may contain crucial information needed to create distinctive clusters. PCA is used in section 2.4.1 to reduce the matrix of sea ice drift end locations. The differences in end locations between years are distinct enough that this small amount of variation shouldn't

make a difference. However, PCA is not used when the entire drift trajectories are considered in section 2.4.2. The full drift trajectories often begin similarly, even in different years, with differences accumulating over time, meaning the percent of variation not captured by PCA is more important. Therefore, PCA is less ideal for clustering the full drift trajectories.

2.3.5 Dimensionality Reduction Deep Autoencoders

An alternative to traditional dimensionality reduction techniques like PCA are deep autoencoders, also known as auto-associative neural networks or bottleneck networks. Unlike PCA, these represent complex, nonlinear data.

Autoencoders are neural networks. Generally, neural networks are similar in structure to a brain, where nodes are aligned in layers, and are connected to other layers of nodes. Information flows through the nodes, where linear algebra operations take place, before passing to the next layer. Different "weights" are associated with each node, which update as input data flows through the neural network. The goal is to tweak the weights until the output of the network minimizes a chosen metric.

Autoencoders consist of two main parts: an encoder and a decoder (figure 2.6). The encoder reduces the input to a lower dimension, while the decoder uses the reduced representation to reconstruct the original input to the encoder. The goal is to minimize the error in reconstruction, meaning the output of the encoder contains the most essential information from the input. The reduced representation from the encoder is the low dimensional representation that can be used for clustering [Lange and Riedmiller, 2010].

While autoencoders can learn complex representations, they are less straight forward than traditional dimensionality reduction methods like PCA. For one, autoencoders are more computationally expensive. Unlike PCA, they must be tweaked for each matrix they reduce, depending on the complexity of the data. More complex data will require more

neurons in each layer to fully capture the variance in the data. However, too many neurons on more simple data may result in the autoencoder learning patterns that don't exist. This "overfitting" can be combated by adding drop-out layers, in which neurons are randomly toggled on and off during the training. Even with drop-out layers, tweaking the number of neurons for each layer is often necessary to learn the optimal low-dimensional representation.

Autoencoders are used in section 2.4.2 to reduce the matrix containing the full trajectory paths for over 4000 pixels. This method is used since the data is nonlinear and complex. However, tweaking the autoencoder was time intensive, and as a result was only used for one location.

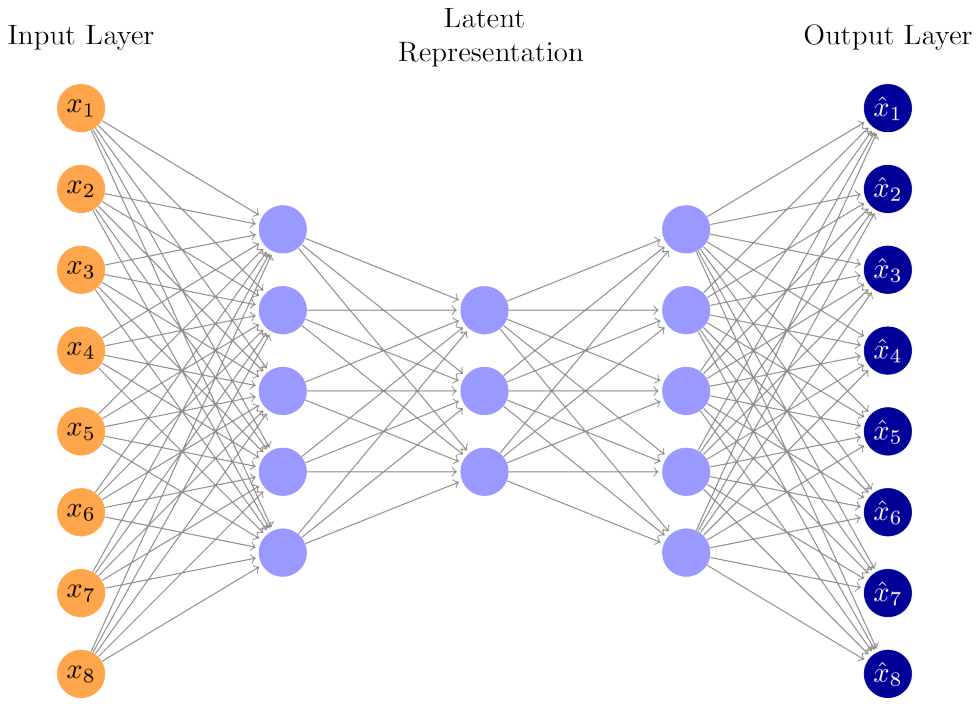

Figure 2.6: Autoencoder illustration. The input/output layers are the size of the original input, while the latent representation is the reduced size input. Reproduced from https://tikz.net/autoencoder/

2.4 Clustering

We used two techniques for clustering. The first considers only the locations where the ice tracking starts and ends. PCA reduces the data, outliers are removed with DBSCAN, and k-means does the clustering. The second approach uses the full ice path trajectories. Data is reduced with an autoencoder, outliers are removed with DBSCAN, and clustering is done by K-medioids. These approaches were developed after experimentation with different clustering algorithms, and knowledge about the strengths and weaknesses of these algorithms on high dimensional data.

2.4.1 Last Ice Locations Clustering

The first clustering approach only uses information about where ice tracking begins and ends each year. This approach sets up a data processing structure which is useful for more complicated clustering approaches.

2.4.1.1 Data Preparation

A 10 x 10 pixel region is selected as the starting location for the tracked ice parcels. Each ice parcel originating in this box is tracked from September 15th of one year until June 25th (270 days later). To reduce the dimension of the data given to the clustering algorithm, these drift paths are reduced to only the last pixel in the path. Therefore, the clustering algorithm knows where the pixel begins each year (in the boxed region) and where each pixel ends up (figure 2.7).

For the starting location, the pixel coordinates of the last locations of ice parcels for one year are turned into a 1D vector. For a single year, there are 100 end locations. This means one location will have 41 1D vectors (one for each year), each storing 100 coordinate pairs (x, y in the EASE-grid) of end locations. The resulting array is illustrated by figure 2.8.

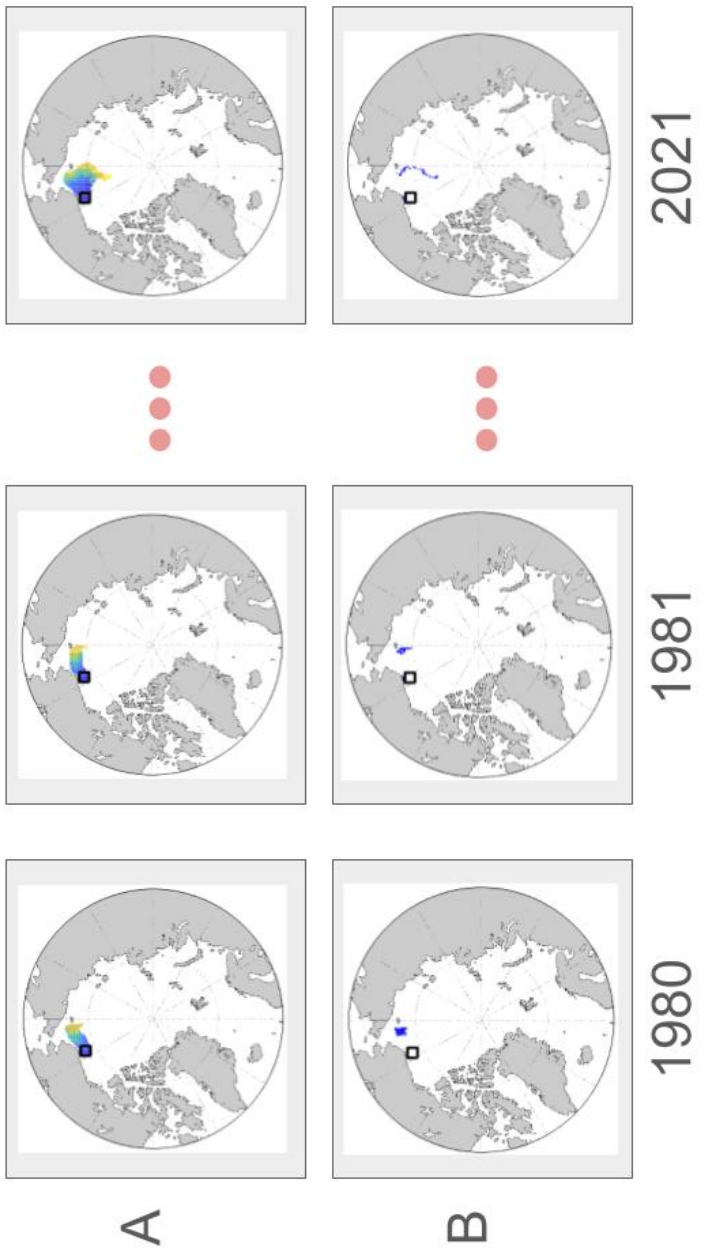

Figure 2.7: Drift paths for ice originating in the boxed region, tracked from September 15th of one year until September 14th of the next or when the ice melts, whichever comes first. Drift was computed for the boxed region over all years of data, from 1980 to 2021. (A) Blue pixels show the ice in the beginning of the tracking time period (fall), while yellow pixels show the ice at the end of the time period (spring/summer). (B) Only the last location of the ice for that year's drift paths are plotted.

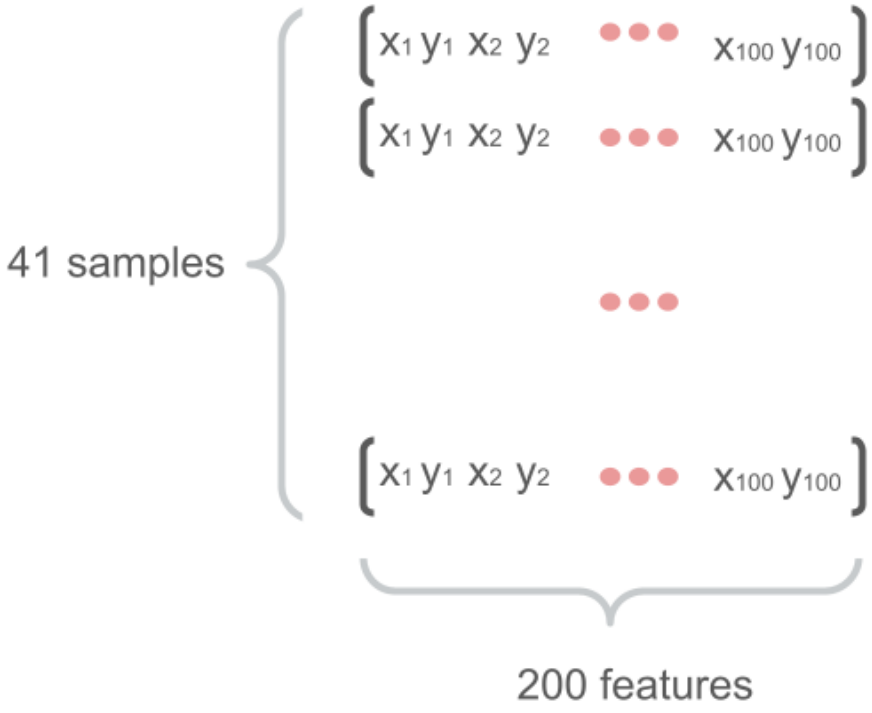

Figure 2.8: The matrix prepared for clustering for one location. Each row in the matrix is 1 year of coordinates showing where the ice that originated in the boxed region ended up. In total there are 41 years of data, each with 100 coordinate pairs showing where ice ended up.

We apply PCA analysis to the ending matrices for each location: 5 principle components capture 95% of the variation. This reduces the dimensions of each sample from 200 to 6, allowing for more accurate clustering and nearest neighbor detection.

Once PCA is applied, DBSCAN identifies years whose end locations that are outliers. A relatively large ϵ and a small MinPts identifies years that are most unlike any other years for that location according to the Euclidean distance metric. An example of identified outliers is shown in figure 2.9.These years of data are completely left out of the clustering analysis. Without removing these outliers, the k-means algorithm's clusters are only marginally more distinct than a random clustering.

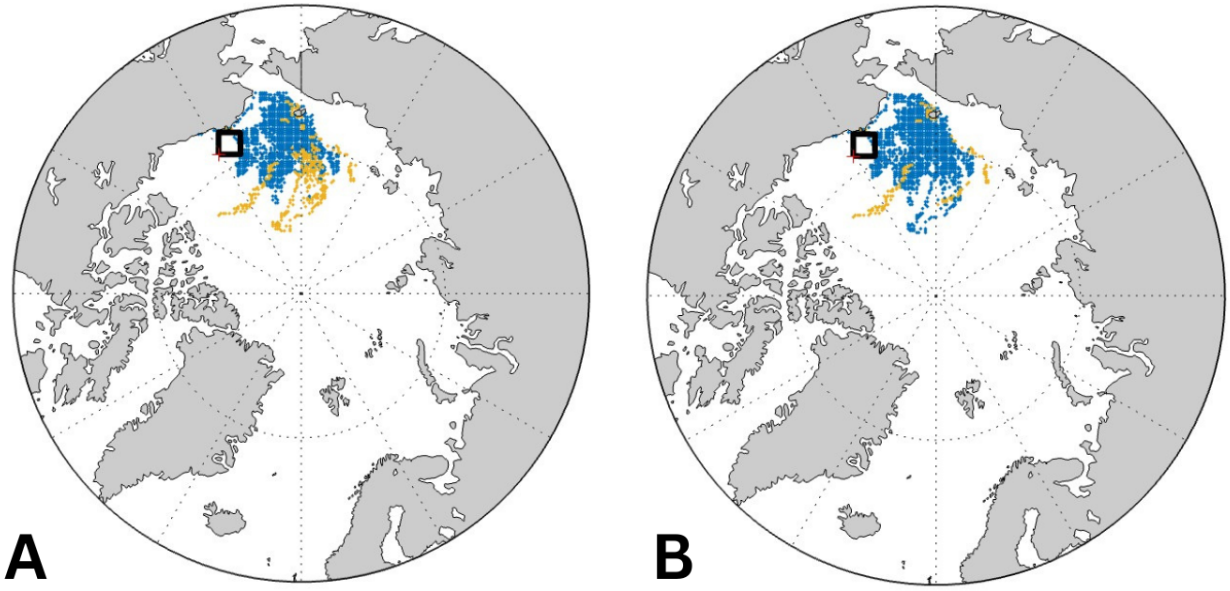

Figure 2.9: DBSCAN identifies outliers using the Euclidean distance function. Composite of all drift end locations originating in the Beaufort location. The outlier groups of drift are shown in orange. (A) Smaller epsilon used, meaning more outliers are detected. (B) Larger epsilon used, so only the most unique groups are identified as outliers.

2.4.1.2 Human Clustering Benchmark

In order to determine whether a clustering algorithm is effective at sorting sea ice drift trajectories, we clustered the 41 years of end location pixels for one location. The goal was to create an upper bound for what distinct clusters should look like by matching the shape, orientation, and location of each year's end location pixels. The result is shown in figure 2.10. The members per group vary from 2 to 10. For analysis of the years in the context of meteorological data like sea level pressure or wind velocity, larger groups are more useful. While these groups are too small for the rest of the analysis in this project, the results show that distinct groups of ice drift exist.

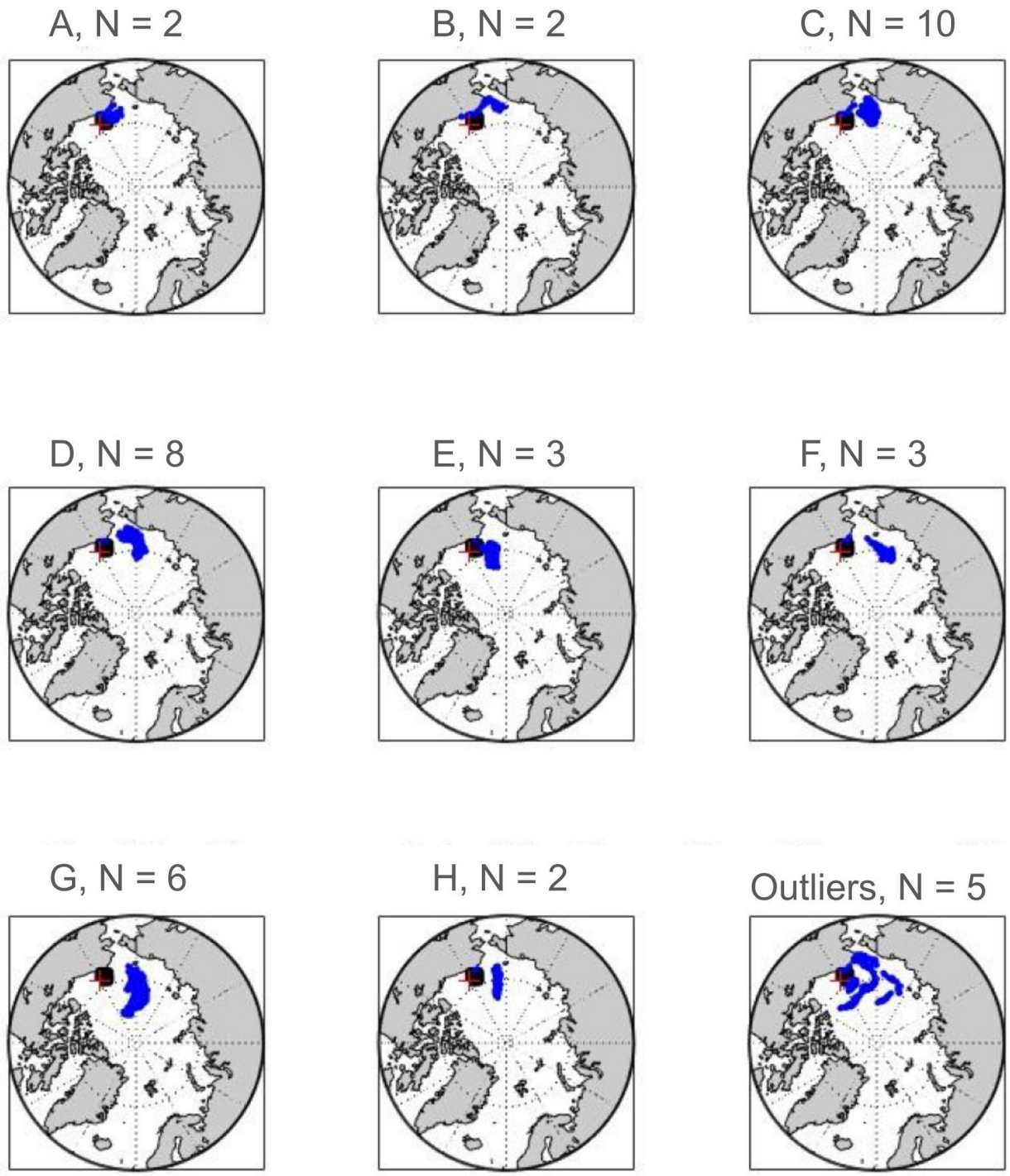

Figure 2.10: Manual clustering of ice end locations originating in the black box. This is an upper bound on how distinct clusters can be.

2.4.1.3 Applying K-Means Clustering to Find Years of Similar Drift

We prepared the data as described in section 2.4.1.1 and illustrated by figure 2.8. We created one matrix for each location, where each "sample" in the matrix contains the end locations of the ice originating in that location. For each location, our goal is to sort each sample into groups with the most similar sea ice end pixel locations. These groups should be circular, meaning the distance of each vector to the center of the cluster (what the "average" year of drift looks like in that cluster) is minimized. Since k-means clustering creates circular groups, we chose this algorithm. To determine the optimal number of k for clustering, we created an elbow plot (see section 2.3.1 for more information about choosing the optimal k). K-means clustering puts all the years of drift belonging to one location into k groups, except for the outlier years identified in the DBSCAN step of the data preparation. Once these clusters are created, we compared the clusters and associated climate patterns across different locations (see section 3.1 and 3.2).

2.4.2 Full Trajectory Clustering

Instead of only giving the clustering algorithm information about where the ice ends up, we use the full ice path trajectories. Section 2.4.2.1 describes how we represent the full path data as a matrix. If the full path trajectories can be successfully clustered, these clusters would be more representative of similar years of ice drift since the entire path is considered. In practice the full path trajectories are massive and cannot be effectively clustered with the Euclidean distance function because of the "curse of dimensionality" (section 2.3.4). To overcome the curse, we reduced the dimensions of the dataset with deep autoencoders (section 2.4.2.2). Data reduction techniques differ for this clustering approach, but the clustering methods are similar to clustering methods used on only the last ice locations in section 2.4.1.

Figure 2.11: Full path trajectory matrix representation. Each sample holds drift trajectories for all 100 pixels in the location's starting box. The 1st pixel's drift trajectory is placed into a row, followed by the 2nd, until all the pixels are placed into the row.

2.4.2.1 Trajectory Data Representation

We select a 10 x 10 pixel region to start tracking sea ice, same as the region selected when just preparing the matrix of end locations in section 2.4.1. We track the ice from September 15th of one year until June 25th, 270 days later. The full trajectory information is recorded, displayed in figure 2.7A. The clustering algorithm considers entire drift paths when creating groups of similar drift.

For one boxed location and one pixel-sized drift trajectory, we turn the pixel coordinates of each point on the trajectory into a 540 x 1 vector: one (x, y) pair of coordinates gives the pixel location for one day, and these are concatenated one after another for 270 days. We create these 540 x 1 vectors for all 100 pixels in the starting box, and concatenate them together to create a 54,000 x 1 vector. Since there are 41 years of data, the final matrix is 54,000 x 41. The data representation is visualized by figure 2.11. This is a massive feature to sample ratio, meaning the unreduced matrix cannot be meaningfully clustered.

2.4.2.2 Dimensionality Reduction and Clustering

The matrix created in the previous section (2.4.2.1) must be reduced in dimensions but we do not use PCA (as with the end locations matrix in section 2.4.1). This is because the internal structure of the data is nonlinear. Instead we use an autoencoder, which is able to reduce the size of complex, nonlinear data (see section 2.3.5 to learn more about autoencoders). The autoencoder is split into two parts, the encoder and the decoder. The job of the encoder is to gradually reduce the input data into a low level representation without losing important information [Lange and Riedmiller, 2010].

We have only applied the autoencoder to the Beaufort location so far. In the future, the autoencoder can be applied to all the locations, but doing so is time intensive because the autoencoder's structure must be customized for each location. The first layer is always fully connected, with one neuron for every feature. The next layers are smaller than the first, and their purpose is to slowly reduce the dimensions of the input from the first layer without losing important information. We add dropout layers (randomly turning a certain percentage of the neurons off in each training run) to reduce over training.

For the Beaufort location, the first layer contains 54,000 neurons, one for each input feature. The next layer contains 800 neurons, and uses 40% dropout. This is followed by a layer with 256 neurons, using 20% dropout. The last layer has 32 neurons and is the bottleneck layer, or the reduced representation of the original matrix. Each hidden layer uses the ReLu activation function. The dropout decreases near the bottleneck layer since the information that is passed through these inner layers is already compressed significantly, and therefore is more likely to be part of the low-dimensional representation of the data, as opposed to learned noise. The decoder is the reverse of the encoder, with hidden layers of size 256 and 800.

To evaluate the performance of the autoencoder, we set aside 20% of the matrix as

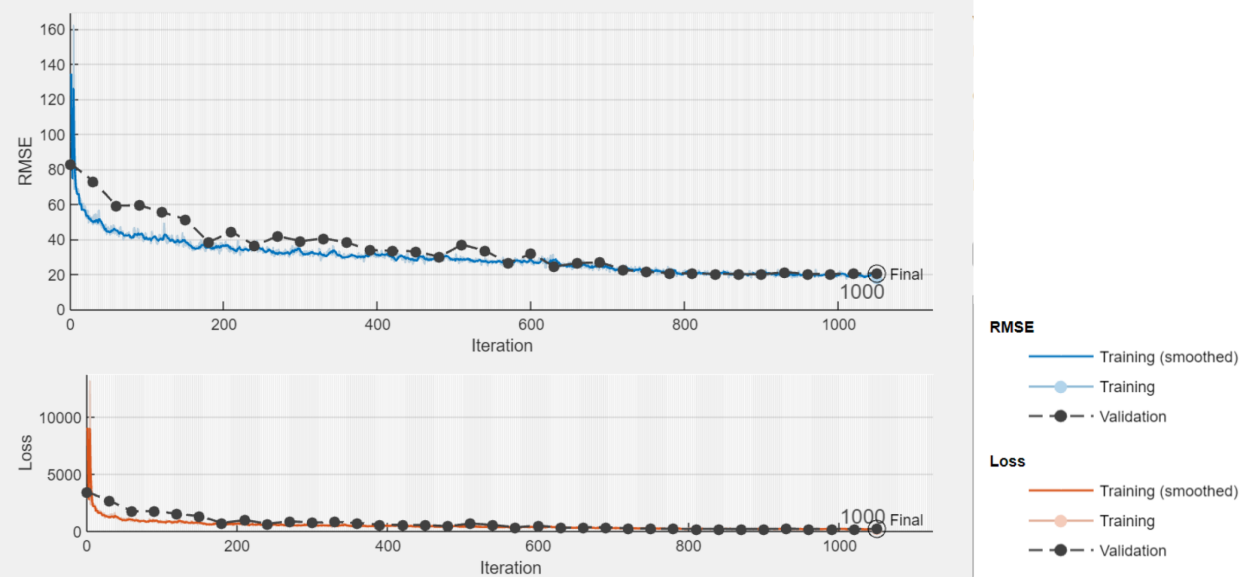

Figure 2.12: Loss and Root Mean Squared Error (RMSE) over 1500 training iterations converge for the training and validation sets.

validation data, while leaving the other 80% as training data. We plot the root mean squared error (RMSE) between the original and reconstructed matrix for each training iteration. We tweak the number of neurons and the dropout percentage so that the total RMSE is minimized, and the RMSE of the training and validation data converges (figure 2.12). After 1500 training iterations, the loss and RMSE on the training and validation sets converge, indicating that the autoencoder learned a low dimensional representation of the data without over-fitting.

The reduced representation of the data (the bottleneck layer of the autoencoder) is a 32 x 41 matrix. Similar to how the end locations matrix was clustered in section 2.4.1, we apply DBSCAN to identify outliers. We create an elbow plot to identify the optimal number of K. We use k-medioids to create clusters, instead of k-means, to create stable clusters. However, k-means may have been more appropriate since the median drift trajectory year is less meaningful than the mean. See section 2.3.2 for more information about k-medioids

clustering.

The elbow plot suggests 2 is the optimal number of clusters for this data. The results of the clustering with k=2 are shown in figures 2.13 and 2.14, where the Beaufort A cluster tends to drift towards the pole and the Beaufort B cluster tends to drift towards Siberia.[1] The results of the full path trajectory are not discussed further but should be analyzed in future work (along with full path trajectory clusters for the other 3 locations).

[1]Plotting the full path trajectories for each year takes a lot of time. To be efficient, I created trajectory plots for each year in the Beaufort location and saved them as separate images in a folder. To create figures 2.13 and 2.14 (showing the trajectories of the years belonging to the cluster), I told MATLAB to plot these images. This made trying out the autoencoder with different layers a lot easier.

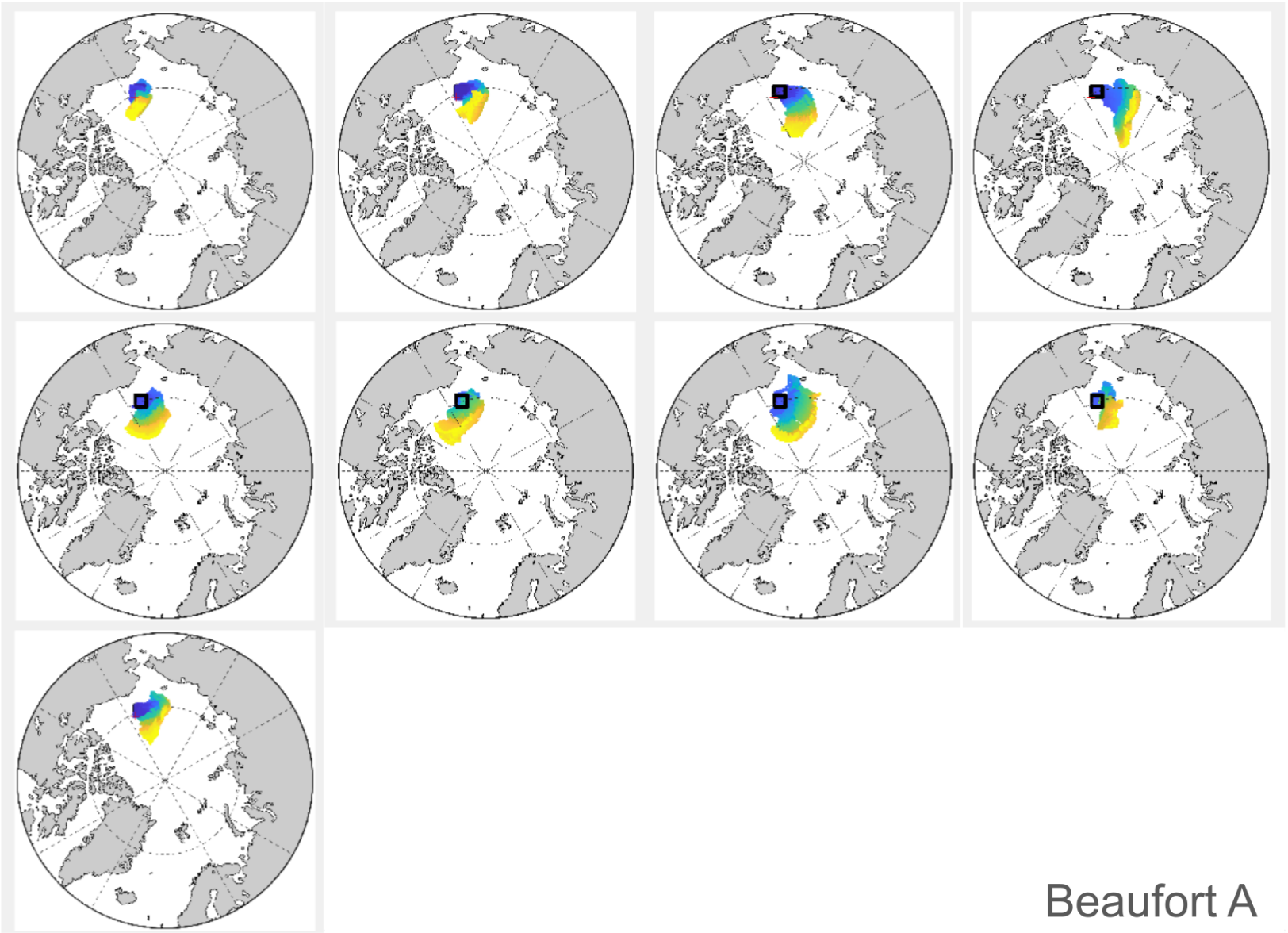

Figure 2.13: Beaufort A. Drift originates in the black box. Trajectories begin on September 15th (blue pixels) and end in October of the next year (yellow pixels).

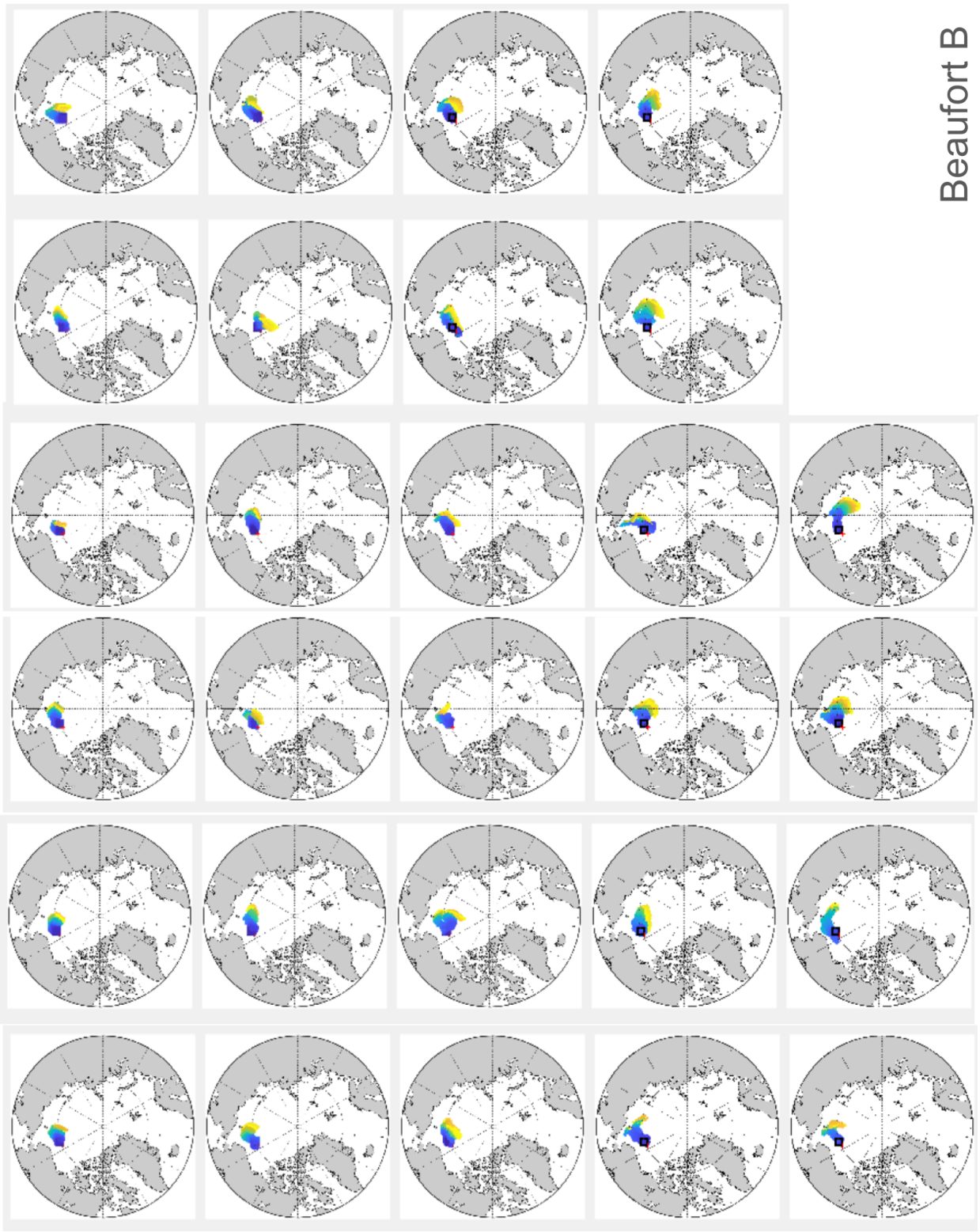

Figure 2.14: Beaufort B. Drift originates in the black box. Trajectories begin on September 15th (blue pixels) and end in October of the next year (yellow pixels).

Chapter 3

Results

The results below are only for the clusters created with knowledge of the end and start locations of the ice, as described in section 2.4.1.

3.1 Clusters

I analyzed four locations throughout the Arctic to find the years of most similar drift. These include two locations whose drift regime is mostly controlled by TPD (Laptev Sea and North Pole), and two locations whose drift regime is mostly controlled by the BG (Beaufort Sea and Ellesmere Island). The elbow plots for each location indicated that 3 clusters were optimal for each of these locations. K-means clustering creates clusters which appear as distinct as my manual benchmark clusters, although there are fewer of them. While smaller clusters have less intercluster variation, we chose larger clusters to facilitate the climate variable analysis.

Clusters for ice originating in each of the four locations are plotted as composites of the drift year end locations belonging to each group. For example, figure 3.2 (first row) plots the clusters created for the Laptev location. There are 18 years of drift that belong to Laptev A.

The end locations for each of these drift years are plotted together as the blue pixels under "Laptev A."

We considered two factors as possible associations with the observed clusters: climate variables, or changes in sea ice properties over time. For each cluster, we plotted average winter sea level pressure and winds. Meridional wind is positive when wind blows south to north, and negative when wind blows north to south. Zonal wind is positive when wind blows west to east, and negative when wind blows east to west. We approximated changes in sea ice properties by the years which belong to each of the clustered groups, with ice in more recent years being thinner and sparser in most locations throughout the Arctic [Zhang et al., 2012].

3.1.1 Laptev Sea

Clusters found for sea ice parcels originating in the Laptev location travel poleward in 3 distinct drift patterns (fig. 3.2). This is the approximate location where Nansen froze the Fram into sea ice in 1893 [Nansen, 1897]. Depending on the year, the speed of poleward drift varies dramatically. Ice in Laptev A moves only a few hundred kilometers from the original location, resembling the slow sea ice drift observed by Nansen. Ice in Laptev B drifts the farthest, taking only one year to drift by Franz Joseph Land (twice as fast as the Fram's voyage).

Laptev A has lower SLP across most of the Arctic compared to Laptev B and Laptev C (fig. 3.3). This weakens TPD, contributing to slow sea ice drift. Conversely, Laptev B and C have higher SLP, resulting in stronger TPD. All three clusters have low SLP in the Barents (fig. 3.2), which creates anticyclonic circulation off the coast of Scandinavia. However, Laptev B is associated with anomalously low SLP in the Barents, creating especially strong counterclockwise circulation which strengthens TPD. This contributes to the long

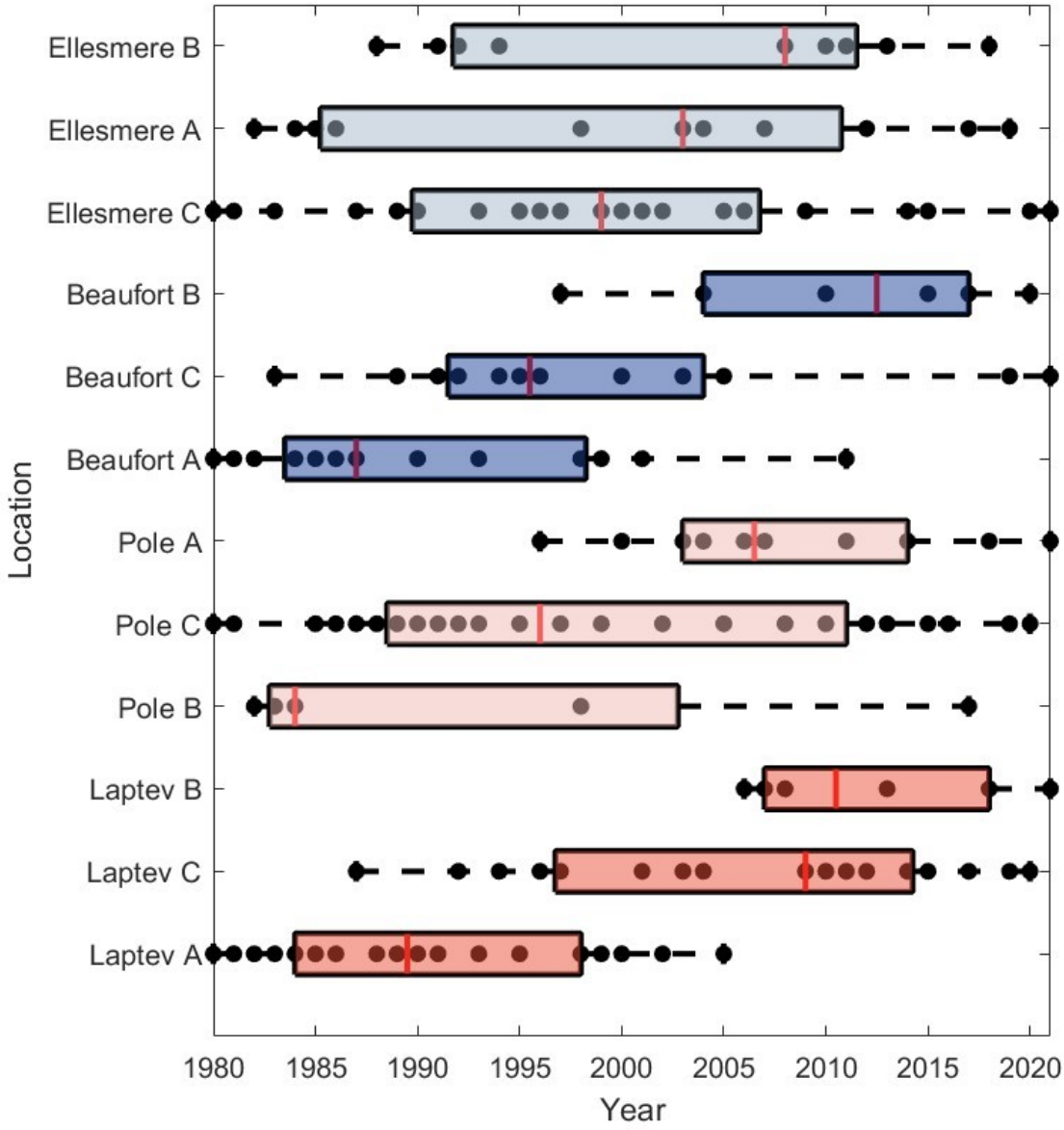

Figure 3.1: Box plots showing years belonging to each group. The black plotted points are the year of drift trajectories included in each cluster. The median for each cluster is shown in red. The groupings for locations where ice originates in the marginal ice zone (Laptev, Pole, Beaufort) are likely correlated with recent changes in ice properties. This is not true of the Ellesmere location, where thick multiyear ice remains.

drift trajectories in Laptev B compared to the other clusters. Laptev C is associated with anomalously high SLP in the Barents, weakening the counterclockwise circulation off the

coast of Scandinavia. As a result, TPD is weaker in C compared to B, contributing to the shorter drift paths in C.

The years in the groups are associated with changes in sea ice properties, such as thickness and total ice extent (how "locked up" the Arctic is). Section 1.2.3 discusses how these changes in ice properties are increasing sea ice drift speeds throughout the Arctic. The Laptev location is in the marginal ice zone, where these changes in sea ice properties are the most extreme [Zhang et al., 2012]. Figure 3.1 plots the distribution of drift years belonging to each of the Laptev groups. There is a clear age difference between the three groups: Laptev A contains the oldest years (median 1989.5), followed by Laptev C (median 2009), while every single year in Laptev B (median 2011) is younger than the years in Laptev A. This corresponds with how far the ice in each group traveled, with the older group traveling the least and the youngest group traveling the most. Given that the meteorological data in figure 3.3 partially explain the differences in drift patterns observed at the Laptev location, the Laptev groups are likely formed in response to recurring SLP patterns and changes in sea ice properties in recent decades.

3.1.2 North Pole

Clusters found for sea ice parcels originating in the Pole location are the least distinct, with Pole A and C appearing the most similar (fig. **??**). Pole B is distinct in that the ice only drifts a few hundred kilometers, while ice in Pole A and C drift out the Fram Strait.

The ice movements are not well explained by meteorological data (fig. **??**). Years belonging to Pole B are characterized by high SLP over the BG and low SLP over the Barents, with a strong pressure gradient across the central Arctic. This should enhance TPD, leading to faster ice motion out the Fram Strait. However, this is the opposite of what is observed: ice in Pole B moves substantially less than ice in Pole A and Pole C. Ice drift in Pole A and

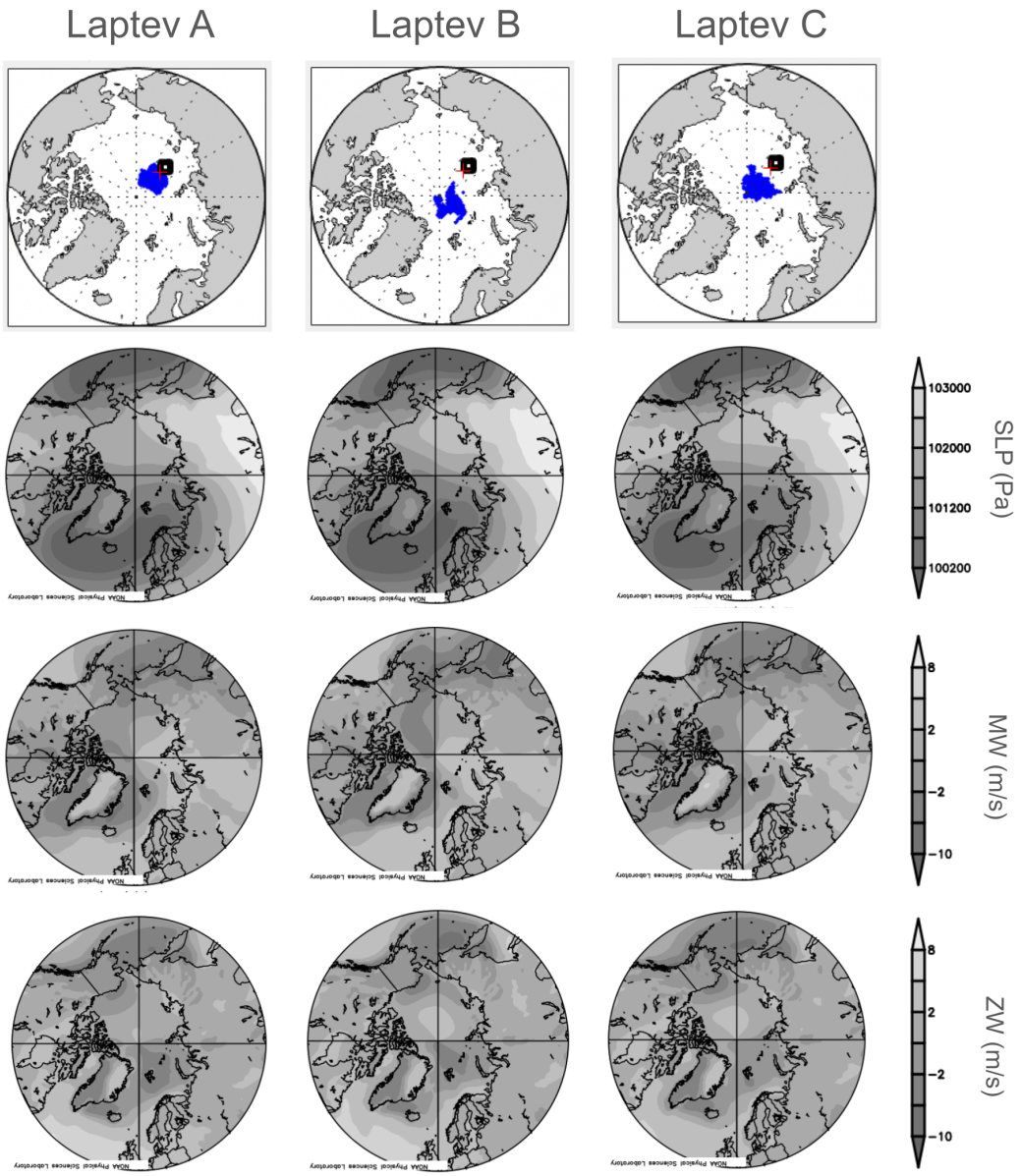

Figure 3.2: Meteorological data from the Laptev location. The black box shows where ice originates every year. Ice is tracked from September of one year until September of the next, or until the ice melts. The blue pixels show the end of the ice trajectory. Each map in the top row is a composite of all the ice end locations for each year in that group. Meteorological data is the composite average for all the years belonging to each group. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

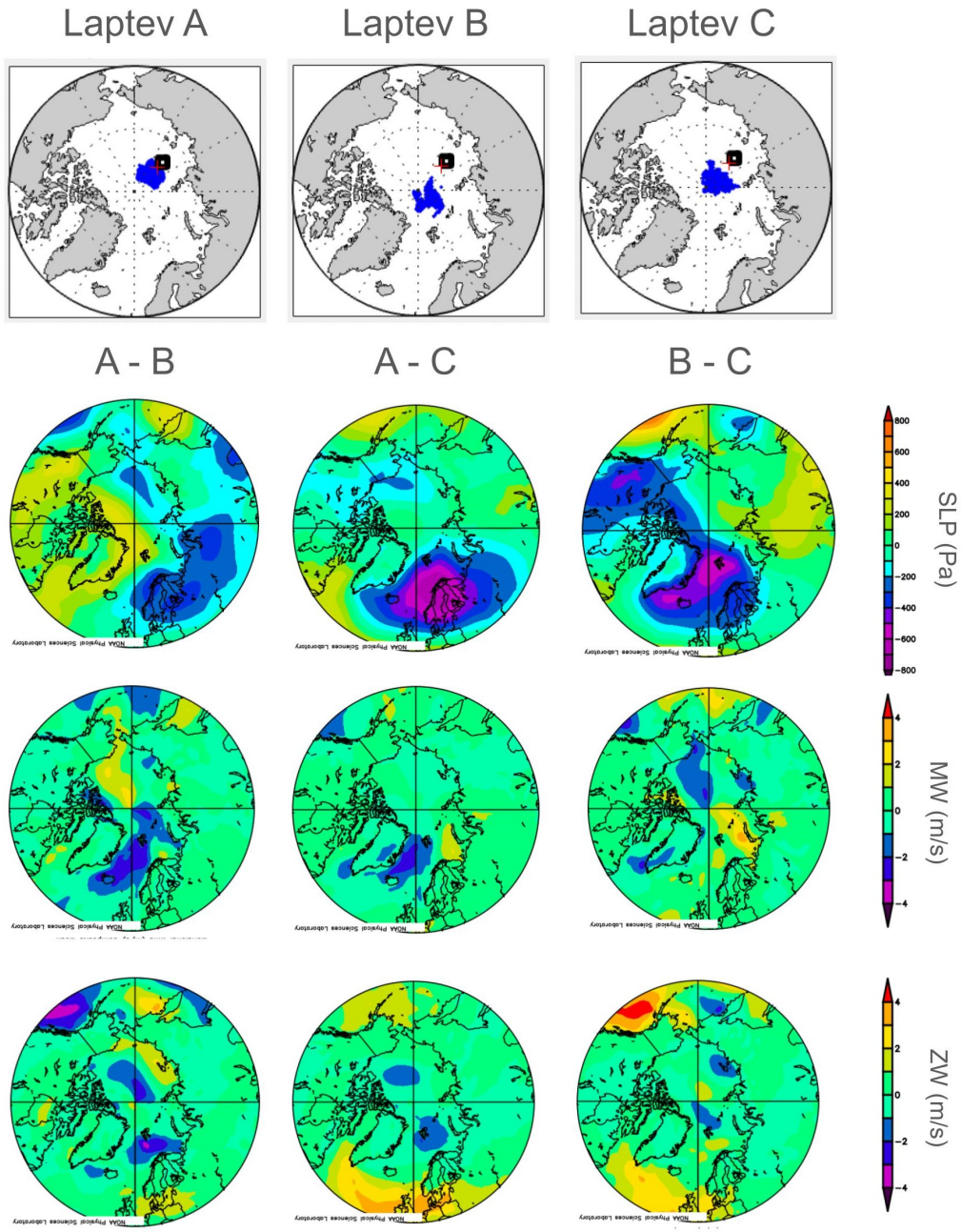

Figure 3.3: Meteorological data difference map from the Laptev location. The difference in meteorological data, for example "A - B", is the composite average for all the years belonging to A minus the composite average for all years belonging to B. Laptev B is associated with anamalously high pressure in the Barents, contributing to faster drift speeds compared to Laptev C. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

Pole C are similar, and these groups also have similar SLP, meridional, and zonal winds. Ice in Pole A moves farther than ice in Pole C, and there are no obvious differences in the meteorological data which would explain these differences.

Ice in the pole location originates in the TPD stream, where changes in ice properties in recent decades have accelerated sea ice drift in response to the same external forces, further explained in section 1.2.3. There is a clear age difference between the three clusters at the Pole location (fig. 3.1): Pole B contains the oldest years (median 1985), followed by Pole C (median 1996) and Pole A (median 2006.5). The first quartile of years in Pole A is younger than the 3rd quartile of years in Pole B, showing a significant age difference between these two clusters. Similar to the Laptev location (section 3.1.1), age corresponds to how far the ice in each group travels, with the oldest group traveling the least and the youngest group traveling the most. Unlike the meteorological data, changes in sea ice properties in recent decades explain drift patterns observed in the Pole location.

3.1.3 Beaufort

Clusters found for sea ice parcels originating in the Beaufort location represent 3 distinct drift patterns, with ice drifting in the direction of the BG (fig. 3.6). Ice in the Beaufort A and B groups drift towards the East Siberian Sea, while ice in Beaufort C drifts poleward. The longest drift trajectories belong to Beaufort B, with ice spreading out in a crescent shape rather than staying packed tightly together, as in A and C.

The meteorological data explains many of the differences in ice motion between the three Beaufort groups (fig. 3.7). Ice motion in Beaufort A is associated with of a strong BG, pushing ice from the Beaufort Sea along the Siberian coast. Higher pressure over the central Arctic in Beaufort A compared to Beaufort B, and especially Beaufort C, creates strong clockwise circulation in the BG. Beaufort A is associated with anomalously low pressure over

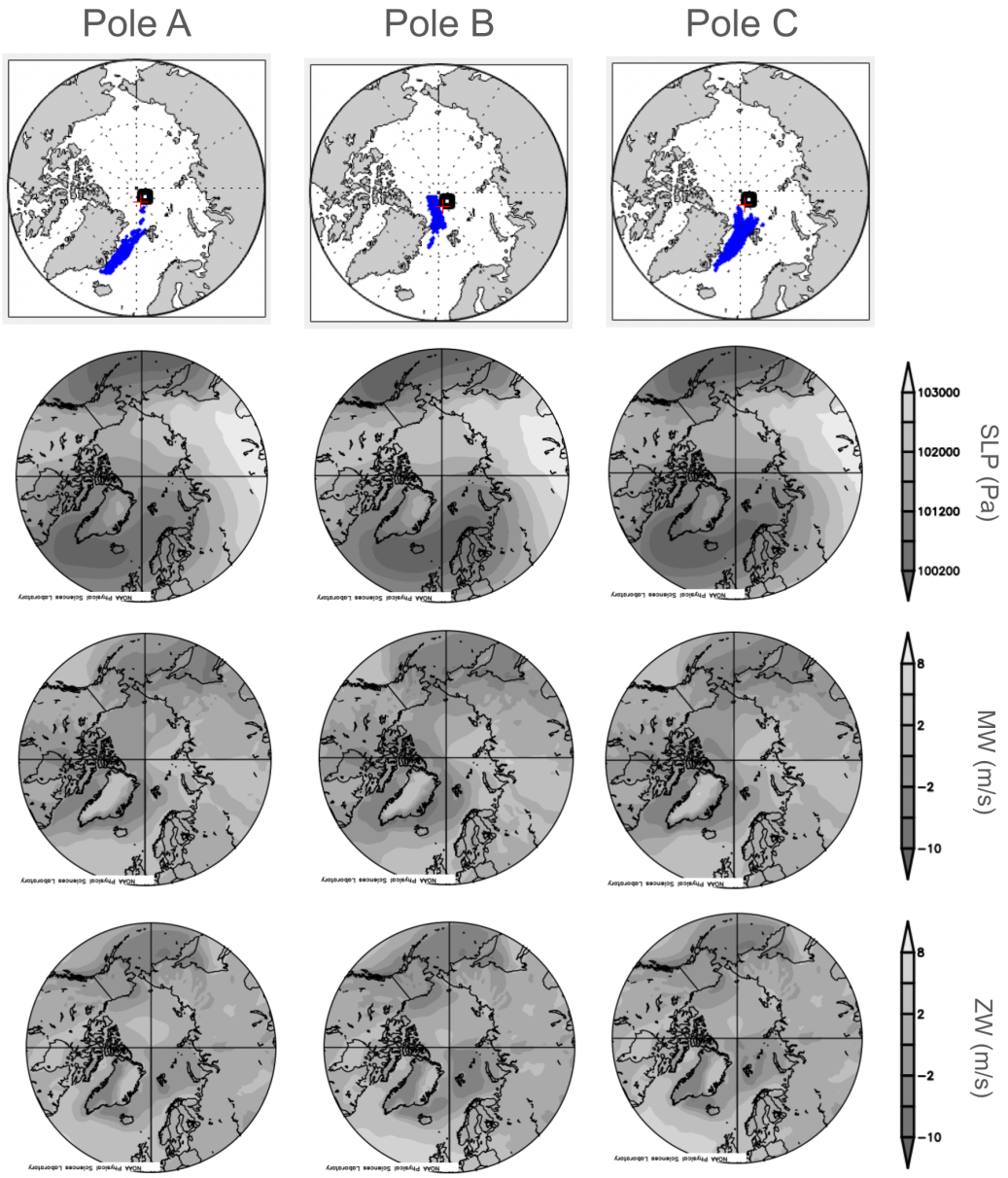

Figure 3.4: Meteorological data from the Pole location. The black box shows where ice originates every year. Ice is tracked from September of one year until September of the next, or until the ice melts. The blue pixels show the end of the ice trajectory. Each map in the top row is a composite of all the ice end locations for each year in that group. Meteorological data is the composite average for all the years belonging to each group. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

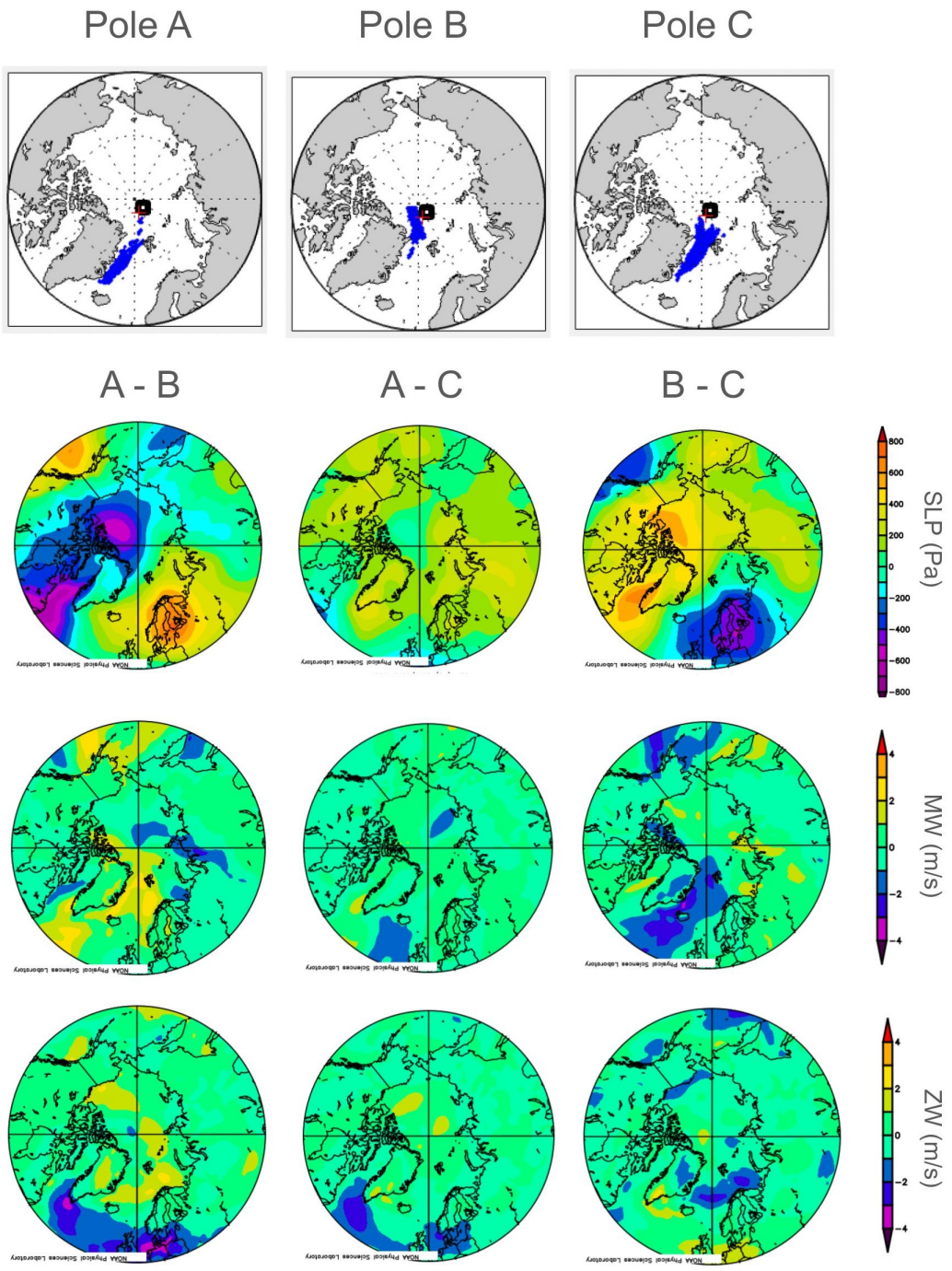

Figure 3.5: Meteorological data difference map from the Pole location. The difference in meteorological data, for example "A - B", is the composite average for all the years belonging to A minus the composite average for all years belonging to B. The meteorological data does not explain the differences in ice drift between the Pole groups. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

Alaska and the Bering Strait, which creates counterclockwise circulation which supplements the BG.

Ice drift in Beaufort B is characteristic of a weaker BG compared to Beaufort A. This can be explained by higher pressure over Alaska and the Bering Strait, creating clockwise circulation that opposes the BG. Ice trajectories in Beaufort B are directed poleward more strongly than trajectories in Beaufort A, suggesting strong TPD. This can be explained by a weaker BG, although meridional wind through the Fram Strait is comparable in Beaufort A and Beaufort B.

Drift in Beaufort C is significantly poleward, indicating strong TPD and a weak BG. Compared to groups A and B, Beaufort C is associated with much lower pressure over the BG. This would weaken the BG, causing the TPD to migrate farther into the central Arctic. Beaufort C also is associated with lower pressure in the Barents Sea, creating counterclockwise circulation which supplements TPD.

Ice in the Beaufort location originates in the marginal ice zone, and like the Pole/Laptev locations, has experienced significant changes in ice properties in recent decades (section 1.2.3). Figure 3.1 plots the distribution of the years of drift belonging to each of the Beaufort groups. Similar to the Pole/Laptev locations, there is a clear age difference between the three groups: Beaufort A contains the oldest years (median 1987), followed by Beaufort C (median 1995.5) and Beaufort A (median 2013.5). The first quartile of years in Beaufort B is younger than the 3rd quartile of years in Beaufort A, showing a significant age difference between these two clusters. Similar to the Pole/Laptev locations, ice travel distance is inversely correlated with age, with the oldest group traveling the least and the youngest group traveling the most.

Age also corresponds to how spread out the drift trajectories are, with ice in Beaufort A (older) staying compact while ice in Beaufort B (younger) is significantly more spread out. This is because ice in the Beaufort location is becoming thinner, younger, and more easily

deformed, leading to more spread out drift trajectories (fig. 3.6).

3.1.4 Ellesmere Island

Clusters found for sea ice parcels originating in the Ellesmere location have short drift trajectories (<1000 km). There are 3 distinct drift patterns: Ellesmere A drifts parallel to the Canadian Arctic Archipelago towards Alaska, Ellesmere C drifts southeast towards the Canadian Arctic Archipelago/Greenland, and Ellesmere B drifts zonally towards the coast of Greenland (fig. 3.8).

Many of the differences in ice trajectories between the three groups can be explained by meteorological data (fig. 3.9). Ellesmere A and Ellesmere B are associated with almost identical SLP and wind velocities, especially where the ice originates, but have the largest difference in ice motion. Ellesmere B is associated with stronger meridional winds through the Fram Strait, which likely drives TPD. This may explain the stronger drift towards the Fram Strait in Ellesmere B. The meteorological data for Ellesmere C shows especially low SLP in the Beaufort Sea. This low pressure creates counterclockwise circulation which supplements the BG, preventing drift towards the Fram Strait.

The ice tracked in the Ellesmere location originates in a reservoir of thick, multiyear ice, mostly unaffected by melt in recent decades. Previous studies have found that this ice has experienced less motion than areas that have been subject to thinning or decrease in ice extent [Kwok, 2018]. Unlike the Laptev/Beaufort/Pole locations, there is not a clear age difference between the three Ellesmere groups: the interquartile data overlaps between all three (fig. 3.1). This supports the hypothesis that recent changes in sea ice properties, as opposed to changes in climate forcing, is responsible for the different ice motion patterns observed in the Laptev/Pole/Beaufort locations.

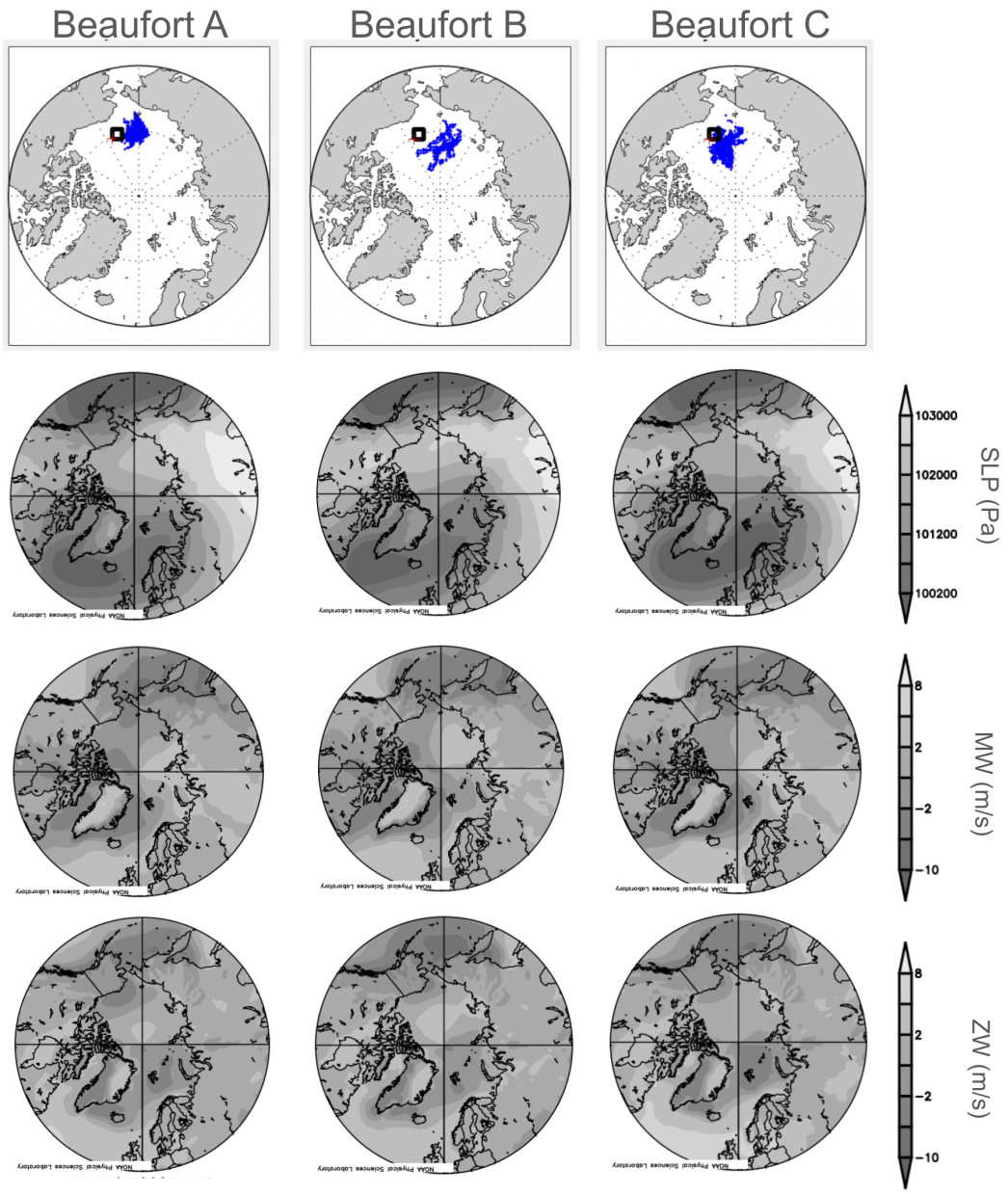

Figure 3.6: Meteorological data from the Beaufort location. The black box shows where ice originates every year. Ice is tracked from September of one year until September of the next, or until the ice melts. The blue pixels show the end of the ice trajectory. Each map in the top row is a composite of all the ice end locations for each year in that group. Meteorological data is the composite average for all the years belonging to each group. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

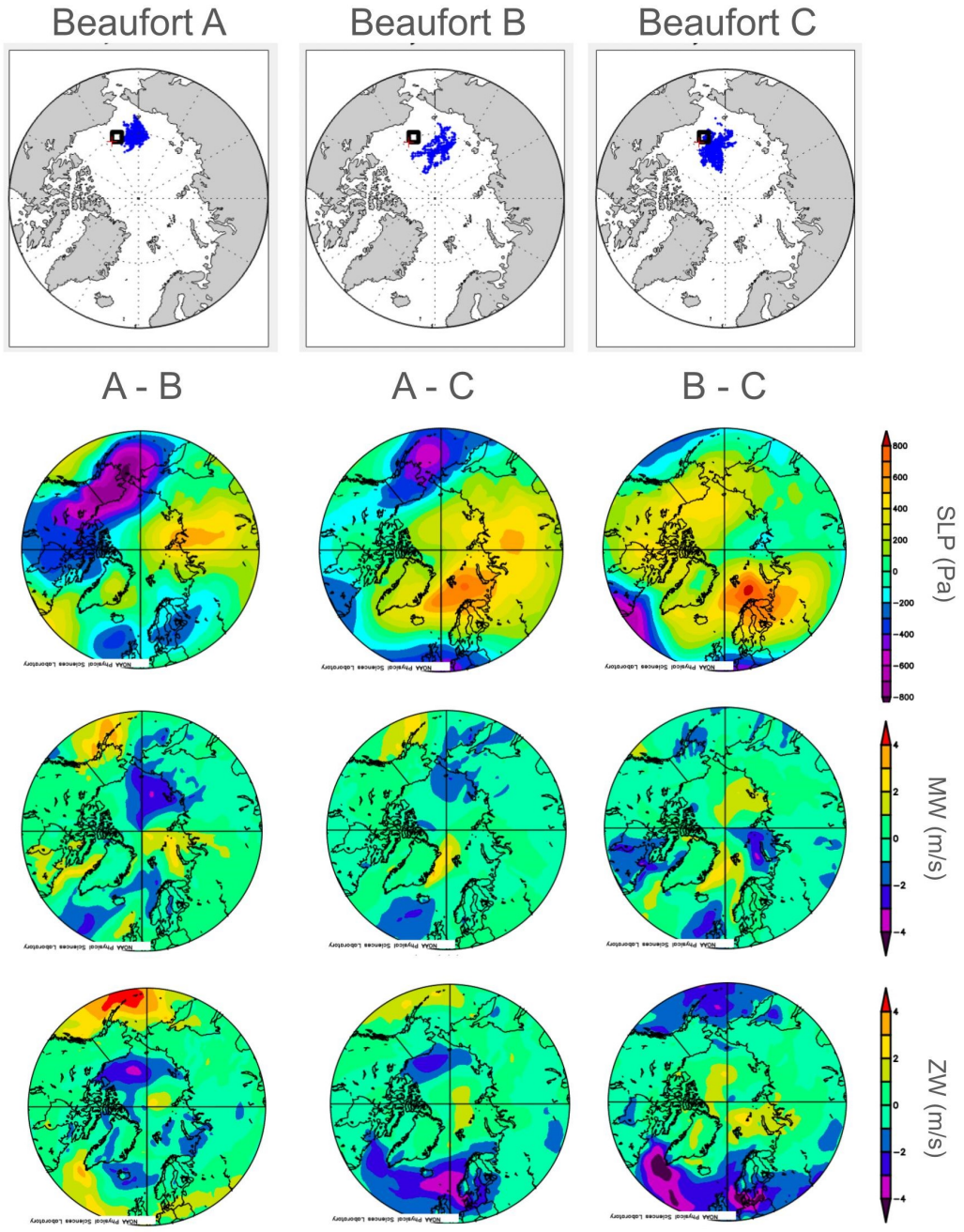

Figure 3.7: Meteorological data difference map from the Beaufort location. The difference in meteorological data, for example "A - B", is the composite average for all the years belonging to A minus the composite average for all years belonging to B. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

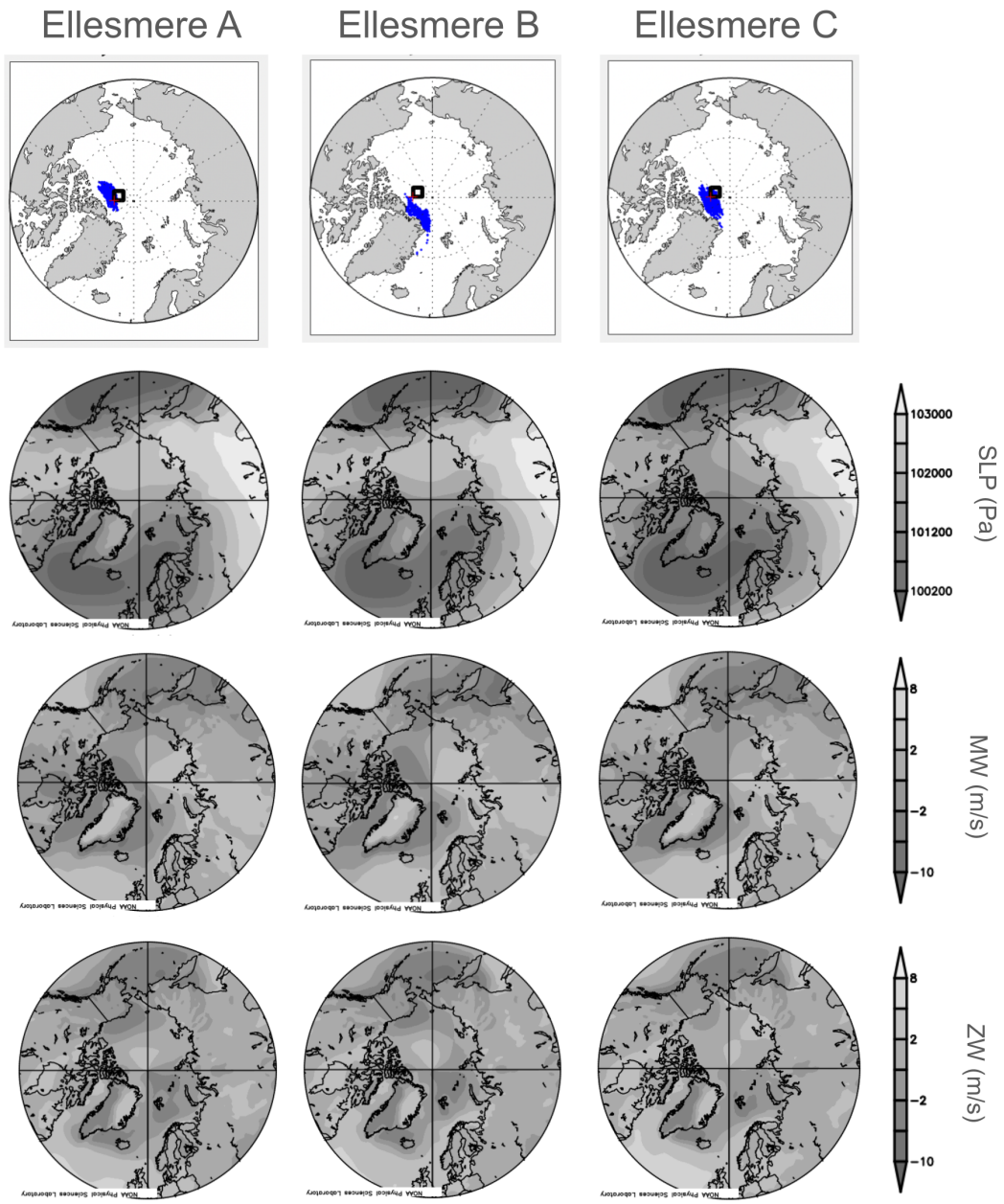

Figure 3.8: Meteorological data from the Ellesmere location. The black box shows where ice originates every year. Ice is tracked from September of one year until September of the next, or until the ice melts. The blue pixels show the end of the ice trajectory. Each map in the top row is a composite of all the ice end locations for each year in that group. Meteorological data is the composite average for all the years belonging to each group. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

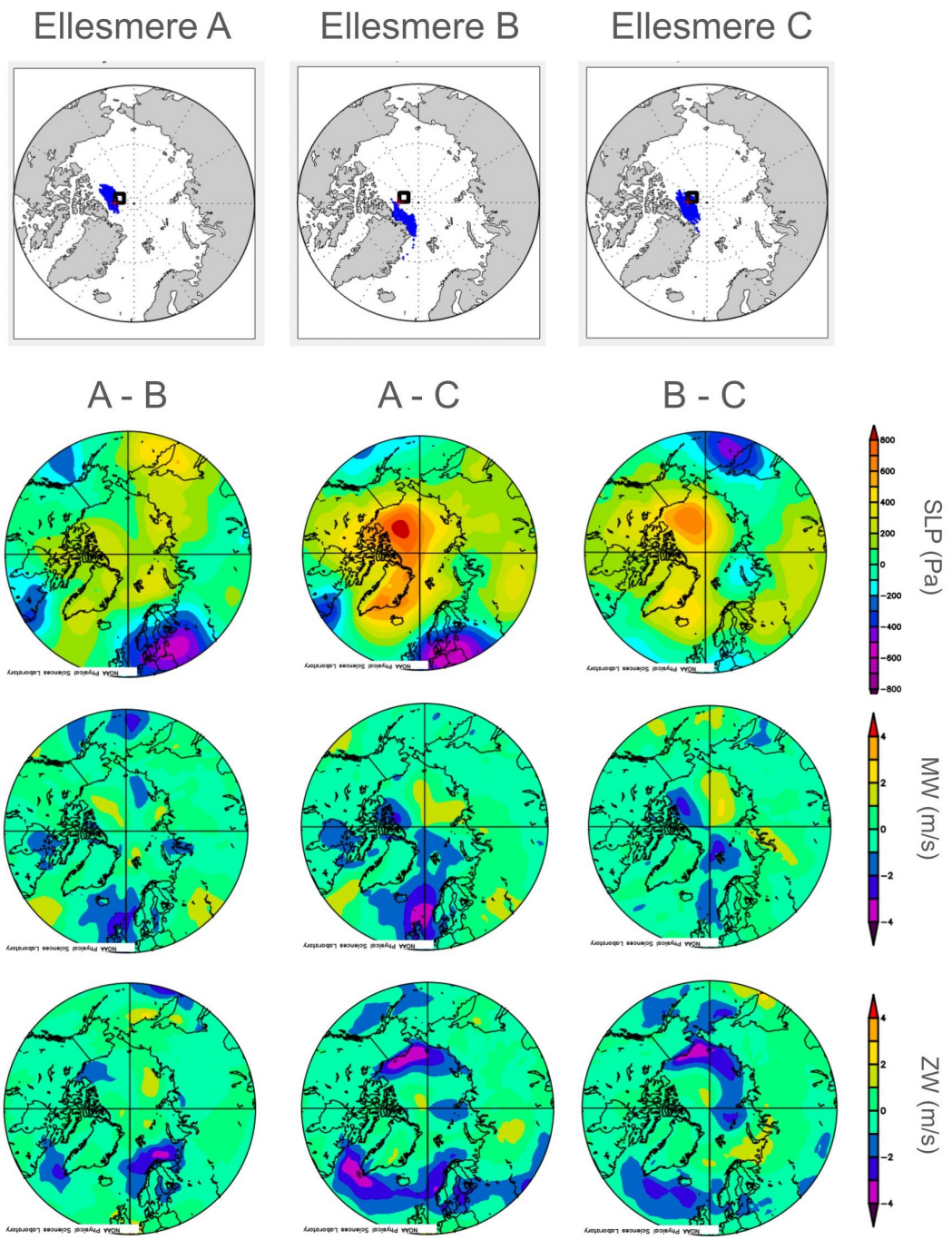

Figure 3.9: Meteorological data difference map from the Ellesmere location. The difference in meteorological data, for example "A - B", is the composite average for all the years belonging to A minus the composite average for all years belonging to B. These plots are provided by the NOAA Physical Sciences Laboratory, Boulder Colorado from their Web site at https://psl.noaa.gov/.

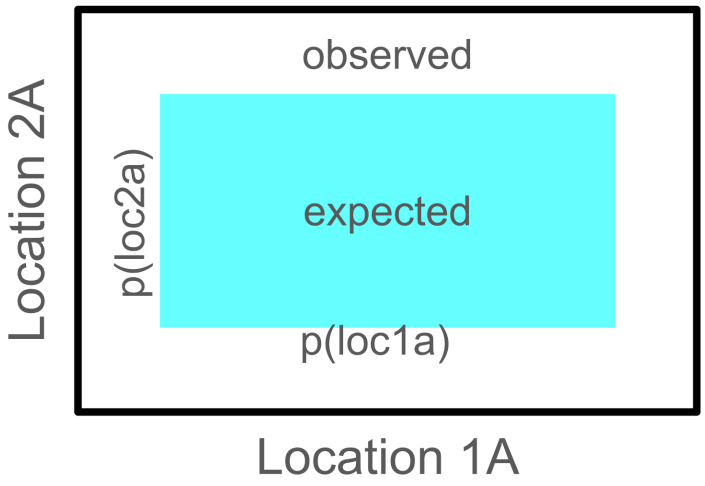

Figure 3.10: Expected vs. observed overlap plotted for years belonging to clusters location 1A and location 2A.

3.2 Associated Clusters

We compared clusters from all four locations to determine whether pan-Arctic drift patterns are visible. This is done by determining which clusters share an unexpectedly large number of years with each other. We represent expected overlap between years belonging to a cluster in different locations with rectangles. To explain this type of plot, a simplified example is shown by figure 3.10. The following notation is used to explain figure 3.10:

- $p(x)$ is the probability of x occurring
- $c(s)$ is the number of items in set s
- loc_{jk} is the set of years belonging to ice originating at location J and cluster k

In this example, we are comparing years belonging to loc_{1a} with years belonging to loc_{2a}. First we calculate $p(loc_{1a})$, the probability of a year randomly belonging to loc_{2a} . This is the number of years in location 1A divided by the sum of all the years in clusters for location 1 (some years are outliers, and don't belong to any of location 1's clusters):

$$E(loc_{1a}) = \frac{c(loc_{1a})}{c(loc_1)} \tag{3.1}$$

On the loc_{1a} axis we plot $p(loc_{1a})$ as the width of the blue rectangle. The height of the blue rectangle is $p(loc_{2a})$. The probability a year randomly ends up in loc_{1a} and loc_{2a}, or $p(loc_{1a} \cap loc_{1b})$ is calculated as:

$$\hat{p}(loc_{1a} \cap loc_{2a}) = p(loc_{1a}) \times p(loc_{2a}) \tag{3.2}$$

Therefore, the area of the rectangle in 3.10 is the expected probability of overlap between years in loc_{1a} and loc_{2a}. The actual overlap is the fraction of years in both loc_{1a} and loc_{2a} given the total number of years in loc_1 and loc_2. This is plotted as the area of the black outline:

$$p(loc_{1a} \cap loc_{2a}) = \frac{|loc_{1a} \cap loc_{2a}|}{|loc_1 + loc_2|} \tag{3.3}$$

In summary, the colored rectangle is the expected overlap between years in clusters if the clusters were created randomly. The black outline is the observed overlap between years in clusters. If the black outline is larger than the colored rectangle, there is more overlap than chance, and vice versa. These results are plotted in figure 3.11. A similar analysis compares each location to AO and NAO indices (figures 3.12 and 3.13). The following sections explore the clusters that were most associated across the four locations.

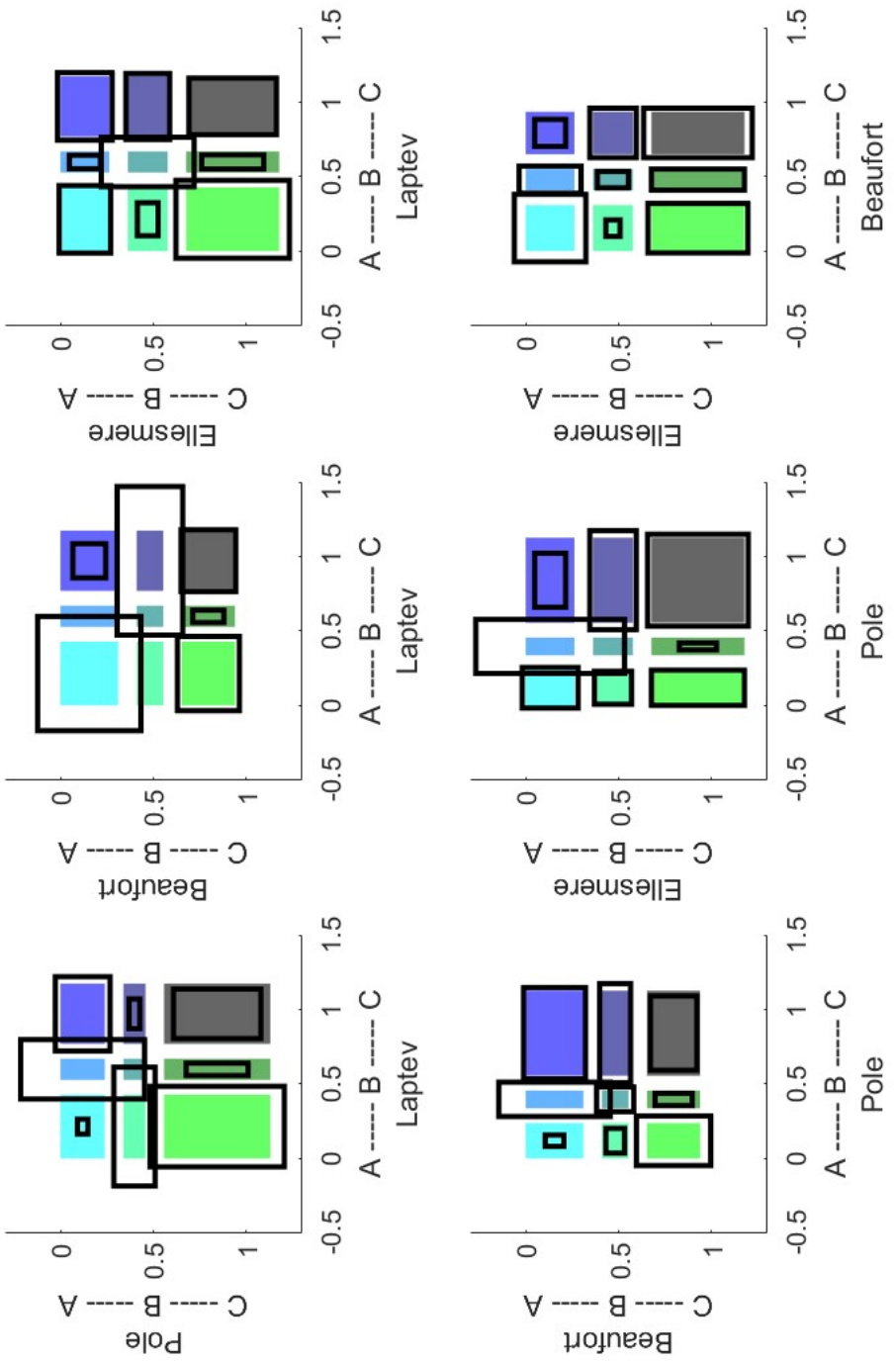

Figure 3.11: Cluster associations between all locations. The probability of belonging to both clusters is represented by the colored rectangles. The observed overlap is represented by the black rectangles. Larger black outlines than the colored rectangles represents more overlap than chance, while rectangles that are smaller than the colored rectangles represent less overlap than chance.

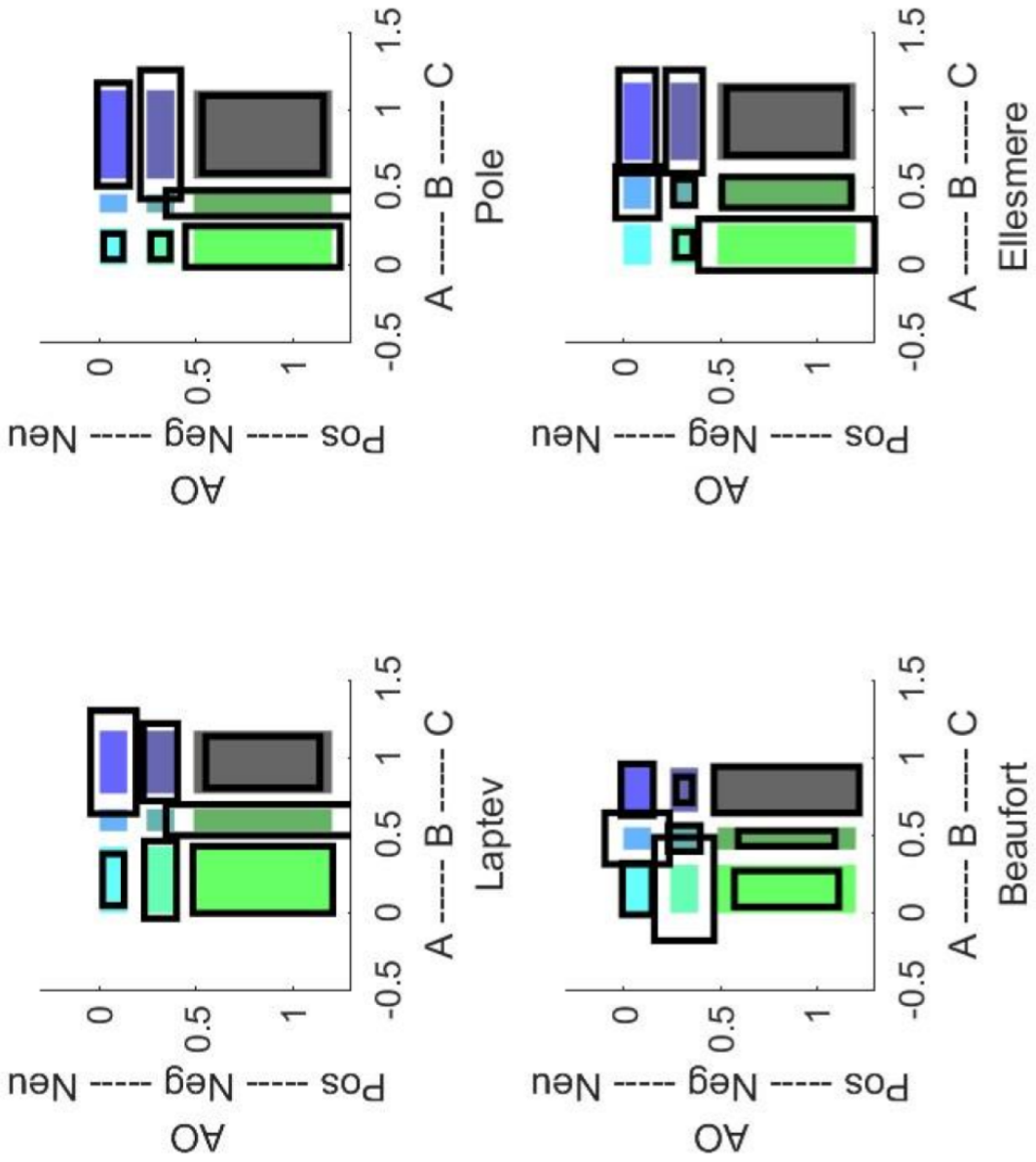

Figure 3.12: AO-Cluster associations between all locations. The probability of belonging to both clusters is represented by the colored rectangles. The observed overlap is represented by the black rectangles. Larger black outlines than the colored rectangles represents more overlap than chance, while rectangles that are smaller than the colored rectangles represent less overlap than chance.

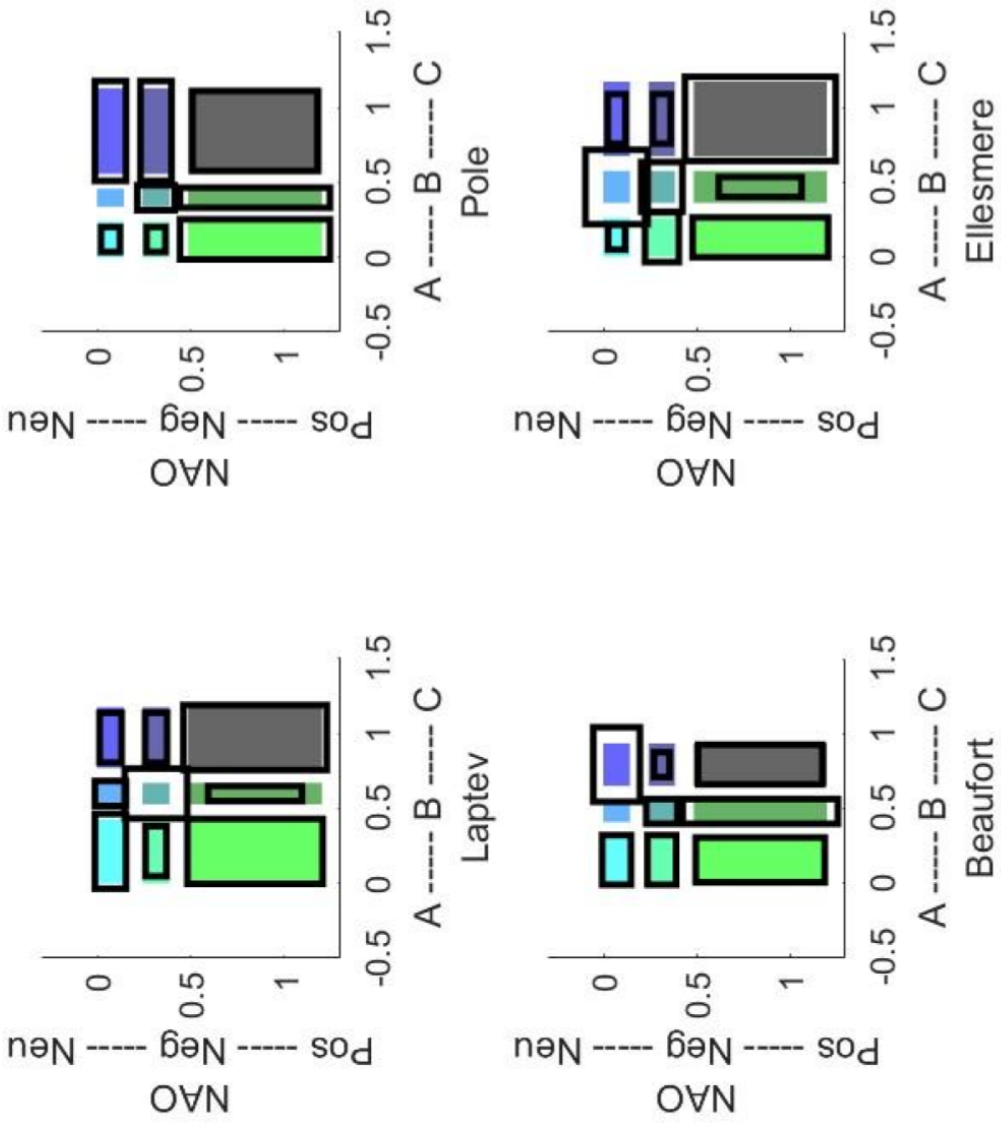

Figure 3.13: NAO-Cluster Associations. The probability of belonging to both clusters is represented by the colored rectangles. The observed overlap is represented by the black rectangles. Larger black outlines than the colored rectangles represents more overlap than chance, while rectangles that are smaller than the colored rectangles represent less overlap than chance.

3.2.1 Pole A and Laptev B

Both of these clusters are from locations on the TPD stream. Overlap between clusters on the TPD is expected, since drift patterns occurring higher up in the drift stream should be similar to those occurring farther down. These clusters contain the most recent years (with Pole A's average age of 2008 and Laptev B's average age of 2012) and trajectories that move the farthest. Pole A's location is not particularly associated to AO or NAO, but Laptev B's location associated with AO positive and NAO negative. This is expected, since AO positive and NAO negative are both expected to cause a strong TPD. The association between Pole A and Laptev B despite the lack of association of Pole A with a climate index suggests that the years are correlated because of recent changes to TPD, specifically the trend towards higher drift speeds, rather than large scale climate forcing.

3.2.2 Old Arctic Group

The Old Arctic Group consists of Pole B, Laptev A, and Beaufort A. Pole B and Laptev A are along the TPD stream, but Beaufort A is most affected by the BG. These clusters represent the oldest groups, with average years of 1992 for pole B, 1996 for Laptev A, and 1991 for Beaufort A. They are all characterized by the least amount of movement for their respective locations. Pole B is associated with AO positive, which is unexpected since previous studies have suggested TPD is stronger during AO positive years. Beaufort A shares overlap with AO negative years, which is also unexpected since previous studies have suggested AO negative years are associated with a stronger BG. There are no other assoiations with AO or NAO between the three clusters. As in the Pole A and Laptev B group (section 3.2.1), the association of the clusters despite shared climate index associations suggests a more locked up and immobile ice pack in the past determines sea ice drift trajectories more significantly than large scale climate forcing.

3.2.3 Beaufort B and Laptev C

The average year in these clusters are more recent, with 2011 for Beaufort B and 2006 for Laptev C. Beaufort B is associated with the most movement of any of the Beaufort clusters, indicating a strong BG. Ice originating in Laptev C moves more than ice originating in Laptev A but less than Laptev B, indicating a medium-strong TPD. Beaufort B and Laptev C both correspond most to AO neutral. The overlap of these two clusters, given the more recent average age, could be representative of years with a stronger BG and a weaker TPD, but thinner ice, causing accelerated ice motion in both the BG and the TPD.

3.2.4 Ellesmere A and Pole B

These locations are geographically close, so overlapping clusters are more expected. Clusters in the Ellesmere location all have about the same average year, and show relatively little movement. This differs from the Pole location, where Pole B cluster is the oldest, with an average year of 1992. Pole B also shows the smallest ice trajectories, which is expected since Pole B contains the oldest years. Movement in Ellesmere A is characterized by running parallel to the Canadian coast, in contrast to the other clusters for that location which tend to move in the opposite direction and bump into the Greenland coast. Similarly, movement in Pole B takes a slightly more zonal trajectory towards Greenland compared to Pole A and Pole C. Both clusters also overlap with AO positive years, when the BG is expected to be weak and TPD strong. This does not match with the observed drift paths, since a strong TPD should increase the change the Ellesmere is brought towards the Fram Strait, and increase the length of the trajectories in Pole B. The resistance to movement by Pole B could be explained by slower ice movements out the Fram strait in the past. However, this does not explain the Ellesmere location, which is moving towards the BG despite the overlap with AO positive years. Ellesmere A's unexpected movement also cannot be explained by

new drift regimes, since the clustered years are not particularly recent, and even today this region still contains some of the oldest ice in the Arctic.

3.2.5 Ellesmere B and Laptev B

The Laptev B group is the most mobile, youngest cluster for the Laptev location, showing long drift trajectories along the TPD. Ellesmere B is characterized by the most movement of the Ellesmere locations, with ice pushed up against the northern Greenland coast. While Ellesmere B does not overlap significantly with the AO, Ellesmere B does have more overlap than expected with NAO neutral years, and to a lesser extent NAO negative years. This is somewhat in line with previous studies which have found NAO negative years produce a more distinct BG and TPD, which would promote longer drift trajectories. Laptev B also corresponds with NAO negative, which is in line with the observed longer drift trajectories. Laptev B is additionally corresponds to AO positive years, which is also associated with a stronger TPD.

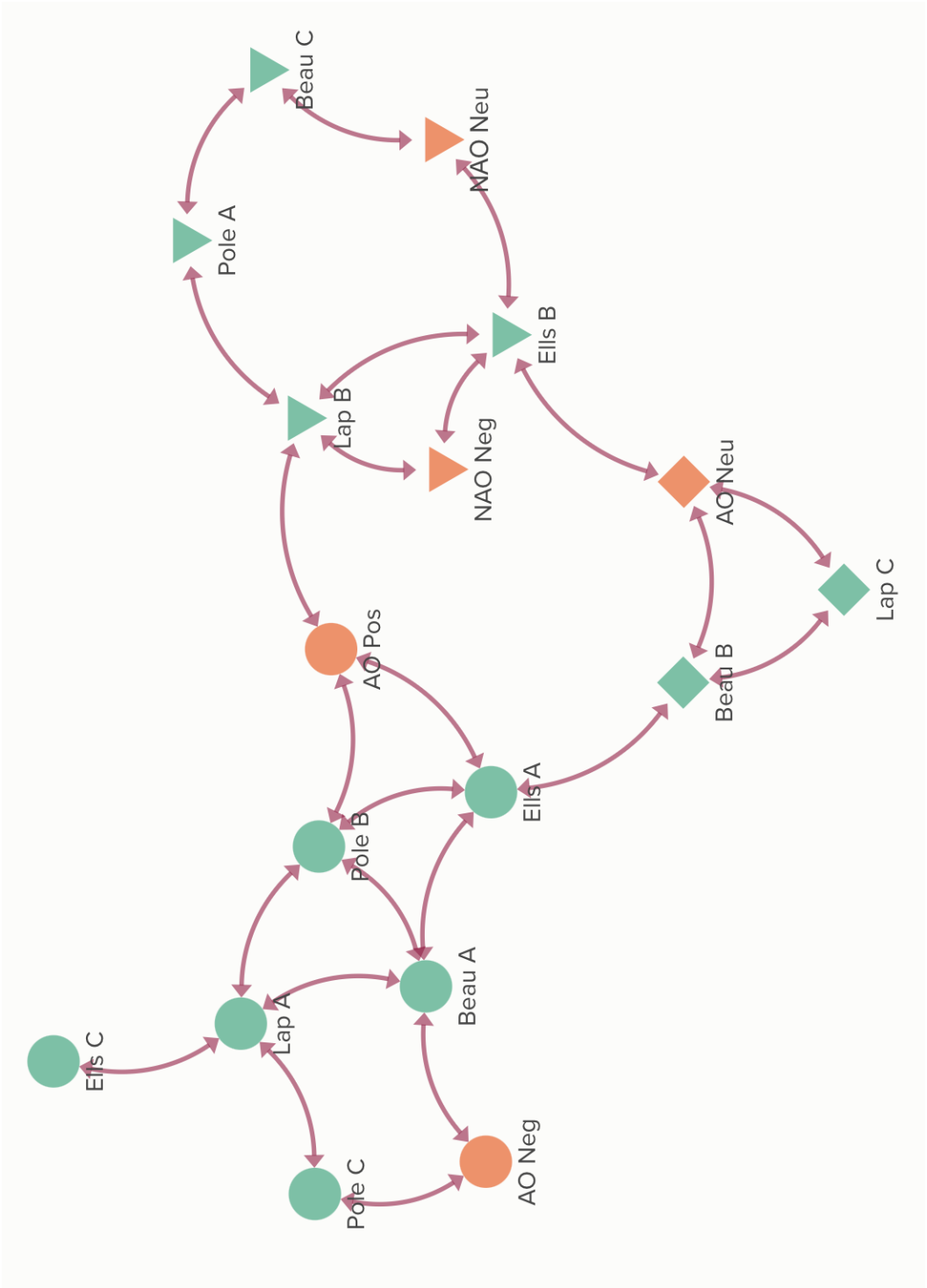

Figure 3.14: Map of connections. An arrow represents a relationship where the expected overlap between the clusters represented by the nodes is greater than random.

Chapter 4

Discussion

Our pan-Arctic analysis of ice drift trajectories from 1980 to 2022 suggest that researchers need to reconsider recent drivers of ice motion as sea ice becomes thinner and more mobile. Ice dynamics research in the 20th century and early 2000s suggests that recurring ice drift patterns are related to large scale climate indices. However, our analysis suggests that changes in ice properties, approximated by the year of drift, are strongly correlated with recurring drift patterns, while these drift patterns are weakly connected (or completely unrelated) to climate indices. Our analysis differed from previous ice drift studies in that we identified ice patterns over small regions, and found that these patterns do repeat simultaneously across the entire Arctic. While ice drift patterns may repeat when the Arctic is considered as a whole, this observation does not hold up when considering drift on smaller scales, and may hint at the complexity of factors involved in determining ice drift, including differential melt trends and local climate forcing. Finally, we have presented novel techniques of clustering geospatial trajectories, which can be used by future studies in the growing field of geospatial machine learning. We also discuss limitations of our study and suggest future research which would help untangle the connection between climate forcing, ice properties, and local sea ice drift patterns.

4.1 Recent Changes in Ice Properties Effect on Drift

Our analysis indicates that recent changes in ice properties have a large influence on local ice drift patterns. Previous studies indicate that thinner and less abundant ice causes increased drift speeds where sea ice is disappearing the fastest. Our results indicate faster drift in recent decades across the Beaufort, Laptev, and Pole locations. This trend is only absent in the Ellesmere location, where thick multiyear ice remains mostly unaffected by the warming Arctic. The trend is strongest in the Beaufort and Laptev locations, where sea ice properties have faced the most extreme changes. Not only is sea ice drifting faster in the Beaufort, Laptev, and Pole locations, but changing ice properties has the largest influence on drift trajectory.

Previous work has explored the relationship between climate indices, such as the NAO or AO, and pan-Arctic ice motion. Our study is the first to consider ice drift patterns on sub-regional scales, and we did not find any associations between the NAO or AO that were shared across clusters in all locations. While climate modes may be connected to large-scale drift patterns, this connection is not replicated when small-scale drift is considered, which may be a result of shifting centers of action even for identical climate indices. This is supported by previous work which found that the NAO fails to capture local variations in climate, and instead the strength and location of the centers of action are significantly more important [Osman et al., 2021]. While we did not find associations with climate indices, most of the drift patterns we identified are explained by sea level pressure and wind composites.

Our results do not indicate that large scale climate forcing causes correlations between local sea ice drift patterns. The Old Arctic Group is the only pan-Arctic correlation in drift years, comprised of the oldest groups from the Laptev, Beaufort, and Pole locations: this indicates a shift in drift regimes in the marginal seas and central Arctic caused by thinner and more mobile sea ice.

4.2 Contributions to Geospatial Clustering Methods

Geospatial machine learning is an emerging field with few published studies. While satellite data is numerous, unlike other research areas with lots of data, geospatial data is high dimensional and evolving over time, creating special challenges for machine learning algorithms [Rolf et al., 2024]. In this project, I developed a novel pipeline for clustering geospatial drift trajectories.

Geospatial trajectories are difficult to cluster because they are represented by high dimensional matrices. Previous studies have successfully used traditional clustering methods to sort trajectories by using alternative distance functions which compute the distance of only a subset of the actual data points, such as the Hausdorf distance [Chen et al., 2011]. Similarly, we decided to only cluster the pixels where the ice ends each year. This was at the risk of losing information about the complete ice drift trajectories. However, a comparison with the full drift trajectories clustered with advanced deep learning methods in section xx shows that the full drift trajectories are not necessary to distinguish between clusters, since sea ice takes similar trajectories to reach the same end locations across different years.

My use of a density based clustering algorithm (DBSCAN) to identify outliers before using k-means clustering is also novel (as far as this author is aware), and led to a significant improvement in the quality and stability of clusters created by k-means. This method of removing outliers may be useful in future geospatial studies, where outliers in high dimensional data are difficult to identify and have a significant impact on the performance of traditional clustering algorithms.

4.3 Limitations and Future Work

Our analysis is the first pan-Arctic comparison of local drift patterns, made possible by the Polar Pathfinder dataset. The main limitations come from inaccuracies in the Pathfinder dataset, but in section x we also suggest future improvements to our ice tracking and clustering methods. Finally I discuss how the connection between climate and small-scale drift should be further explored by moving beyond climate indices, which do not capture enough geospatial information about their centers of action.

4.3.1 Polar Pathfinder Dataset

The Polar Pathfinder Dataset is the largest source of uncertainty in our analysis. The data is less reliable where ice is melting, and doesn't exist in regions with less than 15% ice concentrations. There are also no calculations over the pole, or in regions without enough open ocean, like near the Canadian Archipelago. The dataset is most accurate for meridional drift (common in the TPD), but less accurate for zonal drift (common in the BG). Errors in drift trajectories calculated with the Pathfinder dataset accumulate over time, and generally ice motion is underestimated.

One of the most overlooked sources of error in the Pathfinder dataset is that the wind-derived estimates of ice motion rely on Thorndike and Colony's 1982 observation that summer ice moves at 1% of wind speed and 20 degrees to the wind. Numerous studies have indicating that ice is moving faster in response to the same wind forcing as Arctic sea ice becomes thinner and less locked up [Gascard et al., 2008, Rampal et al., 2009, Kwok et al., 2013b, Tandon et al., 2018]. This may explain why [Gui et al., 2020] found the Pathfinder dataset underestimated drift trajectories in both the years studied, 2014 and 2016. Future versions of the dataset by NSIDC could be improved by taking into account ice concentration and thickness when calculating the affects of wind vectors on sea ice motion.

The inaccuracies in the Pathfinder dataset are likely why a study of local drift trajectories has not yet been explored. However, the trends in ice motion identified in this thesis are supported by previous observations of Arctic sea ice, which supports their significance despite the issues with the underlying dataset, and suggests that the Pathfinder dataset should not be overlooked for small-scale studies of ice motion.

Our results indicate small-scale drift patterns differ from large-scale drift, and is motivation to improve the pathfinder dataset, through changes to the wind-component of the calculation, the addition of buoys to increase the accuracy of ice motion estimates, or improvement to other variables in the calculation.

4.3.2 Ice Tracking and Clustering Methods

The ice tracking code attached in appendix A.3 should be improved by adding a boolean to check whether tracking has ended before the end of the year. The code currently tracks the ice parcels for exactly one year, even if the original parcel has melted. This is likely not a significant source of error, since our ice tracking algorithm sets the velocity of the parcel to zero if the Pathfinder dataset indicates open water. However, there is a chance that an ice floe from another region could drift into the pixel where the original parcel melted, and this new ice floe would then be tracked as if it were the original ice parcel.

The clustering in section 2.4.1 could be improved to increase the stability of the clusters. One common clustering technique I did not implement is to repeat k-means clustering 5 or more times, and save the clustering which repeats the most often.

The full trajectory clustering in section 2.4.2 should be further investigated. Previous studies have similarly used autoencoders to cluster trajectories, and their methods would be worth exploring further [Olive et al., 2020, Zeng et al., 2021].

4.3.3 Explaining Drift Patterns

While we observed as strong trend in changing drift patterns with changes in sea ice properties, future work should further explore the connection to large scale climate forces. Beyond climate indices, metrics that capture the strength and location of centers of actions should be considered, since these are more relevant to local sea ice motion. On small scales, changes in ice properties seem to be more important than pan-Arctic climate patterns. Future studies should consider whether climate indices are still relevant to large-scale sea ice dynamics as the properties of sea ice rapidly change.

A more detailed analysis of the climate data could employ causal discovery techniques to determine whether local or global sea level pressure patterns are most associated with sea ice drift. Sea level pressure matrices can be given to algorithms like FCI, which determine which sea level pressure pixels cause the observed drift trajectory accounting for other confounding variables such as sea ice thickness and concentration.

All future studies would benefit from more years of ice drift data. Some of the clusters only contain 5-6 data points, which makes establishing a trend difficult. More data would allow complex and evolving trends to become apparent.

Chapter 5

Conclusions

This study is the first comprehensive pan-Arctic comparison of local ice drift patterns from 1980 to 2022. By leveraging the Polar Pathfinder dataset, which provides continuous estimates of ice motion across nearly the entire Arctic, we achieved three primary goals: (1) developing clustering techniques to sort geospatial trajectories, (2) investigating the influence of climate modes and recent changes in ice properties on medium-scale drift patterns, and (3) identifying pan-Arctic correlations in local drift trajectories.

We analyzed ice originating in four locations: the Beaufort Sea, the North Pole, Ellesmere Island, and the Laptev Sea. These regions have experienced varying magnitudes of ice melt in recent decades. The Beaufort and Laptev locations are in the marginal seas, where ice has decreased significantly in both concentration and thickness. In contrast, the Ellesmere location retains thick, mulityear ice, largely unaffected by recent melting trends. The Pole location is in the central Arctic, where ice has decreased in thickness with smaller declines in concentration [Zhang et al., 2012, Kwok, 2018]. Using the Pathfinder dataset, we computed year-long drift trajectories from daily velocity vectors for ice originating in the study locations. For each location, we tracked ice originating 250 x 250 km^2.

We developed a clustering pipeline to sort these yearly drift trajectories by similarity.

This task was challenging because each drift trajectory is a large amount of data and we only have 40 years of drift for each location. Some years of drift are also unique, and do not belong in any cluster. To reduce the amount of data representing drift, we used only the final position of the ice to represent a year of drift. We then applied Principle Component Analysis (PCA) for futher dimensionality redution and used a density based clustering algorithm (DBSCAN) to identify and remove outliers. This step was crucial to avoid random cluster formations when applying k-means, which sorted the drift patterns into 3 clusters (the optimal k for each location).

Geospatial machine learning is a new field, and our clustering approach, particularly the use of trajectory endpoints combined with DBSCAN to remove outliers, is a novel clustering approach which may be useful in future studies. We also experimented with clustering full-path trajectories for the Beaufort location using autoencoders for dimensional reduction, but this approach was computationally intensive without producing significantly better results.

We found that local drift is significantly influenced by changes in sea ice properties. In the Beaufort, Laptev, and Pole regions, where ice has weakened considerably, clusters with minimal ice movements contained the oldest years. This suggests a shift towards a more mobile Arctic with faster ice drift. Conversely, in the Ellesmere location, there was no age difference between ice drift clusters, implying that factors other than ice properties determine medium-scale drift here.

Local drift showed little association with climate indices, likely because these indices do not capture the geospatial locations of anomalous sea level pressure relevant to smaller drift regions. Additionally, we found no significant pan-Arctic correlation in local drift, with the exception of the "Old Arctic Group," where ice in the Laptev, Beaufort, and Pole regions exhibited shorter drift trajectories in the past. Further research is needed to establish a causal link between local drift and basin-wide sea level pressure patterns.

Our work introduces new techniques for trajectory clustering, and is also a first look at

medium-scale sea ice drift. More work needs to be done to explain the drift patterns we have identified, but this project lays groundwork for future studies involving local drift dynamics and geospatial machine learning.

Appendix A

An appendix

A.1 Code Graph

A.2 NOAA 20th Century Reanalysis Monthly Composites URL builder

```
% Produces stereo arctic projection for a given variable

% var = variable to plot. slp = Sea Level Pressure, zw = Zonal Wind, mw = meridional wind
% grp = creates a seasonal composite of the variable over these years.
% minus_group = subtracts the seasonal composite of the variable in these years from grp
% rlow = lower bound for scale bar
% rhigh = upper bound for scale bar
% cint = interval for scale bar
% color = color scheme for plot. bw = black and white, while color = rainbow

function copy_noaa_plot_url(var, grp, minus_grp, rlow, rhigh, cint, color)

    % first part of URL is always the same
    prefix = "https://psl.noaa.gov/cgi-bin/data/composites/comp.20thc.v2.pl?";
```

```
15
16      % match color scheme
17      switch color
18          case "bw"
19              color_urlname = "Black+and+white"; % 'Color'
20          otherwise
21              color_urlname = "Color";
22      end
23
24      % match variable name
25      switch var
26          case "slp"
27              var_urlname = "Sea+Level+Pressure";
28          case "zw"
29              var_urlname = "Zonal+Wind";
30          case "mw"
31              var_urlname = "Meridional+Wind";
32          otherwise
33              disp("ERROR: invalid variable name");
34              quit(code)
35      end
36
37      % url before the years are added in
38      prefix = prefix + "var="+ var_urlname + "&level=1000mb&
           version=1&mon1=0&mon2=0";
39
40      % add years in grp to composite
41      years = "";
42      for i = 1:length(grp)
43
44          years = years + "&iy=";
45
46          if i <= length(grp)
47              years = years + grp(i);
48          end
49
50      end
51
52      % years with a negative sign in front will be part of the
           composite
53      % subtracted from grp years
54      for i = 1:length(minus_grp)
```

```
55          years = years + "&iy=-" + minus_grp(i);
56
57      end
58
59      % everything after the years
60      suffix ="&ipos%5B1%5D=&ipos%5B2%5D=&ineg%5B1%5D=&ineg%5B2%5
           D=&timefile0=&tstype=3&timefile1=&value=&typeval=1&
           compval=1&lag=0&labelcolor=" ...
61            + color_urlname + "&labelshaded=Shaded&type=1&ensemble
                 =1&scale=&contourlabel=0&switch=0&cint="+cint+"&
                 lowr="+rlow+"&highr="+rhigh+ ...
62            "&proj=Custom&xlat1=50&xlat2=90&xlon1=0&xlon2=360&
                 custproj=Northern+Hemisphere+Polar+Stereographic&
                 Submit=Create+Plot";
63      url = prefix + years + suffix;
64      clipboard('copy', url)
65  end
```

A.3 Ice Tracking Code

Tracks the location of ice originating in one pixel over the course of one year.

```
1
2  % Given starting pixels and days, tracks the location the ice
       moves over 1
3  % year
4  %Returns tracks in 2 arrays: one for the column pixel locations
       , and one
5  %for the row pixel locations
6
7  function [col_pixels_year, row_pixels_year] =
       get_pixel_path_one_year(col0_pixel, row0_pixel, v, u,
       start_year_day, end_year_day, x_ref, y_ref)
8
9       numDays = length(start_year_day:end_year_day);
10
11      % tracks x/y coordinates of ice over the year, according to
12      % coordinate system used in x_ref/y_ref
13      col_coords = nan(1, numDays);%nan(end_year_day,1);
14      row_coords = nan(1, numDays);%nan(end_year_day,1);
15      col_coords(1) = x_ref(col0_pixel);
16      row_coords(1) = y_ref(row0_pixel);
```

```
17
18      % tracks which x/y pixels ice is in over the year
19      % 1xDaysInYear matrix, where 1 index corresponds to
           location for each
20      % day in the year
21      col_pixels_year = nan(1, numDays);
22      row_pixels_year = nan(1, numDays);
23      col_pixels_year(1) = col0_pixel;
24      row_pixels_year(1) = row0_pixel;
25
26      s_in_day = 24 * 3600; % num seconds in day
27      start_day_offset = start_year_day - 1;
28
29      % Calculates the x and y pixel locations of the ice over an
            entire
30      % year, given the starting day/location.
31      for on_this_day = 1:numDays - 1 % any number from 1 -
           15630, and can be used to index into the dataset (v/u/
           xref... etc)
32
33          % Determine velocities of the ice in m/s
34          %   v = x-velocity: index into with pixels
35          %   u = y-velocity: index into with pixels
36          % NOTE: the coordinate system is such that (for
               whatever reason)
37          % the direction of positive v and u velocity is upside
               down. See
38          % pg. 8 of https://nsidc.org/sites/default/files/nsidc
               -0116-v003-userguide_0.pdf
39          % and compare to an imagesc plot of the unprocessed
               data.
40          %
41          %   * both v/u given in cm/s, so must divide by 100 to
               get m/s!
42          v_ms = -v(row_pixels_year(on_this_day), col_pixels_year
               (on_this_day), on_this_day + start_day_offset)/100;
43          u_ms = -u(row_pixels_year(on_this_day), col_pixels_year
               (on_this_day), on_this_day + start_day_offset)/100;
44
45          % set velocities to 0 if they are nan (ice not moving)
46          v_ms(isnan(v_ms)) = 0; % y velocity
47          u_ms(isnan(u_ms)) = 0; % x velocity
```

```
48
49          % update x and y locations for next day based on
               velocities
50          col_coords(on_this_day + 1) = col_coords(on_this_day) +
               u_ms * s_in_day; % change in x for 1 day
51          row_coords(on_this_day + 1) = row_coords(on_this_day) +
               v_ms * s_in_day; % change in y for 1 day
52
53          % set pixels of the next day to pixels corresponding to
               the
54          % computed x1/y1 coordinate locations
55          [~, col_pixels_year(on_this_day + 1)] = min((col_coords
               (on_this_day + 1) - x_ref).^2);
56          [~, row_pixels_year(on_this_day + 1)] = min((row_coords
               (on_this_day + 1) - y_ref).^2);
57
58      end
59  end
```

Bibliography

Monitoring north atlantic oscillation. URL https://www.ncei.noaa.gov/access/monitoring/nao/.

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Preeti Arora, Deepali, and Shipra Varshney. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512, 12 2016. doi: 10.1016/j.procs.2016.02.095.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer, 1999.

Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105 (9), 2014.

NOAA Climate Prediction Center. Arctic oscillation. https://psl.noaa.gov/data/correlation/ao.data, a.

NOAA Climate Prediction Center. North atlantic oscillation. https://psl.noaa.gov/data/correlation/nao.data, b.

Jinyang Chen, Rangding Wang, Liangxu Liu, and Jiatao Song. Clustering of trajectories based on hausdorff distance. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 1940–1944. IEEE, 2011.

Divam. An overview of deep learning based clustering techniques, Mar 2019. URL `https://divamgupta.com/unsupervised-learning/2019/03/08/an-overview-of-deep-learning-based-clustering-techniques.html#:~:text=The%20deep%20learning%20based%20methods,for%20improvement%20in%20several%20datasets.`

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Andrew Evans. Forward, May 2012. URL `https://www.nationalgeographic.com/travel/article/forward`.

Jean-Claude Gascard, Jean Festy, Hervé le Goff, Matthieu Weber, Burghard Bruemmer, Michael Offermann, Martin Doble, Peter Wadhams, René Forsberg, Susan Hanson, et al. Exploring arctic transpolar drift during dramatic sea ice retreat. *Eos, Transactions American Geophysical Union*, 89(3):21–22, 2008.

Dawei Gui, Ruibo 18, Xiaoping Pang, Jennifer K Hutchings, Guangyu Zuo, and Mengxi Zhai. Validation of remote-sensing products of sea-ice motion: A case study in the western arctic ocean. *Journal of Glaciology*, 66(259):807–821, 2020.

Marika M Holland. The north atlantic oscillation–arctic oscillation in the ccsm2 and its influence on arctic climate variability. *Journal of Climate*, 16(16):2767–2781, 2003.

Satwant Kaur, Jens K. Ehn, and David G. Barber. Pan-arctic winter drift speeds and

changing patterns of sea ice motion: 1979–2015. *Polar Record*, 54(5–6):303–311, 2018. doi: 10.1017/S0032247418000566.

Kristina Kiest. International arctic buoy programme, Jan 2024. URL `https://arctic.noaa.gov/research/international-arctic-buoy-programme/`.

Thomas Krumpen, H. Jakob Belter, Antje Boetius, Ellen Damm, Christian Haas, Stefan Hendricks, Marcel Nicolaus, Eva-Maria Nöthig, Stephan Paul, Ilka Peeken, and et al. Arctic warming interrupts the transpolar drift and affects long-range transport of sea ice and ice-rafted matter. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-41456-y.

R. Kwok, G. Spreen, and S. Pang. Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. *Journal of Geophysical Research: Oceans*, 118(5):2408–2425, 2013a. doi: https://doi.org/10.1002/jgrc.20191. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrc.20191`.

Ron Kwok. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, 13(10):105005, 2018.

Ron Kwok, G Spreen, and S Pang. Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. *Journal of Geophysical Research: Oceans*, 118(5):2408–2425, 2013b.

Ronald Kwok. Recent changes in arctic ocean sea ice motion associated with the north atlantic oscillation. *Geophysical Research Letters*, 27(6):775–778, 2000. doi: https://doi.org/10.1029/1999GL002382. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999GL002382`.

Sascha Lange and Martin Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *The 2010 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.

Ruibo Lei, Mario Hoppmann, Bin Cheng, Guangyu Zuo, Dawei Gui, Qiongqiong Cai, H Jakob Belter, and Wangxiao Yang. Seasonal changes in sea ice kinematics and deformation in the pacific sector of the arctic ocean in 2018/19. *The Cryosphere*, 15(3): 1321–1341, 2021.

Matti Leppäranta. *The drift of sea ice*. Springer Science & Business Media, 2011.

Peigen Lin, Robert S Pickart, Harry Heorton, Michel Tsamados, Motoyo Itoh, and Takashi Kikuchi. Recent state transition of the arctic ocean's beaufort gyre. *Nature Geoscience*, 16(6):485–491, 2023.

Rebecca Lindsey. Climate variability: Arctic oscillation. URL `https://www.climate.gov/news-features/understanding-climate/climate-variability-arctic-oscillation`.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Torge Martin and Rüdiger Gerdes. Sea ice drift variability in arctic ocean model intercomparison project models and observations. *Journal of Geophysical Research: Oceans*, 112 (C4), 2007.

Walter N. Meier. Satellite passive microwave observations of sea ice. In J. Kirk Cochran, Henry J. Bokuniewicz, and Patricia L. Yager, editors, *Encyclopedia of Ocean Sciences (Third Edition)*, pages 402–414. Academic Press, Oxford, third edition edition, 2019. ISBN 978-0-12-813082-7. doi: https://doi.org/10.1016/B978-0-12-409548-9.11461-7. URL `https://www.sciencedirect.com/science/article/pii/B9780124095489114617`.

Walter N. Meier. Sea ice in the satellite era. *Past Global Changes*, 30(2):70–71, 2022. URL https://doi.org/10.22498/pages.30.2.70.

Fridtjof Nansen. *Farthest North: Being the Record of a Voyage of Exploration of the Ship Fram, 1893-96, and of a Fifteen Months' Sleigh Journey*, volume 2. Cambridge University Press, 1897.

NSIDC. National snow and ice data center, 2024. URL https://nsidc.org/learn/parts-cryosphere/sea-ice/science-sea-ice#anchor-6.

Xavier Olive, Luis Basora, Benoit Viry, and Richard Alligier. Deep trajectory clustering with autoencoders. In *ICRAT 2020, 9th International Conference for Research in Air Transportation*, 2020.

Mahmoud Osman, Benjamin Zaitchik, Hamada Badr, and Sultan Hameed. North atlantic centers of action and seasonal to subseasonal temperature variability in europe and eastern north america. *International Journal of Climatology*, 41:E1775–E1790, 2021.

Leonid Polyak, Richard B Alley, John T Andrews, Julie Brigham-Grette, Thomas M Cronin, Dennis A Darby, Arthur S Dyke, Joan J Fitzpatrick, Svend Funder, Marika Holland, et al. History of sea ice in the arctic. *Quaternary Science Reviews*, 29(15-16):1757–1778, 2010.

A. Preußer, G. Heinemann, S. Willmes, and S. Paul. Circumpolar polynya regions and ice production in the arctic: results from modis thermal infrared imagery from 2002/2003 to 2014/2015 with a regional focus on the laptev sea. *The Cryosphere*, 10(6):3021–3042, 2016. doi: 10.5194/tc-10-3021-2016. URL https://tc.copernicus.org/articles/10/3021/2016/.

Pierre Rampal, Jérôme Weiss, and David Marsan. Positive trend in the mean speed and

deformation rate of arctic sea ice, 1979–2007. *Journal of Geophysical Research: Oceans*, 114(C5), 2009.

Ignatius Rigor and Mark Ortmeyer. The international arctic buoy programme–monitoring the arctic ocean for forecasting and research. *Arctic Research of the United States*, 18: 21–21, 2004.

Ignatius G Rigor, John M Wallace, and Roger L Colony. Response of sea ice to the arctic oscillation. *Journal of Climate*, 15(18):2648–2663, 2002.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024.

Mark C Serreze and Walter N Meier. The arctic's sea ice cover: trends, variability, predictability, and comparisons to the antarctic. *Annals of the New York Academy of Sciences*, 1436(1):36–53, 2019.

Pulkit Sharma. The ultimate guide to k-means clustering: Definition, methods and applications, Feb 2024. URL `https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/`.

Gunnar Spreen, Ron Kwok, and Dimitris Menemenlis. Trends in arctic sea ice drift and role of wind forcing: 1992–2009. *Geophysical Research Letters*, 38(19), 2011. doi: https://doi.org/10.1029/2011GL048970. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GL048970`.

Hiroshi Sumata, Laura de Steur, Dmitry V Divine, Mats A Granskog, and Sebastian Gerland. Regime shift in arctic ocean sea ice thickness. *Nature*, 615(7952):443–449, 2023.

Neil F Tandon, Paul J Kushner, David Docquier, Justin J Wettstein, and Camille Li. Reassessing sea ice drift and its relationship to long-term arctic sea ice loss in coupled climate models. *Journal of Geophysical Research: Oceans*, 123(6):4338–4359, 2018.

David WJ Thompson and John M Wallace. The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, 25(9):1297–1300, 1998.

AS Thorndike and R Colony. Sea ice motion in response to geostrophic winds. *Journal of Geophysical Research: Oceans*, 87(C8):5845–5852, 1982.

Mary-Louise Timmermans and John M Toole. The arctic ocean's beaufort gyre. *Annual Review of Marine Science*, 15:223–248, 2023.

M. Tschudi, Fowler C., Maslanik J., J. S. Stewart, and W. N. Meier. Polar pathfinder daily 25 km ease-grid sea ice motion vectors, version 3, 2016. URL `https://nsidc.org/data/NSIDC-0116/versions/3`.

M. Tschudi, W. N Meier, J. S. Stewart, C. Fowler, and J. Maslanik. Measuring sea ice motion, 2019.

Timo Vihma, Priit Tisler, and Petteri Uotila. Atmospheric forcing on the drift of arctic sea ice in 1989–2009. *Geophysical Research Letters*, 39(2), 2012.

Eiji Watanabe, Jia Wang, Akimasa Sumi, and Hiroyasu Hasumi. Arctic dipole anomaly and its contribution to sea ice export from the arctic ocean in the 20th century. *Geophysical research letters*, 33(23), 2006.

W. Weeks. *On Sea Ice*. University of Alaska Press, 2010. ISBN 9781602231016. URL `https://books.google.com/books?id=9S55O6WzuL8C`.

Chris Wilson, Yevgeny Aksenov, Stefanie Rynders, Stephen J Kelly, Thomas Krumpen, and Andrew C Coward. Significant variability of structure and predictability of arctic ocean surface pathways affects basin-wide connectivity. *Communications Earth & Environment*, 2(1):164, 2021.

Bingyi Wu, Jia Wang, and John E Walsh. Dipole anomaly in the winter arctic atmosphere and its association with sea ice motion. *Journal of Climate*, 19(2):210–225, 2006.

Tian Yin, Bai Xuezhi, and Huang Yingqi. Analysis of the variation in intensity and source region of the arctic transpolar drift. *Chinese Journal of Polar Research*, 33(4):529, 2021.

Weili Zeng, Zhengfeng Xu, Zhipeng Cai, Xiao Chu, and Xiaobo Lu. Aircraft trajectory clustering in terminal airspace based on deep autoencoder and gaussian mixture model. *Aerospace*, 8(9):266, 2021.

Fanyi Zhang, Xiaoping Pang, Ruibo Lei, Mengxi Zhai, Xi Zhao, and Qiongqiong Cai. Arctic sea ice motion change and response to atmospheric forcing between 1979 and 2019. *International Journal of Climatology*, 42(3):1854–1876, 2022.

Jinlun Zhang, Ron Lindsay, Axel Schweiger, and Ignatius Rigor. Recent changes in the dynamic properties of declining arctic sea ice: A model study. *Geophysical Research Letters*, 39(20), 2012.