Morphological Approaches To Understanding Antarctic Sea Ice Thickness

by

M. Jeffrey Mei

B.S., New York University (2015)

Submitted to the Department of Mechanical Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Oceanographic Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and

WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2020

Author .....

Joint Program in Oceanography/Applied Ocean Physics & Engineering
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
August 21, 2020

Certified by .....

Ted Maksym
Associate Scientist, Department of Applied Ocean Physics & Engineering
Woods Hole Oceanographic Institution
Thesis Supervisor

Accepted by .....

Nicolas Hadjiconstantinou
Professor, Mechanical Engineering
Chair, Department Committee on Graduate Students,
Massachusetts Institute of Technology

Accepted by .....

David Ralston
Associate Scientist, Applied Ocean Physics & Engineering Department
Chair, Joint Committee for Applied Ocean Physics and Engineering,
Woods Hole Oceanographic Institution

Morphological Approaches To Understanding Antarctic Sea Ice Thickness

by

M. Jeffrey Mei

Submitted to the Department of Mechanical Engineering
on August 21, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Oceanographic Engineering

Abstract

Sea ice thickness has long been an under-measured quantity, even in the satellite era. The snow surface elevation, which is far easier to measure, cannot be directly converted into sea ice thickness estimates without knowledge or assumption of what proportion of the snow surface consists of snow and ice. We do not fully understand how snow is distributed upon sea ice, in particular around areas with surface deformation. Here, we show that deep learning methods can be used to directly predict snow depth, as well as sea ice thickness, from measurements of surface topography obtained from laser altimetry. We also show that snow surfaces can be texturally distinguished, and that texturally-similar segments have similar snow depths. This can be used to predict snow depth at both local (sub-kilometer) and satellite (25 km) scales with much lower error and bias, and with greater ability to distinguish inter-annual and regional variability than current methods using linear regressions. We find that sea ice thickness can be estimated to $\sim$20% error at the kilometer scale. The success of deep learning methods to predict snow depth and sea ice thickness suggests that such methods may be also applied to temporally/spatially larger datasets like ICESat-2.

Thesis Supervisor: Ted Maksym
Title: Associate Scientist, Department of Applied Ocean Physics & Engineering
Woods Hole Oceanographic Institution

Acknowledgments

Firstly, thanks to my adviser, Dr. Ted Maksym for constantly encouraging me, in particular through the disappointing results of the first qualifying exam. Thank you for letting me find my own direction while keeping me on track. I would also like to thank my committee: Prof. Alexandra Techet, Prof. Brent Minchew and Dr. Yogi Girdhar for their advice and suggestions that greatly improved this thesis.

I am grateful for my peers in the Joint Program, EAPS and MechE for the conversations and good cheer. Ryan, Harry, Rohini, Lizzie, Chris, Scott, Rui, Shourav, Rachel: over lunch chats, Zoom socials and many a board game, you have all kept me sane and made me feel like I was never alone on this journey.

To my Kiwi group: Nicole, Ben, Michael, Richard. Thank you for reminding me of home (and occasionally bringing me NZ snacks) and the riveting debates we have over domestic politics.

To Daphne, Neil, Tejas, Reetik, Amy, Alex, Matt, Chen, Sarah, Vishal: thank you for the games, training and the friendships that carried off-court.

To my parents, thank you for instilling the value of perseverance and independence from a young age. To my sisters, Debra and Judy, thank you for your unconditional love and support throughout my life.

Finally, to Jon: thank you for putting up with my antics while I finished this chapter of my life. Here's to our next chapter together!

This research was funded by National Aeronautics and Space Administration grant numbers NNX15AC69G and 80NSSC20K0972, the US National Science Foundation grant numbers ANT-1341513, ANT-1341606, ANT-1142075 and ANT-1341717, and the WHOI Academic Programs Office.

I would also like to thank Ron Kwok for providing the IceBridge snow depth data; Blake Weissling, Jeff Anderson, Guy Williams, Alek Razdan, Hanumant Singh and the crew of the *RV N. B. Palmer* were instrumental in collecting data during PIPERS, which was led by Steve Ackley as Chief Scientist.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Chapter 1

Introduction

1.1 Importance and Implications of Sea Ice

Sea ice is one of the most sensitive indicators of our changing climate. A layer of sea ice insulates the ocean from radiative heat loss while its high reflectivity (albedo) further reduces the amount of absorbed radiation. In particular, the ice-albedo feedback loop makes the polar regions very sensitive to warming (Holland et al., 2001). Sea ice also plays a large role in regulating heat transfer across the ocean-air interface (Allison et al., 1981; Maykut, 1982). Sea ice drift modifies this heat exchange, and in addition, affects freshwater fluxes and the rate of new sea ice formation by creating open areas of water for new sea ice formation. This has consequences for ocean warming rates and the meridional overturning circulation, with important consequences for climate and biodiversity (Goosse and Fichefet, 1999; Orsi et al., 1999; Massom and Stammerjohn, 2010; Marshall and Speer, 2012; Liu et al., 2020).

Satellites have documented changes in sea ice extent (SIE) for decades (Parkinson and Cavalieri, 2012). General Circulation Models (GCMs), commonly used to model the climate as a fully-coupled system, are generally able to reproduce the decline in Arctic SIE, although they tend to underestimate the true rate, around 4% per decade (Stroeve et al., 2012; Cavalieri and Parkinson, 2012). However, for the Antarctic, GCMs tend to show an erroneous decrease in Antarctic SIE, which actually has a positive decadal trend of around 2% per decade (Zhang and Walsh,

2006; Turner et al., 2013; Comiso et al., 2017). This overall increasing trend is composed of stronger, opposing regional trends (Fig. 1-1). There is a decreasing trend in the Bellingshausen/Amundsen Seas (-2.5% per decade) and an increasing trend in the Ross (4.5% per decade) and Weddell (2.5% per decade) Seas (Comiso et al., 2017). Due to the recent record-low Antarctic SIE from 2016-2018, the overall increasing trend is no longer statistically significant and only the regional decline in the Bell/Am Seas remains significant (Ludescher et al., 2019). In any case, the reasons for the failure of GCMs to resolve Antarctic SIE are not yet fully known, but one issue is the apparent inability of many fully-coupled models to produce accurate atmospheric forcings, leading to incorrect ice drift (and thus incorrect SIE) (Uotila et al., 2014). Using multi-model averaging, Polvani and Smith (2013); Swart and Fyfe (2013) found that the increased Antarctic SIE is statistically consistent with just internal variability. Holland and Kwok (2012) found that wind-driven changes in ice advection can be linked to the decreased SIE in the Weddell (due to deceleration of the Weddell Gyre) and increased SIE in the Bellingshausen/Amundsen and Ross Seas (due to acceleration of the Ross Gyre). Further exploration of possible causes of variability in the Bellingshausen, Amundsen and Weddell Seas is given in Sect. 5.2.2.

Because GCMs typically use a small number of discrete ice thickness bins, this means they do not represent the non-uniform and continuous distribution of sea ice thickness (SIT) well, particularly when the ice is deformed. Sea ice models, which can operate at a higher resolution, can include deformation processes, but typically use a basic thickness redistribution function to represent deformation events. It is therefore unclear if the resultant prediction of SIT distribution is accurate. In tandem with SIE, knowing the SIT distribution allows for the calculation of sea ice volume and mass, which is crucial for working out energy balances, as a change in sea ice volume is equivalent to a specific change in latent heat, and also for estimation of the freshwater exchange between the ice and ocean (e.g. Allison et al., 1981; Maykut, 1982; Holland et al., 1997). Aside from climate implications, sea ice thickness also has important consequences for marine life (e.g. Jenouvrier et al., 2006) and shipping (e.g. Mussells et al., 2017).

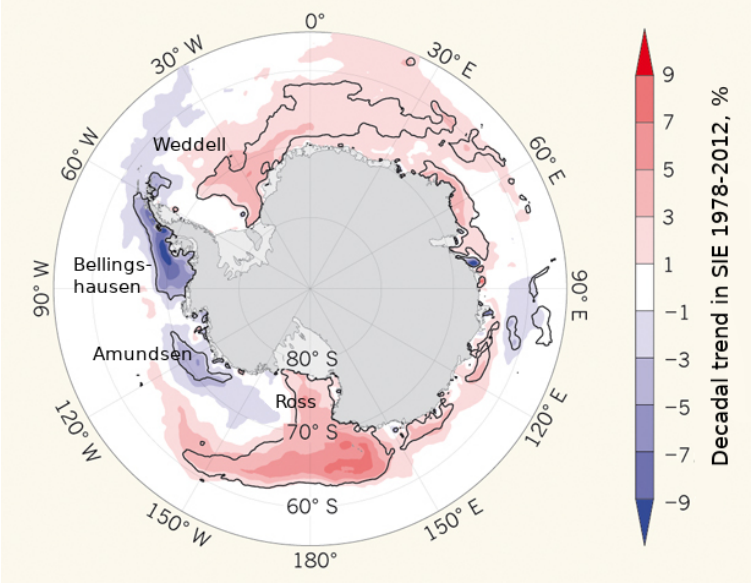

Figure 1-1: The decadal trend in sea ice extent by region. As of 2012, the Ross and Weddell Seas had increasing decadal trends in sea ice extent, whereas the Bellingshausen/Amundsen Seas had decreasing decadal trends. Adapted with permission from Nature Publishing Group: *Nature*, 'Climate science: A resolution of the Antarctic paradox' (King, 2014).

SIT is much harder than SIE to measure remotely. Declines in Arctic SIT over the past several decades have been detected in under-ice upward-looking sonar (ULS) surveys and satellite observations (Rothrock et al., 2008; Kwok and Rothrock, 2009). Arctic ice thickness has been observed with satellite altimetry to continue to decline over the past decade (Kwok and Haas, 2015), with the average ice thickness reducing by 66% (Kwok, 2018). Any possible trends in Antarctic SIT are difficult to detect because of the presumably relatively small changes, and difficulties in estimating SIT in the Antarctic (Kurtz and Markus, 2012; Zwally et al., 2008). Because fully-coupled models (such as GCMS) generally fail to reproduce the observed multi-decadal increase in Antarctic SIE, it is likely that their simulated decrease in Antarctic SIT is also incorrect (Turner et al., 2013; Shu et al., 2015). However, ocean-ice models forced with atmospheric reanalysis correctly reproduce an increasing Antarctic SIE and suggest an increasing SIT (Holland et al., 2014). Massonnet et al. (2013) found that assimilating sea ice models with sea ice concentration shows that SIT covaries positively with SIE at the multi-decadal time scale, and thus implies an increasing

sea ice volume in the Antarctic. Detection of variations in SIT and volume are important to understanding a variety of climate feedbacks (e.g. Holland et al., 2006; Stammerjohn et al., 2008); for example, they are critical to understanding trends and variability in Southern Ocean salinity (e.g. Haumann et al., 2016). At present, large-scale ice thickness cannot be retrieved with sufficient accuracy to detect with any confidence the relatively small trends in thickness expected (Massonnet et al., 2013), or even interannual variability (Kern and Spreen, 2015).

The main source of Antarctic SIT measurements comes from ship-based visual observations (ASPeCt, the Antarctic Sea Ice Processes and Climate program, compiled in Worby et al. (2008)), drill-line measurements (e.g. Tin and Jeffries, 2003; Özsoy-Çiçek et al., 2013), aerial surveys with electromagnetic induction (e.g. Haas et al., 2009) and sporadic data from moored ULS (e.g. Worby et al., 2001; Harms et al., 2001; Behrendt et al., 2013). These are all sparsely conducted, with significant gaps in both time and space, making it hard to infer any variability or trends. There is also some evidence of a sampling bias towards thinner ice due to logistical constraints of ships traversing areas of thick and deformed ice (Williams et al., 2015). Typical mean sea ice thicknesses, based on ASPeCT, range from 0.5-0.8 m in winter (as sea ice begins to form) to 0.9-2.1 m in summer (where the thin sea ice has melted away) (Worby et al., 2008). There is also some regional variability: the western Weddell Sea and (to a lesser extent) the Bellingshausen/Amundsen Seas can contain multi-year ice that can survive the summer melt, and so have thicker ice than other regions.

The only currently-feasible means of obtaining SIT data on a large enough scale to examine thickness variability is through remotely-sensed data, either from large-scale airborne campaigns such as Operation IceBridge (OIB) (Kurtz, 2013), or more broadly from satellite altimetry, (e.g. ICESat (Zwally et al., 2008), or more recently, ICESat-2 (Markus et al., 2017)). Here, SIT is derived from either the measured snow surface (i.e. surface elevation referenced to local sea level) in the case of laser altimeters (ICESat and OIB), or from a measure of the ice surface freeboard (CryoSat-2) (Wingham et al., 2006). The measurement of the surface elevation itself has some error, due to the error in estimating the local sea surface height (Kurtz et al., 2012). When

using radar altimetry, the ice-snow interface may be hard to detect as observations suggest that the radar return can occur from within the snowpack (e.g. Willatt et al., 2009), possibly due to scattering from brine wicked up into the overlying snow, or melt-freeze cycles creating ice lenses, or from the snow-ice interface (Fons and Kurtz, 2019). However, even with an accurate measurement of the snow/ice freeboard, there are challenges with converting this to a SIT estimate.

Assuming hydrostatic equilibrium, the ice thickness T may be related to the snow freeboard F (i.e. snow depth + ice freeboard, see Fig. 1-2) and snow depth D measurements using the relation

$$T = \frac{\rho_w}{\rho_w - \rho_i} F - \frac{\rho_w - \rho_s}{\rho_w - \rho_i} D \tag{1.1}$$

for some densities of ice, water and snow ρ_i, ρ_w, ρ_s (Fig. 1-2). Without simultaneous snow depth estimates (e.g. from passive microwave radiometry (Markus and Cavalieri, 1998) or from ultrawideband snow radar such as that used on OIB (Kwok and Maksym, 2014)), some assumption of snow depth has to be made, or an estimate using empirical fits to field observations is needed (e.g. Özsoy-Çiçek et al., 2013). When averaging over multiple kilometers, and in particular during spring, it is common to assume that there is sufficient snowfall that leads to no ice component in the snow freeboard, i.e. $F = D$ in Eq. 1.1 (Xie et al., 2013; Yi et al., 2011; Kurtz and Markus, 2012). However, this assumption is likely not valid near areas of deformed ice, which may have significant non-zero ice freeboard, and OIB data suggest this is not true at least for much of the spring sea ice pack (Kwok and Maksym, 2014). More generally, empirical fits of SIT to F can be used (Özsoy-Çiçek et al., 2013), but these implicitly assume a constant proportion of snow within the snow freeboard and a constant snow and ice density. These are not likely to be true, particularly at smaller scales and for deformed ice. Moreover, detecting variability with such methods is prone to error because these relationships may change seasonally and interannually. Kern and Spreen (2015) suggested a ballpark error of 50% from ICESat-derived thickness estimates. Kern et al. (2016), following Worby et al. (2008), looked at the snow freeboard as one

layer with some effective density taken as some linear combination of sea ice and snow densities. More recently, Li et al. (2018) has used a regionally- and temporally-varying density (equivalently, a variable proportion of snow in snow freeboard) inferred from the empirical fits of Özsoy-Çiçek et al. (2013), which is equivalent to a more complex, regime-dependent set of snow assumptions. SIT estimates using satellite-based measurements suggest considerable sampling bias in the ship-based observations from Worby et al. (2008) presented earlier this section. Kurtz and Markus (2012) found that ICESat-derived estimates of SIT were consistently higher, up to 44% in some cases, than ship-based observations from the same time and location.

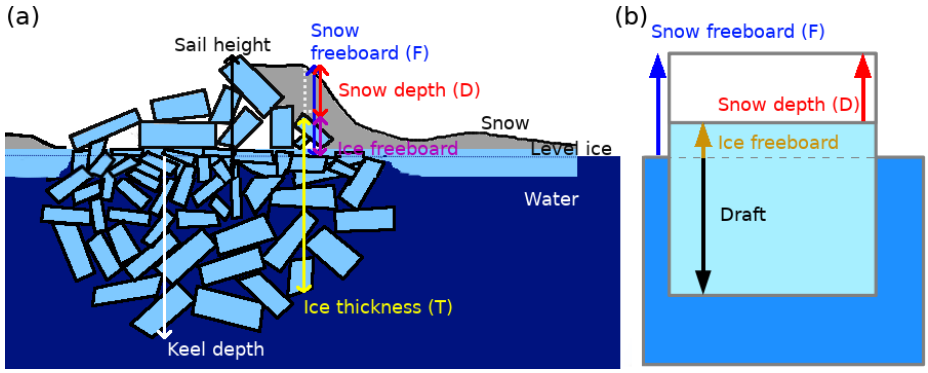

Figure 1-2: (a) A schematic diagram of a typical first-year ridge. The ridge may not be symmetric, and peaks of the sail and keel may not coincide. The effective density of the ice is affected by the air gaps above water and the water gaps below water. T, D and F may be linked by assuming hydrostatic balance (Eq. 1.1). (b) a simplified diagram for level ice to show the different terms we use in this thesis.

A key question is how much the sea ice morphology affects these relationships between surface measurements and thickness. Pressure ridges, which form when sea ice collides, fractures and forms a mound-like structure (Fig. 1-2), are a primary source of deformed ice. Although only a minority of the sea ice surface is deformed, ridges occur at a spatial frequency of 3-30 per km and so may account for a majority of the total sea ice volume (Worby et al., 1996; Haas et al., 1999). The sea ice surface naturally has a varying proportion of deformed ice, which affects the sampling required to faithfully represent the distribution (Weissling et al., 2011). Around deformed areas, both the ice freeboard and snow depth may be high, and we do not yet know the statistical distribution of snow around such deformation features. In this respect,

local estimates of SIT are likely biased low as the average ice freeboard cannot be assumed to be zero. Deformed sea ice may peak at thicknesses exceeding 10-15 m, far higher than the level ice surrounding it (typically 0.5-2.0 m) (Williams et al., 2015). Moreover, the effective density of deformed ice (i.e. the density of the deformed ice including snow-, air- and seawater-filled gaps) may differ significantly from level ice areas due to drained brine and trapped snow in ridge sails, and seawater in large pore spaces in ridge keels (Fig. 1-2; also discussed in Hutchings et al. (2015)). Because these densities affect the empirical fits, it is important to quantify how SIT predictions should be adjusted to account for morphological differences in snow freeboard measurements.

In order to account for the varying effective density of a ridge, we need to be able to characterize different deformed surfaces. The analysis of ridge morphology is currently very simplistic. As summarized in Strub-Klein and Sudom (2012), the geometry of the above-water (sail) and below-water (keel) heights is typically analyzed, traditionally by calculating the sail-keel ratios and sail angles (Timco and Burden, 1997). There are known morphological differences between Arctic and Antarctic ridges, such as sail heights of Antarctic ridges being generally lower than those of Arctic ridges, but these are not known comprehensively (Tin and Jeffries, 2003). According to drilling data and shipboard underway observations, Antarctic ridges have typical sail heights of less than 1 m (Worby et al., 2008) and keel depths of order 2-4 m (Tin and Jeffries, 2003), though much thicker (maximum keel depths >15 m) ridges have also been observed with autonomous underwater vehicles (Williams et al., 2015). Metrics like sail/keel angle are less meaningful in the presence of non-triangular, irregular or highly deformed ridges (e.g. Fig. 1-3), which are underrepresented in literature due to selection bias. Arctic ridges are somewhat more well-studied, with Tucker III and Govoni (1981) finding a square-root relationship between block size and above-water (sail) height, and Timco and Burden (1997) finding a linear relationship between sail height and keel depth but no relationship between sail height and level ice thickness. Ekeberg et al. (2015) found that first-year (Arctic) ridge keels are better characterized by a trapezoid than a triangle, and Petty et al. (2016) found that ice thickness could be

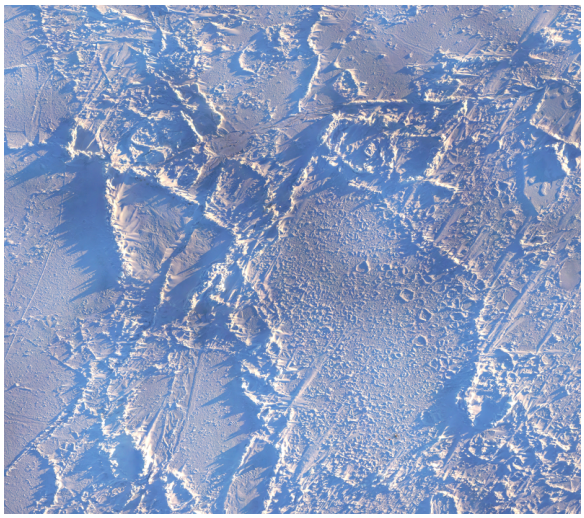

Figure 1-3: Drone imagery (180 m x 180 m) of heavily deformed ice in the Ross Sea, Antarctica. There are multiple ridges which cannot be easily separated. The ridge widths and slopes are varying and must be arbitrarily defined, leading to a variety of possible values. Image provided by Guy Williams.

predicted (with considerable error) from metrics taken from lidar-derived topography of deformed ice. These results may or may not hold for Antarctic ridges. For Antarctic ridges, Tin and Jeffries (2003) found the keel depth was proportional to the level ice thickness around a ridge, and Tin and Jeffries (2001b) found a linear relationship between the ice thickness and snow surface roughness. It is possible that other, more complex metrics may be more relevant for characterizing the relationship between pressure ridge morphology and its corresponding SIT distribution. Tan et al. (2018) found that the ridge shape, along with other metrics like roughness length and ridging intensity, may affect its form drag and hence its drift. Identifying how the morphology of deformed ice can inform estimates of SIT is important for reducing errors on SIT estimates. This is necessary to understanding temporal-spatial variations in SIT using existing measurements of surface elevation.

The uncertainty in sea ice density is also a significant contributing factor to the high uncertainty of SIT estimates (Kern and Spreen, 2015). For example, if assuming zero ice freeboard ($F = D$ in Eq. 1.1) with some known snow density, a 10% uncertainty in the sea ice density can lead to a 50% uncertainty in the SIT. As mentioned before, the effective density may also vary locally, particularly in deformed ice. On

previous Antarctic fieldwork such as SIPEX-II in spring 2012, Hutchings et al. (2015) found the density of first-year ice in the presence of porous granular ice to be as low as 800 kg m^{-3}, a difference of more than 10% from the standard assumption of 900-920 kg m^{-3} (e.g. Worby et al., 2008; Xie et al., 2013; Maksym and Markus, 2008; Zwally et al., 2008; Timco and Weeks, 2010), but in line with the 750-900 kg m^{-3} range found by Urabe and Inoue (1988). This effective density could vary regionally and seasonally in line with ridging frequency, and knowing these variations with greater certainty would decrease the errors in SIT estimations. The effective density may also vary locally around areas of deformed ice, which have varying gap volumes. This means that the scatter in any given linear fit of T and F, and the variability between different fits for different datasets, can be interpreted as differences in effective densities; alternatively, this points out that linear fits will have an irreducible error due to local effective density variations.

1.2 Snow on Sea Ice

Snow on sea ice has important consequences for sea ice thickness changes. In spring and summer, the snow on the sea ice reflects a lot of solar radiation (having a higher albedo than the underlying ice), slowing down the melt. However, in winter, the snow insulates the ice from the cold air, reducing the rate of sea ice formation, while also depressing the ice and exposing it to more melt from the surrounding warmer water (Ledley, 1991; Sturm and Massom, 2009). In the presence of heavy snow, which is particularly the case in the Antarctic, the ice floe can become entirely submerged, leading to the flooding of the snow on top, which can then refreeze, forming snow-ice (e.g. Lange et al., 1990; Adolphs, 1998; Maksym and Jeffries, 2001), which introduces an additional complication to estimating SIT.

When estimating SIT from large-scale snow freeboard data from laser altimetry where no coincident observations of snow depth are available (such as from airborne campaigns including some Operation IceBridge flights where no snow radar data is available (Kurtz, 2013), or more broadly from satellites, e.g. ICESat (Zwally et al.,

2008) or ICESat-2 (Markus et al., 2017)), one must estimate snow depth independently. Common approaches include using an empirical fit to *in-situ* data (Xie et al., 2013) or assuming that there is negligible ice freeboard (i.e. all the ice has been depressed by snow cover, and so $F = D$ in Eq. 1) (Kurtz and Markus, 2012). This is reasonable on a large scale (averaged over many kilometers), over which the majority of the sea ice surface may consist of relatively thin, undeformed ice.Moreover, linear methods to convert surface elevation to snow depth (or ice thickness) implicitly assume a constant snow/ice ratio and also constant snow and ice densities, which may be valid when averaged over large scales but not at local scales (e.g. Özsoy-Çiçek et al., 2011; Xie et al., 2013). To address this, Steer et al. (2016) used empirical fits for two different regimes (high and low ice freeboard), and found that there was considerable variability in the empirical fits for high ice freeboard regimes, in particular at local (sub-kilometer) scales, suggesting that high-freeboard snow surfaces may be morphologically complex. Steer et al. (2016) also noted that linear regressions are much better suited to large-scale mean snow depth predictions, and speculated that sophisticated models that could estimate the proportion of snow within the snow freeboard would improve snow depth estimates. Indeed, most studies of snow on sea ice use large length scales (typically 12.5 km or 25 km), in order to match the resolution of satellite-measured sea ice concentration datasets (e.g. Kwok and Kacimi, 2018). However, higher resolution SIT estimation are not possible without better knowledge of the snow distribution.

Neither the distribution of snow, nor its relationship with the distribution of ice freeboard, is well understood. Kwok and Maksym (2014) found a weak correlation between snow depth and surface roughness (standard deviation of F), which may be the result of snow being blown and accumulating around deformed ice features. This may also be simply because both snow freeboard and roughness, as well as snow depth and snow freeboard, are themselves correlated, especially at large length scales (e.g. Xie et al., 2011; Markus et al., 2011). Snow depth and snow freeboard measurements have been analyzed in both the Arctic and Antarctic. Farrell et al. (2012) showed that the uncertainty in snow depth was the largest contributor to the uncertainty in SIT,

and also noting that the snow depth over heavily ridged (Arctic) ice was sometimes not retrievable. Kwok and Maksym (2014) found that snow depths sampled *in situ* may be biased due to excessively thin and thick ice not being logistically feasible to sample, while radar-based measurements cannot resolve snow depths below 8 cm, and often have poor returns around deformed ice. Zhou et al. (2020) recently showed that different models for snow depth can give high variations for snow density and depth, and in some cases this can be higher than climatology. Similarly, Mallett et al. (2020) showed that assuming a constant snow density, and hence a constant radar penetration speed, biases the snow depth estimate. It is well known that the snow depth distribution is complex and snow features observed *in situ* on sea ice suggest variations in snow depth (Massom et al., 2001; Filhol and Sturm, 2015; Trujillo et al., 2016). Improving our understanding of small-scale snow depth distribution would therefore greatly improve SIT estimates. This also requires more analysis of deformed sea ice surfaces, for which the snow depth distribution is both less understood and harder to observe.

Excluding drill line studies, which are logistically constrained to ice that is easily accessed by ship (not to mention easily drilled) (e.g. Özsoy-Çiçek et al., 2011; Weissling and Ackley, 2011), there are relatively few studies that look at the relationship between snow depth and SIT at the sub-kilometer scale, in particular in the Antarctic. For the Arctic, Petty et al. (2016) used the Operation IceBridge (OIB) Airborne Topographic Mapper (ATM) altimetry data to show that the sea ice surface topography could be related to the sea ice thickness, by using a segmentation approach with high-resolution, three-dimensional lidar data. There have been, to date, no analogous work for the Antarctic.

Based on ASPeCT observations, typical snow depth values for the Antarctic range from 8-15 cm in winter to 16-50 cm in the summer. As with SIT, the Bellingshausen/Amundsen and Weddell Seas, along with the Ross Sea, have deeper snow than other regions. Using airborne radar to measure snow depth also suggests that ship-based observations of snow depth, much as with SIT, are biased low. Kwok and Maksym (2014) found average snow depths in late spring for both the Belling-

shausen/Amundsen and Weddell sectors of around 40 cm, which is around twice the corresponding averages from ASPeCT. In the Bellingshausen/Amundsen sector, the thickest snow (sometimes exceeding 1.5 m) is typically found nearshore around the multiyear ice near the Abbot Ice Shelf. Similarly, the thickest snow in the Weddell (also sometimes exceeding 1.5 m) is found among the multiyear ice in the western Weddell. These areas often lack ship-based observations due to the logistical difficulties of traversing heavily-deformed multiyear ice.

1.3 Remote Sensing of Sea Ice Deformation

As argued in the previous section, deformed sea ice often collects deeper snow, which is both less frequently sampled in drill lines due to logistical constraints, as well as being less likely to give a snow radar return due to the signal noise. Although the snow distribution around such deformation features is not well understood, we can nevertheless observe the deformation features (e.g. pressure ridges).

Many pressure ridges can be observed from above using airborne or terrestrial lidar scans (e.g. Dierking, 1995). However, it is difficult to derive SIT of deformed areas from these scans due to the difficulty in determining the contribution of snow to the snow freeboard measured by a lidar scan. Furthermore, the corresponding (underwater) keel morphology given some surface (lidar) scan, and its effect on the SIT distribution, is not known.

Sea ice deformation can also be observed from below using sonar on autonomous underwater vehicles (AUVs) (e.g. Williams et al., 2015). The lack of snow also makes the deformation structure much more prominent. Although AUV datasets of deformed ice have higher resolution than air- and satellite-borne lidar datasets, they are much more sparsely conducted and fewer such datasets of Antarctic ice exist. This makes it hard to generalize conclusions of deformed sea ice from empirical datasets. It is therefore important to understand how the morphology of deformed ice relates to its thickness distribution. By using coincident, high-resolution and three-dimensional AUV and lidar surveys of deformed ice, we can characterize areas of deformation and

surface morphology and its relationship to ice thickness and snow freeboard much better than with linear, low-resolution drilling profiles.

Surface roughness, among other factors, is known to affect radar-based estimates of snow depth (Stroeve et al., 2006; Markus and Cavalieri, 1998). Özsoy-Çiçek et al. (2011) and Markus et al. (2011) found that snow depth measured by the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E) around deformed ice is underestimated by a factor of two or more. Kern and Spreen (2015) also showed that the error estimate in the SIT is considerably affected by the snow depth error, with a conservative estimate of 30% error in snow depth leading to a relative ice thickness error up to 80%.

Although originally intended for analyzing surface winds, scatterometry, which uses radar backscatter to characterize surface roughness, has also found application in sea ice, namely in distinguishing different sea ice surfaces, both in the Arctic and Antarctic (e.g. Nghiem et al., 1995; Geldsetzer et al., 2007; Gupta et al., 2014; Fraser et al., 2014). In particular, the C-band advanced scatterometer (ASCAT) has been used to distinguish melt-freeze transitions on sea ice as well as surface-based classification (Mortin et al., 2014; Lindell and Long, 2016). It is also possible to coarsely relate these surfaces types to snow depths, using the variance of the backscatter (e.g Yackel et al., 2019), although these are typically done with linear regressions which may not capture the complexity in snow depth variability.

1.4 Deep Learning For Sea Ice Problems

Increased modeling complexity can naturally be achieved with deep learning, which excels in modeling non-linear relationships. Recently, deep learning techniques have been applied to predicting snow depths based of remotely-sensed data. Liu et al. (2019) used an ensemble-based deep neural network using brightness temperature from passive microwave radiometry to predict snow depths with higher accuracy than linear regressions. Braakmann-Folgmann and Donlon (2019) used a simple neural network with gradient ratios from microwave radiometry data to predict snow depth

in the Arctic, with good comparison to OIB snow depths. Wang et al. (2020a) used global navigation satellite system reflectometry and local meteorological station data to predict Arctic snow depths with a mean absolute error of $\sim$10 cm. Convolutional neural networks have also been used to estimate sea ice concentration from SAR imagery (Wang et al., 2016). However, SAR data is better suited for classification of sea ice types than for SIT estimation due to the ambiguity of the backscatter, as it is not sensitive to snow depth (Aldenhoff et al., 2019).

The textural analysis of snow surfaces is a nascent field, even though textural segmentation has been a long-studied problem in computer vision (e.g Beck et al., 1983). A variety of methods have been developed for this purpose, with differing advantages and disadvantages. As the evaluation of good textural segmentation results ultimately requires comparing to the subjective human eye, there is no single metric used to compare different methods, and as a result there is no single best method. Popular methods to achieve textural segmentation commonly include the use of either spatial or color-based metrics, or some combination thereof (Chen et al., 2005). As our goal is to apply textural segmentation to lidar scans (analogous to grayscale imagery, as it has one input channel), we focus here on spatial metrics only.

The most common spatial metrics involve decomposing the image using some kind of convolutional filter. This essentially decomposes the image based on multiple frequencies of interest. Common examples include the Gabor transform (Porat and Zeevi, 1989), steerable pyramid decomposition (Simoncelli et al., 1992; Simoncelli and Freeman, 1995), and the discrete wavelet transform (Cohen et al., 1992). In the context of sea ice, textural analysis has been largely limited to synthetic aperture radar (SAR) imagery for classification of the surfaces into some fixed number of classes. This includes the use of gray level co-occurrence matrices (GLCM) (e.g. Soh and Tsatsoulis, 1999; Clausi and Yue, 2004), wavelet transforms (e.g. Yu et al., 2002), and neural networks (e.g. Zakhvatkina et al., 2013; Ressel et al., 2015). A detailed review of the state of sea ice classification using SAR data can be found in Zakhvatkina et al. (2019). There are few studies linking SIT to SAR images: Kim et al. (2012) found a weak linear relationship between SIT and depolarization of return

radar signatures may exist for deformed ice surfaces; Shi et al. (2014) used a linear model with various SAR parameters to predict SIT; and Zhang et al. (2016) found that the thickness of undeformed first-year ice < 0.8 m could be exponentially related to a ratio of the polarimetric scattering returns. Deep learning techniques show great promise for sea ice problems and can have a large variety of input data.

1.5 Thesis structure

Sea ice thickness is hard to estimate due to the uncertainty in estimating the amount of snow within the snow surface. Current methods using linear fits to predict snow depths from the snow surface have high errors ($\sim$50%) and do not generalize well to other regions/datasets due to varying snow/ice proportions. However, snow depths form features, which implies that the snow surface morphology contains information that can be used to infer the snow depth. This thesis therefore seeks to quantify this relationship, i.e. to identify how to predict snow depth (and ice thickness) from measurements of the snow surface. In Chap. 2, we present the data we will use to solve this problem. In Chap. 3, we show that sea ice thickness can be directly predicted (with an error $<$20%) with a convolutional neural network, which plausibly learns to account for differences in the effective density of the snow/ice surface layer for different surface types, for a small Ross Sea dataset. We extend this analysis to the much larger Operation IceBridge dataset in Chap. 4, and show that the snow depth in the Weddell can be predicted with a convolutional neural network, with much better generalization than a linear fit to surface elevation data from a different year. This is extended even further in Chap. 5, where we show that convolutional neural networks trained on Weddell datasets can generalize to Bellingshausen/Amundsen datasets, suggesting that they have similar morphological features. We also show that predictions from convolutional neural networks have lower bias than linear fits and hence predict the overall snow depth distribution more accurately, which means that they can resolve interannual and regional variability whereas linear fits cannot. We also find that our predicted snow depth, with error $\sim$15%, results in a sea ice

thickness uncertainty of ∼20%, which is dominated by the uncertainty in sea ice density. Possibilities to extend this work are discussed in Sect. 6.

Chapter 2

Data

In order to assess the relationship between snow surface morphology and the sea ice thickness associated with it, we need to compile snow surface, snow depth and sea ice thickness data. These are described below. They constitute high-resolution 'layer-cakes' with snow depth, ice freeboard and ice draft (Sect. 2.1) to test the viability of surface features to predict SIT, and slightly lower-resolution, but much larger-scale surface morphology and snow depth data from Operation IceBridge (Sect. 2.2) to identify if the method can be extended to larger temporal-spatial scales. As the raw data is not gridded, some preprocessing is needed to convert the various data into gridded format (Sect 2.3).

2.1 PIPERS cruise

The author took part in the PIPERS (Polynyas, Ice Production, and seasonal Evolution in the Ross Sea) expedition from early April to early June 2017 (Fig. 2-2). In total, 6 AUV surveys of the ice draft (below-water ice thickness) were conducted, specifically targeting deformed surfaces. Of these, 4 coincided with snow depth measurements and a lidar survey of the snow freeboard (= sea level referenced surface elevation), thus providing a 'layer-cake' of snow depth, ice freeboard and ice draft data (following Williams et al. (2013)). These 4 layer cakes are shown in Fig. 2-3. There are two other AUV scans which lack lidar/snow measurements so are not included

in our analysis. The AUV surveys were done with a Seabed-class AUV from the Woods Hole Oceanographic Institution (Fig. 2-1a) following Williams et al. (2015), with a swath multibeam sonar (Imagenex 837 DeltaT) at a depth of 15-20 m in a lawnmower pattern (equally spaced passes under the ice in alternating directions). Adjacent passes were spaced to provide approximately 50% overlap in consecutive swaths, with at least one pass across the grid in the transverse direction to allow corrections for sonar orientation in the stitching together of the final sonar map. The AUV multibeam data were processed to correct for vehicle pose, then individual swaths were stitched together, with empirical corrections to pitch and roll offsets due to the physical placement of the sensors. These corrections were determined by minimizing differences in drafts for overlapping portions of adjacent swaths. This largely follows the methodology in Williams et al. (2015), although Simultaneous Localization and Mapping (SLAM) algorithms were not applied here as the quality of the multibeam maps were determined to be comparable to those without SLAM processing, and any improvements in resolving small-scale features would not affect the analysis here. Vehicle navigation used a combination of a Long-baseline transponder array placed outside of the survey area to position the AUV survey within the surveyed floe reference frame (accuracy $\sim$1-5 m), while high-accuracy relative positioning (accuracy better than 1 m) was achieved by ice-bottom tracking using an acoustic doppler current profiler Doppler velocity log. Any offsets in the AUV and surface survey reference frames were corrected as described below. The vertical error in draft is estimated at 10 cm over deformed areas and $<$3 cm for level areas (Williams et al., 2015). The scans were ultimately binned at 0.2 m horizontal resolution. The snow freeboard scans were done with a Riegl VZ-1000 terrestrial lidar scanner, using 3-5 scans from different sides of a 100 m x 100 m grid to minimize shadows, which were stitched together using tripod-mounted reflective targets placed around the grid. We scanned at the highest laser pulse repetition rate of 300 kHz, with an effective maximum range of 450 m. The accuracy and precision at this pulse rate are 8 mm and 5 mm respectively. All composited and registered scans for a particular site were height-adjusted to a sea-level datum using a minimum of 3 drill holes for sea level

references. The output point cloud was binned at 0.2 m resolution, and any small shadows were interpolated over with natural neighbor interpolation (Sibson, 1981). The snow depth measurements were done last, using a MagnaProbe, a commercial probe by Snow-Hydro LLC with negligible vertical error when measuring snow depth on top of ice (Sturm and Holmgren, 2018; Eicken and Salganek, 2010). The probe penetrates the snow and automatically records the snow depth. The ice thickness can then be calculated by taking (draft) + (snow freeboard) - (snow depth). Note that because of thin snow, a negligible portion of the ice had negative freeboard. Where they do tend to occur (in deeper snow adjacent to ridges), the effect on isostacy at the spatial scales considered here will also be negligible because of the much thicker ice.

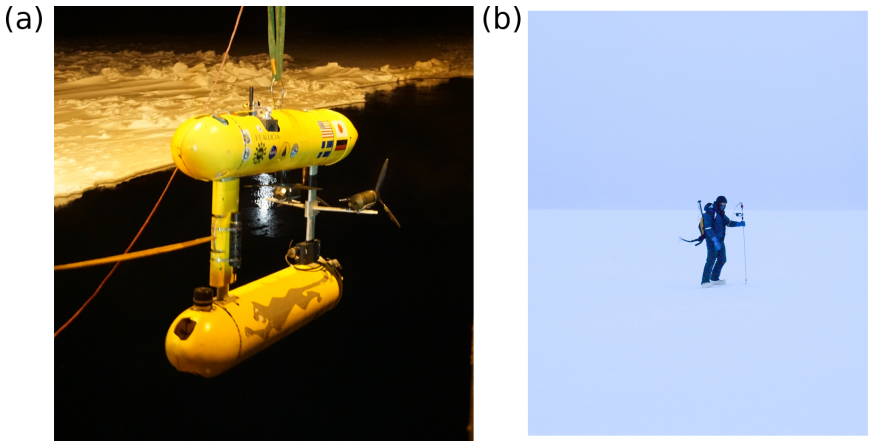

Figure 2-1: (a) The autonomous underwater vehicle (AUV) used to collect sea ice draft data. (b) An example of the author conducting a snow depth survey with the MagnaProbe.

The lidar and AUV data were corrected with a constant offset, estimated by aligning with the mean measurements of the level areas of the drill line for each floe (Fig. 2-4). It is important to use the level areas only as drill line measurements are likely to be biased low due to the difficulties of getting the drill on top of sails, potential small errors in alignment of the drilling line relative to the AUV survey, differences in thickness measurement in highly deformed areas (the drilling line samples at a point, while the AUV will be some average over the sonar footprint) and the presence of seawater-filled gaps that may be confused with the ice-ocean interface when drilling.

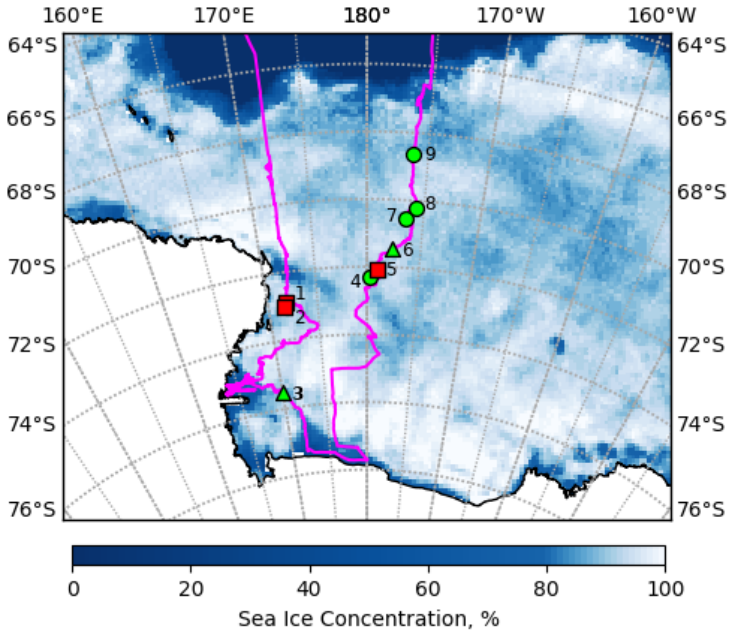

Figure 2-2: PIPERS track (magenta) with locations of ice stations labeled. Stations with AUV scans are shown in green (3, 4, 6, 7, 8 and 9) and the other stations (1, 2 and 5) are shown with red squares. Stations 4, 7, 8 and 9 (green circles) also have a snow freeboard scan and snow depth measurements; these are shown in Fig. 2-3. Other stations have some combination of missing lidar/AUV/snow data. Station dates were 05/14 for station 3, 05/24 for station 4, 05/27 for station 6, 05/29 for station 7, 05/31 for station 8 and 06/02 for station 9. Overlain is the sea ice concentration data (5-day median) for 06/02/2017 from ASI-SSMI (Kaleschke et al., 2017).

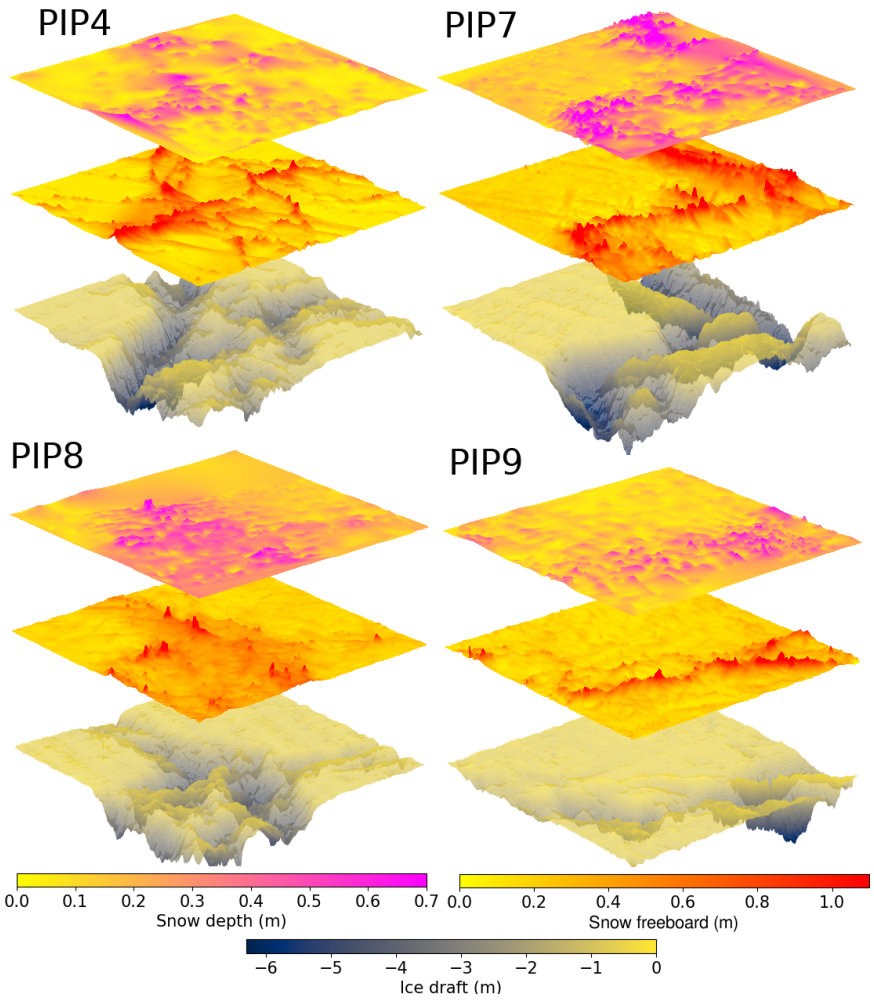

Figure 2-3: Sea ice/snow layer cakes from PIPERS. The top layer is the snow depth (D), the middle layer is the lidar scan of the snow freeboard (F), and the bottom layer is the AUV scan of the ice draft. The ice thickness is therefore given by ice draft + snow freeboard - snow depth.

The order of the lidar correction is ∼1 cm and the order of the AUV correction is ∼10 cm. This offset accounts for errors in estimating the sea level at lidar scan reference points and the AUV depth sensor and vehicle trim. As the AUV surface is similar to the lidar surface, we also use a Enhanced Correlation Coefficient method to align the surfaces (Evangelidis and Psarakis, 2008). This assumes that, on average, the lidar surface features coincide with AUV surface features, and that any spatial offsets are therefore isotropic and hence there is no resultant bias in any single direction. This correction is generally of order a few meters; because we are concerned with the mean sea ice thickness over a 20 m (window) scale, there is little impact on our results.

Figure 2-4: An example of a drill line during PIPERS. At each drill location (spaced 2 m apart), a small area is cleared of snow for ease of drilling, clearly visible here. Drill sites tend to be selected for their ease of access, and so are biased towards thinner and flatter ice.

Summary statistics for the floes sampled during PIPERS are in Table 2.1. The PIPERS surveys comprised floes with ridges that had sails and keels significantly thicker than those that are typically sampled in drilling transects (e.g. Tin and Jeffries, 2003; Worby et al., 2008). The sail/keel angles (the angle of the sail/keel slope relative to vertical) are not as well-defined for complex, non-linear ridges, so a range of angles is given, based on the variety of slopes measured across the deformed area. The 99th percentile for the sail/keel height is also reported to inhibit the effect of outliers from the lidar/AUV scans. We found the sail/keel ratio was much more consistent when using the 99th percentile values. Our sail angles are typically $< 10^o$ and our keel

angles are typically $< 20^o$, in line with averaged values from Tin and Jeffries (2003). However, our sail heights and keel depths are slightly larger in magnitude than the averaged Antarctic values from Tin and Jeffries (2003), and are more similar to their reported values for temperate Arctic ridges. Although our sampled ridges seem to be morphologically typical of Antarctic ridges, they are somewhat thicker than those typically sampled in drilling transects, which is consistent with Williams et al. (2015), who suggested that drilling transects may undersample thicker ice. The ratio of keel depth and snow-sail height for our PIPERS dataset is 3.9, in line with a ratio of 3.6 from 204 drill profiles of Antarctic sea ice examined by Tin and Jeffries (2001a), and also consistent with a ratio of 4.4 for first-year Arctic ridges from Timco and Burden (1997).

Table 2.1: Standard metrics calculated for PIPERS dataset: Sail height (H_S), sail angle (A_S), the surface roughness (here taken as the standard deviation of the snow freeboard, σ), mean snow freeboard ($\bar{F}$), keel depth (H_K), keel angle (A_K), mean thickness ($\bar{I}$), mean level ice thickness ($\bar{I}_L$), mean deformed ice thickness ($\bar{I}_D$), sail-to-keel ratio (H_S/H_K) and % deformation. For H_S and H_K, the absolute maximum is given, along with the 99th percentile value of the deformed section draft (in brackets). The amount of deformed ice in each scan is generally high as the survey grids were deliberately chosen for their deformation. The sail/keel angles are not precisely defined because the deformed surfaces are complex and non-linear, and a range of slopes across the deformed surface are given.

	H_S (m)	A_S (^{o})	σ (m)	$\bar{F}$ (m)	H_K (m)	A_K(^{o})	$\bar{I}$ (m)	$\bar{I}_L$ (m)	$\bar{I}_D$ (m)	H_S/H_K	DEF
PIP4	1.64 (1.33)	6-40	0.20	0.28	7.43 (6.53)	15-25	1.72	0.65	2.19	0.22 (0.20)	71%
PIP7	2.02 (1.53)	3-7	0.26	0.37	7.30 (6.84)	13-17	2.20	0.47	3.49	0.28 (0.22)	57%
PIP8	1.95 (1.16)	1-6	0.15	0.27	5.70 (5.32)	6-14	1.33	0.57	2.08	0.34 (0.22)	50%
PIP9	1.82 (1.27)	6-13	0.15	0.24	6.57 (5.93)	9-34	0.91	0.59	2.01	0.28 (0.21)	23%

2.1.1 Accuracy of drill line

Drill lines are known to have a sampling bias, and this is apparent in our PIPERS data too. Fig. 2-5 shows a representative floe. Drill lines tend to be the most accurate measure of snow/freeboard/ice thickness on level areas, although they may have sampling issues over deformed surfaces, as discussed below. Comparing the lidar and AUV scans of the level areas allows us to correct for any biases. We therefore compare the scan-derived ice thickness to the drill lines conducted during the ice

stations. As can be seen in Fig. 2-5, the drill line largely agrees with the lidar/AUV scans, with some notable caveats:

- The way the drill line is sampled leads to a considerable bias against high freeboards. As it is harder to drill vertically down near sails, the drill locations are often chosen to be more accessible points located 'nearby'. These skew the drilled freeboards lower, as is very noticeable in Fig. 2-5. This does not affect the drilled drafts as much, though they may also be undermeasured due to incomplete drilling (e.g. hitting a pocket of water and not drilling beyond that).

- Due to the above selection bias, the drill locations are not necessarily spaced exactly 2 m apart. Near deformation areas, the freeboard can sharply jump between high and low values, which often leads to choosing an easier drill location by shifting the point slightly, which can distort the measurements laterally.

- The tape measure used for reference is forced to stretch over/around deformed ice areas, which means the drill line (nominally 100 m long) is actually only 90-100 m long, but is recorded as fitting a 100 m line. This compresses the deformation features, as is evident in the drilled draft measurements in Fig. 2-5 near x = 60 m when compared to the AUV-derived draft.

These issues with drill lines are most prominent near areas of deformed ice; we expect the level ice areas to have more precisely-spaced and less biased points. Because each drill line point cannot be directly compared with lidar elevations chosen to be 2 m apart to mimic the drill line, a more useful measure of error is the difference in means of the drilled freeboards/drafts vs. the instrument-surveyed freeboards/drafts for the same patch. For example, in Fig. 2-5, the difference between the measured and interpolated mean elevation of the level areas for the ice freeboard and ice draft are 0.08 cm (3.5%) and 0.5 cm (0.8%), but for the entire drill line, the difference is 6.0 cm (73%) for the freeboard and 12.1 cm (9.5%) for the draft. This reveals the considerable effects of the freeboard sampling bias in comparison to the draft samples.

That drill lines tend to underestimate sea ice thickness is consistent with conclusions from Williams et al. (2015).

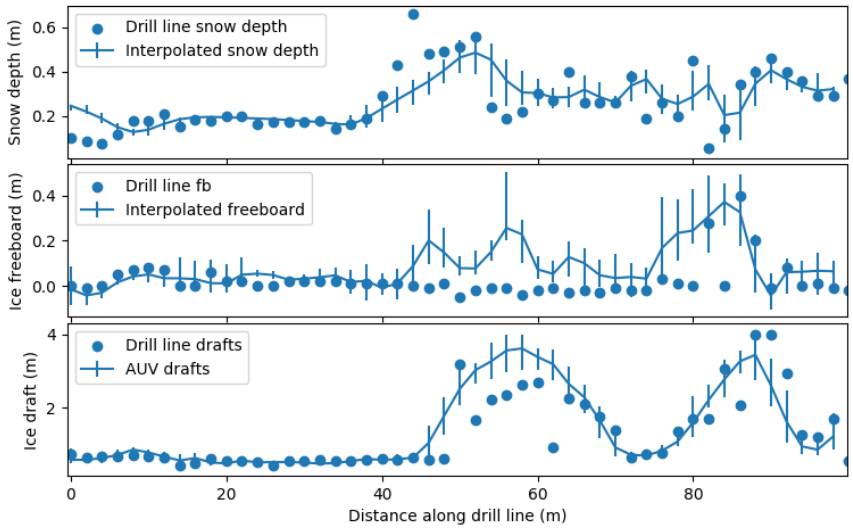

Figure 2-5: Comparison of drill line measurements vs. surface interpolations for snow depth (top), ice freeboard (middle) and ice draft (bottom) for an example ice station (PIP4). The vertical lines represent the maximum and minimum values within a ± 1 m window, as the drill line measurements are not spaced precisely 2 m apart. The freeboards, and to a lesser extent the drafts, are typically undermeasured. The deformation features are typically stretched.

2.1.2 Empirical fits to SIT metrics

Past analyses of the relationship between various metrics and ice thickness have relied on low-resolution drill lines. We therefore repeat the analyses with our higher-resolution dataset to see if the relationships still hold.

Tucker III et al. (1984) found the thickness of the level ice forming a sail (L) and its sail height (S), assuming buckling failure, could be related as $S \propto L^{0.5}$. Tin and Jeffries (2003), following Melling and Riedel (1996), assumed that the sail height (S) could be related to the keel depth (H) as $H = 5S$, and thus the keel depth could be related to the level ice thickness as $H = aL^{0.5}$, and found $a = 5$ for a dataset from the Ross Sea. This coefficient of 5 is lower than the coefficients (15-20) for a variety of ridges in the Beaufort Sea (Tucker III et al., 1984; Melling and Riedel, 1996). When applied to our PIPERS dataset, we get $a = 6.7 \pm 0.7$. Following

Leppäranta and Hakala (1992), Tucker III et al. (1984) and Timco and Sayed (1986), which found the range for the exponent could not be narrowed beyond beyond 0.5-1.0, we also try fitting a linear regression (with no intercept), giving $a = 9.3 \pm 1.5$. We expand this regression to include additional AUV scans from other expeditions (SIPEX-II, East Antarctica, September-October 2012 (Williams et al., 2015); IceBell, Weddell and Bellingshausen Seas, November 2010 (Williams et al., 2015); SeaState, Arctic, October 2015 (Williams et al., 2018)) for a total of 20 scans spanning a much wider range of keel depths (Fig. 2-6). We obtain $a = 6.8 \pm 0.4$ for the square-root relationship and $a = 6.5 \pm 0.7$ for the linear relationship. As the scatter is high, we select the best model using the Akaike Information Criterion (Akaike, 1974) as the R^2 is not well-defined for a fit with no constant term. In both the PIPERS-only and full-AUV datasets, the square-root relationship was a better model than the linear relationship, even if an intercept was included in the linear regression. Our coefficient of $a = 6.8$ is similar to the coefficient of 5 from Tin and Jeffries (2003), and is similarly lower than the coefficients found for Arctic ridges, suggesting a possible morphological difference between Arctic and Antarctic ridges. We also performed a monomial fit to identify the best exponent of L, which gave $H = 6.4L^{0.38}$. This had a marginally smaller AIC than the square-root fit, though this exponent is not within the range of 0.5-1 suggested by Timco and Sayed (1986) and Tucker III et al. (1984). In any case, both the square-root and monomial fits have considerably lower AICs than the linear fit, which suggests that the exponent is likely closer to 0.5 than 1.0.

It is reasonable to expect that rougher ice, which is generally older and more deformed, should be thicker. Tin and Jeffries (2001b) found a linear relationship between the large-scale (1 m resolution, over 150 m) RMS roughness of snow surface and its thickness, and also a linear relationship between the RMS of the snow surface and the RMS of the draft. Taking their ratio, we can estimate the linear relationship between the RMS of the draft and the ice thickness, which we can compare to our AUV data (Fig. 2-7). Tin and Jeffries (2001b) found that the ice thickness was 5.5× the snow surface roughness, and their snow surface roughness was 1/3.7 of the ice bottom roughness. So, approximately, the ice thickness would be 5.5/3.7 ×

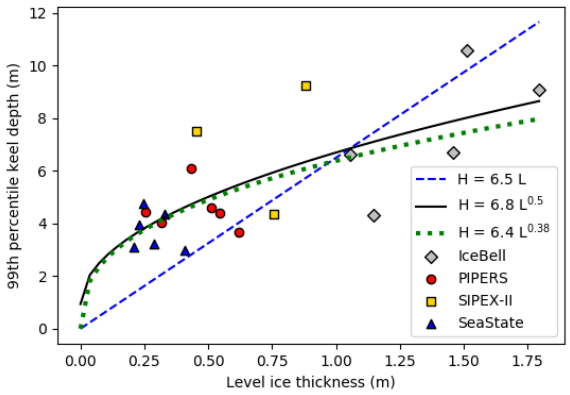

Figure 2-6: Level ice thickness vs keel depth (defined as the 99th percentile draft), following theoretical relationships described in Tucker III et al. (1984); Tin and Jeffries (2003). The square-root fit (black) has a much lower AIC (75.9) than the linear fit (blue, AIC = 92.7), and the monomial fit (green) has a slightly lower AIC (75.4) than both of them. The mean relative errors in predicting keel depths compares similarly, with mean relative errors of 41%, 24% and 22% for the linear, square-root and monomial fits.

the ice bottom roughness, giving a factor of 1.5. As our non-PIPERS AUV data do not come with corresponding surface measurements, we use the draft mean as an estimate of the mean thickness. We find a ratio of (floe thickness)/(floe bottom roughness) = 1.5±0.1 for our AUV dataset (Fig. 2-7). As only PIPERS data have lidar scans, we can only analyze the thickness/snow-surface roughness ratio for the four full ice stations from Fig. 2-3 in the PIPERS dataset. This gives a ratio of 8.2±0.5, with a mean relative prediction error of the mean thickness of 12%. The ratio of surface to bottom roughness is $1/(6.5 \pm 0.5)$. This gives a ratio of 1.3±0.1, which agrees well with the value of 1.5 from Tin and Jeffries (2001b), even though the surface roughness/thickness (8.2) and surface roughness/bottom roughness (6.5) ratios are both significantly different to those from Tin and Jeffries (2001b) (5.5 and 3.7 respectively). This is likely because the surface roughness is often obscured by snow, which may be inconsistent between different floes, whereas the bottom roughness is consistently formed from breaking ice. This suggests the above-water RMS roughness may not have as consistent a relationship with thickness as the below-water roughness. We repeat the analysis using local roughness (the RMS of a 20m x 20m window at resolution 0.2 m) and local mean draft (Fig. 2-7). This has mean

relative errors of 30-50%, which is considerably more than the scan-averaged mean relative errors (10-20%). The window size was chosen by examining the range of the semivariogram for the floes, which was near 25 m, which we expect to represent the maximum feature length scale. We chose 20 m windows to balance this with the need for a smaller window size to ensure a larger number of windows (= data points) for our analysis.

Similarly, we may expect rougher areas to trap more snow (Kwok and Maksym, 2014). Although Kwok and Maksym (2014) averaged the snow depth and surface roughness over a much larger scale than our PIPERS dataset (4 km scale at resolutions 1-10 m), we also find snow accumulates preferentially in areas of deformation. We find snow depth at a 20 m scale is approximated by 0.80 $\times$ the surface roughness + a constant of 0.12 m. Our dataset, which is from early winter in the Ross Sea, has a similar relationship to their dataset from the Weddell Sea in Spring (slope: 0.83, intercept: 0.16 m), and our correlation ($R = 0.66$) is also comparable to theirs ($R = 0.71$). It is not surprising that snow accumulates in areas of deformation, but the relatively high scatter in using a simple linear model motivates more advanced analysis of the topography. RMS roughness is too simple, and does not account for spatial features, as any permutation of points within a grid would have the same RMS.

Worby et al. (2008) identified a relationship for estimating sea ice thickness, assuming a triangular sail and keel, as $T = 2.7RS + Z_u$ for some deformed proportion R, sail height S and level thickness Z_u. Given how non-triangular our ridges are, we are not surprised to find that this does not compare well at all to our PIPERS dataset, with average errors for the mean thickness >100%. (Using the 99th percentile values for the sail height, which would be closer to the effective height of an equivalent-volume triangular sail, is slightly better, with an average error of 56%).

Our results relating ridge morphology to other metrics largely agree with prior literature, despite the difference in resolutions, though the high scatters suggest that these general relationships may be oversimplified.

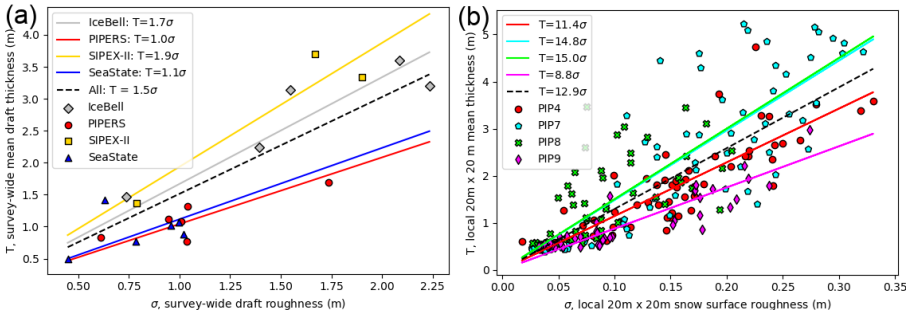

Figure 2-7: (Left) Floe-wide RMS roughness vs. floe draft mean thickness for the different AUV datasets, to be compared against a slope of 1.5 from Tin and Jeffries (2001b). Our fit for all data (black line) also has a slope of 1.5. The resolution is 0.5 m. PIPERS and SeaState largely focused on first-year ridges, whereas Icebell data is from consolidated late Spring (potentially with multi-year ice), and SIPEX is from early Spring. Fits to the individual datasets are color-coded. The mean relative error in the predicted mean thicknesses are 11%, 17%, 12%, and 18% for IceBell, PIPERS, SIPEX-II and SeaState respectively, and 33% for all data. (Right) The same analysis but for the local (20 m x 20 m) RMS roughness plotted against the local mean draft for the different AUV datasets. The mean relative errors in the predicted mean thickness are 38%, 49%, 29% and 39% for Icebell, PIPERS, SIPEX-II and SeaState respectively, and 50% for all datasets combined.

2.2 Operation IceBridge

The flights from OIB for the Weddell and Bellingshausen/Amundsen Seas (Bell/Am) each have two different flight tracks. These 12 flights are shown in Fig. 2-8; details on which flights correspond to which tracks are in Table 2.2. The OIB data we are using consists of the lidar surface elevation data, and the snow radar (processed into snow depth estimates following Kwok et al. (2011); Kwok and Maksym (2014)). There are additional Weddell flights which are not yet processed, which could be potentially added to this analysis (2010/10/26, 2011/10/12, 2011/10/13, 2011/10/18, 2012/11/06, 2016/10/27, 2018/10/19).

2.2.1 Lidar elevation data

The ATM surface elevation data consists of a conically-scanning laser altimeter with a footprint of ∼1 m, swath width of ∼250 m, vertical precision/accuracy of 3 cm/6.6 cm and shot spacing of a few meters (Studinger, 2018). Before the snow surface can be

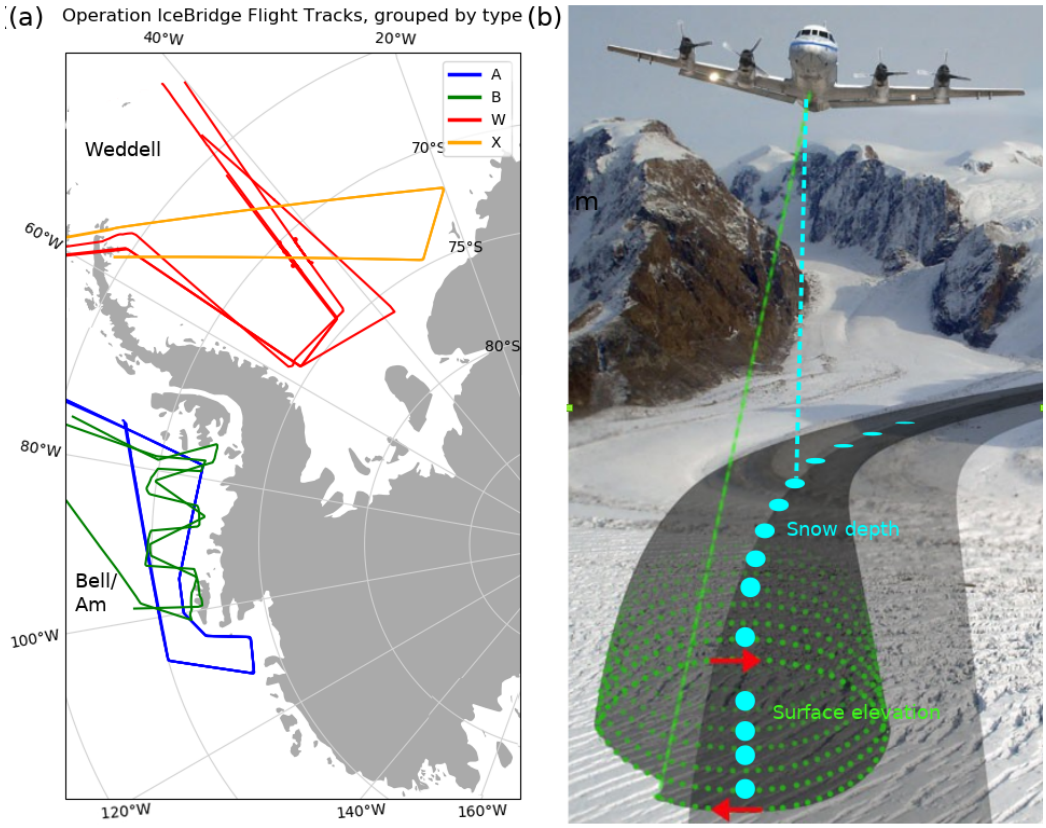

Figure 2-8: (a) The OIB flights in the Bellingshausen/Amundsen (Bell/Am) seas typically have two tracks - these will be referred to as A and B. The Weddell flights also have two typical tracks - these will be referred to as W and X. A list of all flights and their track types is given in Table 2.2, (b) A diagram of the Operation IceBridge flight showing the conical-scanning laser for surface elevation (F) data and snow depth radar for snow depth (D) data (image adapted from `https://atm.wff.nasa.gov/`).

Table 2.2: A list of all flights used in this study and their abbreviated flight type.

Flight Date (YYYY/MM/DD)	Abbreviation
2010/10/28	2010W
2010/10/30	2010A
2011/10/23	2011A
2011/10/25	2011X
2012/10/13	2012A
2012/10/19	2012B
2012/11/07	2012W
2014/10/16	2014W
2014/10/20	2014X
2014/11/13	2014A
2016/10/17	2016W
2016/10/22	2016B
2018/11/05	2018B

analyzed, the ellipsoid-referenced (World Geodetic System 1984) surface elevations must first be adjusted for the local sea surface height. This is done by using a reference geoid (EGM2008), which is accurate to a few meters, and then fine-tuned using nearby areas of open water (leads). This is further described in Sect. 2.3.2 and also Appendix B.

2.2.2 Snow depth data

The OIB snow depth measurements are collected using a frequency-modulated continuous-wave snow radar (Panzer et al., 2013). This type of radar has been demonstrated in one field study to accurately retrieve snow depth (Kanagaratnam et al., 2007). The footprint (corresponding to the First Fresnel Zone) is ∼9 m, with ∼1 m along-track resolution, though the effective resolution is 5.6 m due to decimation and boxcar-filtering of the signal (Kurtz and Farrell, 2011). More details can be found in Panzer et al. (2013); Kwok and Maksym (2014).

Snow depths are retrieved from the OIB L1B Radar Echo Strength Profiles (Paden et al., 2014). Snow radars were first proposed by Vickers and Rose (1972), but this had a short pulse and hence poor vertical resolution. Kanagaratnam et al. (2007) used a linear sweep over a wide bandwidth to dramatically improve vertical resolution. the OIB Frequency-Modulated Continuous Wavelength snow radar is based on this, using a linear sweep between 2-8 GHz, giving a range resolution of ∼3 cm. The radar simultaneously receives and transmits slightly different frequencies, leading to a 'beat' phenomenon. Of the two beat frequencies, the higher one is removed via a low-pass filter, and then the beat frequency is linearly proportional to the range (distance). There are typically two peaks, corresponding to the air-snow and snow-ice interfaces; their difference is hence the snow depth. To account for the different travel speed of electromagnetic waves in snow, the return paths were scaled using the first-order approximation of the relationship between $\frac{v}{c}$ (the ratio of its speed compared to the speed of light in vacuum) and density ρ (g cm^{-3}), $\frac{v}{c} = \frac{1}{\sqrt{2\rho+1}}$ (Tiuri et al., 1984). Taking a typical density of 300 kg m^{-3} following studies like Massom et al. (1997); Arndt and Paul (2018), this gives $\frac{v}{c} = 0.79$. For our datasets, 30-50% of the radar

returns were successfully processed into snow depths. This may introduce a sampling bias; see Sect. 4.4.1 for more details. The average number of successful snow radar returns is 14 per 180 m, or equivalently an average spacing of 12.5 m between returns.

There are several processing methods, differing in how they identify the air-snow and snow-ice interfaces, recently summarized by Kwok et al. (2017). To discuss a few, Kurtz and Farrell (2011) uses empirically-derived thresholds for peak-finding, such that when an unambiguous maximum is not found, the nearest range bin with a return power above a threshold is assigned. Newman et al. (2014) developed a wavelet-based method to distinguish the interfaces, using Haar wavelets to find sudden transitions (i.e. peaks) within signals. Kwok et al. (2011); Kwok and Haas (2015), checks the radar returns to find the signal peaks 6 dB above local noise threshold. All methods involve quality-control to identify 'bad' snow depths, however defined, and to account for sidelobes. Sidelobes occur due to the radar chirp being a finite-time signal. These may be confused with the secondary return from the air-snow interface (typically weaker than the return from the snow-ice interface). These can be reduced by reducing the bandwidth, at the cost of reducing range-resolution, or otherwise there are *ad hoc* algorithms using signal-to-noise ratio from previous returns, such as the algorithm in Kwok and Haas (2015). Kwok and Haas (2015) also uses the OIB lidar data to identify air-snow interfaces, and discards those snow depths whose air-snow interfaces are a poor match to the OIB lidar. Recently, deep learning methods have also been used to process the radar returns (e.g. Holt et al., 2015).

Although we expect similar results using any snow depth dataset, because our algorithm will extrapolate snow depths, we choose the methodology in Kwok and Haas (2015) as this has stricter quality controls on the snow returns, leading to fewer data points but with more certainty about those data points. Fewer data is of less concern for our analysis as we will be extrapolating the snow depth data anyway.

One cause of ambiguity in the interface identification is (upward) brine rejection from sea ice formation, as well as meltponds from melting snow (primarily in the Arctic), which make the snow-ice and air-snow interfaces, respectively, ambiguous. The water induces dielectric loss, which causes the returned snow depth to be under-

estimated. Similarly, the presence of snow-ice (formed by flooded snow refreezing) complicates the snow radar return, as the resulting snow-slush interface can give a stronger return than the snow-ice interface (Panzer et al., 2013).

2.3 Preprocessing

2.3.1 Floe motion correction for *in situ* data

The snow depth data during PIPERS was collected by the author on a moving (rotating and translating) floe. In order to account for the floe rotation, the snow probe was mounted with an Emlid Reach Real-Time Kinematic GPS, referenced to base stations on the floe, which allowed for more precise localization of snow depth. Using Post-Processed Kinematic (PPK) techniques with the open-source RTKLIB library and correcting for floe displacement/rotation, the localization accuracy was ∼10 cm (Fig. 2-9). The snow was sampled by walking back and forth in a lawnmower pattern, with higher sampling clusters around deformed ice. A typical survey over the 100 m x 100 m area had ∼2000 points, with higher resolution (∼10 cm) near areas of deformed ice and lower resolution (∼5 m) over flat, level topography (Fig. 2-9). These measurements were converted into a surface by using natural neighbor interpolation (Sibson, 1981), binned at 20 cm to match the lidar and AUV data.

2.3.2 Lead detection and referencing

The OIB lidar data consists of ellipsoid-referenced surface elevations, which need to first be converted into local lead-referenced snow surface freeboards. To do this, we first subtract the geoid (EGM-2008 at 1 minute resolution), and then we use local lead elevations as references. Identifying leads in the lidar altimetry is difficult, with various algorithms proposed. As presented in Kurtz (2013), Onana et al. (2013) used coincident camera imagery to look for leads directly, with extra filters to remove clouds and shadows. Kwok et al. (2012) used the raw reflectivity data to isolate leads, which have lower reflectivity than ice. Wang et al. (2013) found that in the

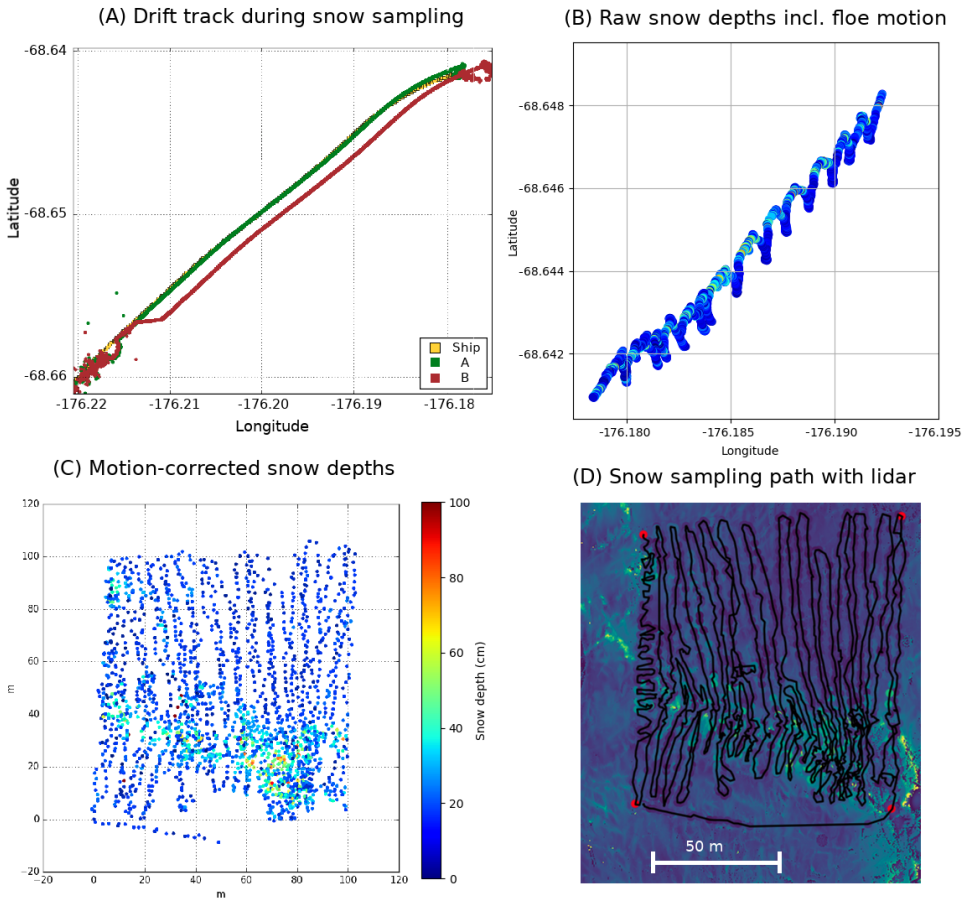

Figure 2-9: An example of the sea ice floe motion correction, as applied to PIP9 (2017/06/02). (A) The motion of two beacons, as well as the ship GPS data, for the duration of the ice station.(B) The raw snow depths without correction for floe motion, showing considerable drift - the total drift during the snow sampling in (B) is 2.7 km, which is far larger than the survey size (100 m), which is why the floe motion needs to be corrected. (C) The sampling frequency is higher around the deformed surface, where snow depths vary over small scales, and is more sparse around level areas that have smaller snow depth variations. (D) The characteristic back-and-forth path, with minor deviations to account for local topography.

absence of lead detections, using the lowest 0.2% percentile elevation values of the lidar freeboard gives a reasonable approximation to the lead elevation, when taking large (30 km) segments. These lead-referenced surface elevations, hence called 'snow freeboard' measurements, can then be combined with measured or assumed snow depths to estimate sea ice thickness. A more detailed discussion of lead-detection errors is in Appendix B.

For our preprocessing, leads are identified using the Digital Mapping System onboard the OIB flights that takes synchronous digital imagery while the lidar/radar are operating (Dominguez, 2010). The method is similar to the SILDAMS algorithm presented by Onana et al. (2013). We use just the grayscale imagery, and identify the peaks corresponding to open water, gray ice, and white ice, and then find the corresponding lidar points in each area (noting that some images may not have all three of these). An example is shown in Fig. 2-10. We also exclude shadows and clouds from the lead detection, and manually verify the detected leads. A list of lead elevations and their longitude/latitude are compiled, and then the geoid-corrected ATM data is referenced to the local sea surface height. The local sea surface height is determined by an inverse-distance weighted elevation for all the leads within ± 5 km. This distance is chosen in order to be within the first baroclinic Rossby radius of deformation ($\sim$10 km at polar latitudes) for a given point, which is the typical length scale of eddies that create nonlinear perturbations on the sea surface height (Chelton et al., 1998). If there are not at least two leads identified within a $\pm$ 5 km, the lidar point is discarded. The lidar is interpolated onto a grid with 1 m spacing using natural neighbor (Sibson, 1981), and then windowed into 180 m x 180 m windows for later use in deep learning (Fig. 2-11). Lidar points over water are masked and discarded from analysis, as there are relatively few returns which can introduce artifacts into the interpolation. These were excluded by identifying windows that had lead-referenced snow freeboards at the 3rd percentile of 0.0 m or lower. Our lead accuracy is typically better than 3 cm, which is comparable to other studies like Kwok and Kacimi (2018).

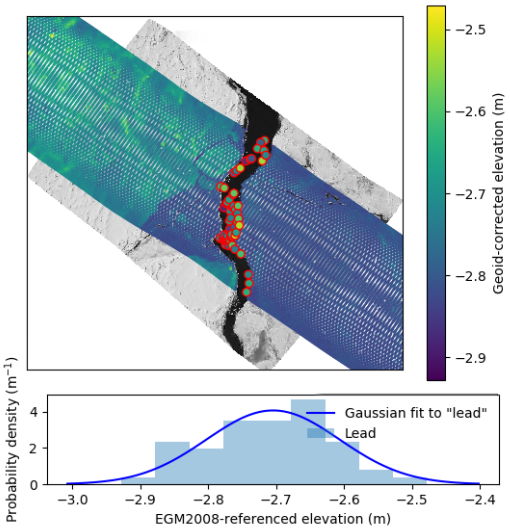

Figure 2-10: An example of the lead-finding algorithm. The DMS camera imagery is shown with the lidar points overlain. The lidar points that are within the lead are circled in red, and their distribution shown in the bottom panel. The elevations have been geoid-corrected from the OIB L1B-ATM data. Note the distribution of lead elevations is approximately Gaussian (best Gaussian fit shown in blue).

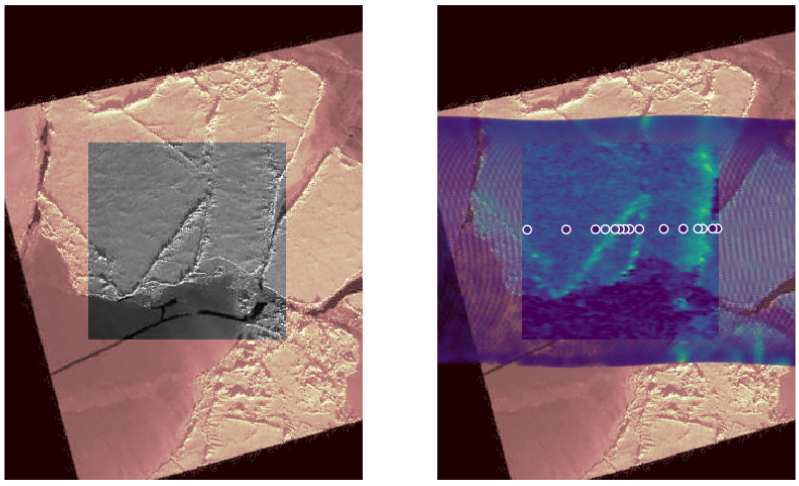

Figure 2-11: An example lidar window, made by taking the lead-referenced lidar data, selecting a 180 x 180 m window (highlighted) and then interpolating using natural neighbor interpolation at 1 m resolution. Overlain is the collinear (1-D) snow depth measurements (white-edged circles). The conical scanning pattern of the laser altimeter is also clearly shown here.

Chapter 3

High-resolution SIT predictions

This work was published in *The Cryosphere* (Mei et al., 2019) and the content is reproduced here, with some formatting edits and some additional analysis in Sect. 3.4.2.

Abstract

Satellites have documented variability in sea ice areal extent for decades, but there are significant challenges in obtaining analogous measurements for sea ice thickness data in the Antarctic, primarily due to difficulties in estimating snow cover on sea ice. Sea ice thickness (SIT) can be estimated from snow freeboard measurements, such as those from airborne/satellite lidar, by assuming some snow depth distribution or empirically fitting with limited data from drilled transects from various field studies. Current estimates for large-scale Antarctic SIT have errors as high as ∼50%, and simple statistical models of small-scale mean thickness have similarly high errors. Averaging measurements over hundreds of meters can improve the model fits to existing data, though these results do not necessarily generalize to other floes. At present, we do not have algorithms that accurately estimate SIT at high resolutions. We use a convolutional neural network with laser altimetry profiles of sea ice surfaces at 0.2 m resolution to show that it is possible to estimate SIT at 20 m resolution with better accuracy and generalization than current methods (mean relative errors ∼15%).

Moreover, the neural network does not require specifying snow depth/density, which increases its potential applications to other lidar datasets. The neural network may be accounting for the variable snow depth and variable snow and ice densities as a common layer with varying density. The learned features appear to correspond to basic morphological features, and these features appear to be common to other floes with the same climatology. This suggests that there is a relationship between the surface morphology and the ice thickness. The model has a mean relative error of 20% when applied to a new floe from the region and season. This method may be extended to lower-resolution, larger-footprint data such as such as Operation IceBridge, and suggests a possible avenue to reduce errors in satellite estimates of Antarctic SIT from ICESat-2 over current methods, especially at smaller scales.

3.1 Objectives

As first stated in Sect. 1.1 and Sect. 1.3, it is much easier to measure sea ice from above than from below. However, it is hard to convert measurements of the above-water snow surface to estimates of sea ice thickness, due to the varying densities of ice and snow and the varying proportion of snow and ice in the surface snow freeboard. Because the interannual variations in sea ice thickness are not known with much confidence, it is important to figure out how to estimate sea ice thickness from surface elevation measurements.

Following from Sect. 1.4, the goals here are to identify whether surface morphology features, as measured via surface laser altimetry, can be used to estimate sea ice thickness. The data for this comes from PIPERS (Sect. 2.1), using the 'layer-cakes' with collocated snow depth, snow surface and sea ice draft. We compare the accuracy of deep learning predictions of SIT as opposed to traditional linear regressions that are used in literature (Sect. 3.3). Then, in Sect. 3.4 we discuss possible reasons why the linear fits may not generalize well from floe-to-floe, as well as why the ConvNet seems to generalize well.

3.2 Methods

3.2.1 Linear regression approach

We attempt to statistically model SIT using surface-measurable metrics (e.g. mean and standard deviation of the snow freeboard), in order to see the limitations of this method. To accurately calculate SIT without making assumptions of snow distribution, we need to use combined measurements of ice draft (AUV), snow freeboard (lidar) and snow depth (probe). Here, we primarily use PIPERS data to focus on early-winter Ross Sea floes, and also because this is the largest such dataset from one season/region, which is important so that the ridges have consistent morphology.

We use simple (multi)linear least-squares regression with either one (snow freeboard, F) or two (F and snow depth, D) variables with a constant term, such that $T = c_1F + c_2D + c_0$.

For the two-variable fit, we do an additional fit with the constant forced to be zero, in order to obtain coefficients that can be used, following Eq. 1.1, to estimate the snow/ice densities.

To measure the fit accuracy, we use the mean relative error (MRE), as this avoids weighting errors from thin/thick ice differently. The R^2_{adj} value, adjusted for different number of variables, is also reported where possible (it is not defined for a fit forced through the origin). When comparing the generalization of the fits to test data excluded from the fit data, we also report the relative error of predicting the mean survey-wide thickness (REM), as often researchers are interested in the aggregate statistics of a survey. These fit errors in estimating mean SIT are compared to both prior relationships derived from drilling data to highlight uncertainty when used with different ice conditions, and to our ConvNet predictions of ice thickness.

In order to motivate more complex methods in subsequent sections, we also use surface roughness (standard deviation, σ) to predict thickness, to demonstrate that surface morphological characteristics have some information that can be used to predict thickness.

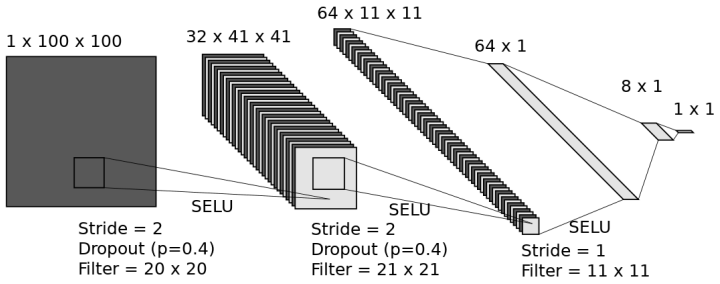

Figure 3-1: ConvNet architecture, using 3 convolutional layers and 2 fully-connected layers, for predicting the mean thickness (1 x 1 output) of a 20 m x 20 m (100 x 100 input) lidar scan window at 0.2 m resolution (LeNail, 2019). The (64 x 1) layer is made by reshaping the (64 x 1 x 1) output of the final convolutional layer, and so is visually combined into one layer.

3.2.2 Deep learning approach

One advantage of deep learning techniques is that they are able to learn complex relationships between the input variables and a desired output, even if the relationships are not obvious to a human. Although they are commonly used for image classification purposes, they can also be used for regression (e.g. Li and Chan, 2014). We expect a convolutional neural network (ConvNet) to achieve lower errors in estimating SIT, as they are able to learn complex structural metrics, in addition to simplistic roughness metrics like σ. Our input is a windowed lidar scan (snow freeboard) and an output of mean ice thickness. Notably, there is no input of snow depth, nor any input of ice/snow densities. This allows the ConvNet to infer these parameters by itself, and more importantly, to potentially use different density values for different areas.

Our architecture is shown in Fig. 3-1. The input consists of 20 x 20 m (100 x 100 pixel) windows, with 3 convolutional layers, with a stride of 2 in the first 2 layers, and two fully-connected layers. We used scaled exponential linear units (SELU) to create non-linearity (Klambauer et al., 2017). The loss function used was mean squared error. The optimizer used was Adam with weight decay 1.0×10^{-5} (Kingma and Ba, 2014). The initial learning rate was $\eta = 3 \times 10^{-3}$ and reduced by a factor of 0.3 every 100 epochs until it reached 9×10^{-5}. We also used dropout (p=0.4) and augmentation

(random 90° rotations, horizontal/vertical flipping) to reduce overfitting (Srivastava et al., 2014). The output is the mean thickness, scaled by 5. The scaling here is because, for our dataset, the maximum thickness was just under 5.0 m, and normalizing the outputs to be between $0-1$ allows the gradients for the backpropagation of error to neither vanish nor blow up.

The training/validation set consisted of randomly-selected windows from three PIPERS ice stations, each on a different floe. We chose 20 m as the window size by using the range of the semivariogram for the floes (25 m), which we expect to represent the maximum feature length scale. This compares well to an average snow feature size of 23.3 m from early-winter Ross Sea drill lines from Sturm et al. (1998). We chose 20 m instead of 25 m windows to balance this with the need for a smaller window size to ensure a larger number of windows (= data inputs) for our analysis. These data were randomly divided into 80%-20% to make the training and validation sets. The remaining floe (divided into windows) was kept as a test set, in case the training and validation windows had similar morphology and the validation set was thus not entirely independent of the training set. To prevent cherry-picking, the ConvNet was trained four times, with a different floe used as the test floe each time. Results are shown in Table 3.2. Although the training error is directly analogous to the fit error for linear models for some dataset, it is much easier to overfit with a ConvNet as the training error can be made arbitrarily low. As a result, we compare our validation error to the linear fit errors, and also use our test errors as a test of the generalization of our model. For the rest of this chapter, analysis of the ConvNet refers to the one using PIP8 as a test set, though using a different one would yield qualitatively similar analysis.

3.3 Results

3.3.1 Linear model results

Fitting to snow freeboard only

Although we have snow depth measurements in addition to snow freeboard measurements, in general there are far fewer snow data and so we first try to fit with just snow freeboard, by making some snow depth assumptions. This approach has been applied by Özsoy-Çiçek et al. (2013) and Xie et al. (2013) in order to obtain empirical relationships between SIT and snow freeboard. All our fitted coefficients are shown in Table 3.1. Because the R^2 is not well-defined for a fit with no constant term, we can compare all the model fits with the AIC (Akaike Information Criterion (lower is better, see Akaike (1974)). For all categories except for 'Level', the $\{F, D$, constant$\}$ fit is indisputably best; for example, a difference in AIC of 70 between the two best models in the 'All' category implies that the likelihood that the model with $\{F$, constant$\}$ is better than the one with $\{F, D$, constant$\}$ is $e^{-70} = 4 \times 10^{-31}$. For the 'Level' category, the difference in AIC suggests that linear fits with $\{F, D$, constant$\}$ and $\{F$, constant$\}$ are very similar (the latter has a 50% likelihood of being better than the former), which is consistent with the idea that level ice probably has a constant ice/snow ratio such that introducing D as a variable does not improve much on using only F.

Fitting $T = c_1 F + c_0$ gives a mean relative error (MRE) of 23%. However, the slope is much higher (7.7), and the intercept is also larger and different in sign (-0.7 m) to existing fits in the literature (e.g. Özsoy-Çiçek et al. (2013) found that $T = 2.45F + 0.21$ for a early-spring Ross Sea dataset). Using the fitted relationship from Özsoy-Çiçek et al. (2013) on our dataset, the MRE is 36%, and the relative error in estimating the overall survey mean thickness (REM) is 41%. This is perhaps partly due to the seasonal difference in these datasets, which itself implies that the proportion of deformed ice (and hence nonzero ice freeboard) is variable. Reasons for the difference in slope and intercept are given in Section 3.4.1.

Table 3.1: Fitted coefficients for SIT T as a multilinear regression of the snow freeboard F and snow depth D (Section 3.2.2), and also fitting for F only (Section 3.2.1). The variable 'const.' refers to a constant term being included in the fit. Surfaces are also categorized (Fig. 3-3) to incorporate roughness into the fits (Section 3.3.1). As the R^2 is not well-defined for a fit with no constant term, the Akaike Information Criterion (a metric that minimizes information loss) is used to compare the models (Akaike, 1974). The R^2 is reported for the with-constant fits only and is adjusted for the different sample sizes in each fit. For each dataset, the smallest AIC value is **bolded**, and the second-lowest underlined. The absolute value of the AIC does not matter; only the relative differences between AICs for different models that use the same dataset matter, with the lowest being the best model. For individual floe fits, only PIP8 is shown for brevity as the other floes have comparable errors/coefficients.

	Fitted variables	R^2_{adj}	AIC	MRE, m [%]	F coeff.	D coeff.	Constant (m)
PIP8	F, const.	0.91	10.2	0.20 [16]	7.07 ± 0.30	N/A	-0.81 ± 0.10
	F, D	N/A	37.3	0.26 [24]	9.03 ± 1.0	-5.45 ± 1.25	N/A
	F, D, const.	0.92	**5.30**	0.18 [15]	8.85 ± 0.73	-2.70 ± 1.02	-0.70 ± 0.11
Ridged	F, const.	0.91	128	0.31 [21]	7.59 ± 0.20	N/A	-0.65 ± 0.08
	F, D	N/A	111	0.29 [22]	10.33 ± 0.44	-6.53 ± 0.67	N/A
	F, D, const.	0.94	**75.5**	0.25 [17]	10.42 ± 0.39	-5.06 ± 0.63	-0.45 ± 0.07
Level	F, const.	0.00	-71.6	0.07 [13]	0.02 ± 0.67	N/A	0.50 ± 0.11
	F, D	N/A	-56.5	0.07 [13]	3.58 ± 0.77	-0.82 ± 0.96	N/A
	F, D, const.	0.07	**-72.3**	0.06 [12]	0.87 ± 0.85	-1.22 ± 0.76	0.52 ± 0.11
Snowy	F , const.	0.81	32.3	0.27 [24]	7.74 ± 0.59	N/A	-0.72 ± 0.16
	F, D	N/A	36.4	0.29 [34]	10.45 ± 1.37	-6.29 ± 1.63	N/A
	F, D, const.	0.87	**19.9**	0.22 [23]	11.88 ± 1.15	-5.33 ± 1.33	-0.63 ± 0.14
All	F, const.	0.92	179	0.28 [23]	7.67 ± 0.15	N/A	-0.73 ± 0.05
	F, D	N/A	194	0.30 [31]	10.42 ± 0.37	-6.81 ± 0.53	N/A
	F, D, const.	0.94	**109**	0.24 [20]	10.19 ± 0.31	-4.51 ± 0.49	-0.52 ± 0.05

We also test how well-generalized the fits are by fitting only 3 of our 4 surveys at a time, then testing the fitted coefficients on the remaining survey. These results are summarized in Table 3.2. The average fit error was 24%, but the average test error was 31%, which means that empirical fits to the snow freeboard may have errors of 31% when applied to new datasets.

Table 3.2: A compilation of the MRE of different fitting methods. Coefficients for the linear fits are shown in Table 3.1 and details are in Sections 3.2.1-2. The leftmost column indicates the floe that was excluded from the fitting data (e.g. the first row indicates fits that were done over the PIP7-9 data and then tested on PIP4). The ConvNet validation error was used for comparison with the linear model fits, as the training error can be made artificially low by overfitting. On average, the ConvNet achieves the best generalization in the fit, even though there are individual anomalous cases. For example, the F-only fit using PIP7 as a test set has a low test error than fit error, which simply means that the average snow/ice ratio for PIP7 is similar to the averaged snow/ice ratio for the other floes. The F only fit is most comparable to our ConvNet as neither use the snow depth as an input.

	Linear (no constant)		Linear (with constant)		F only (with constant)		ConvNet	
Test set	Fit MRE	Test MRE	Fit MRE	Test MRE	Fit MRE	Test MRE	Val. MRE	Test MRE
PIP4	36%	12%	17%	31%	19%	39%	14%	20%
PIP7	25%	33%	20%	24%	26%	23%	14%	18%
PIP8	33%	32%	22%	23%	25%	32%	16%	20%
PIP9	27%	59%	20%	34%	24%	30%	14%	20%
Average	**30%**	**34%**	**20%**	**28%**	**24%**	**31%**	**15%**	**20%**

Fitting to snow freeboard and snow depth

For this section, we do two different regressions: one with a constant, and one without. The with-constant fit is intended to test whether introducing additional information improves the empirical fits, following Özsoy-Çiçek et al. (2013), and the without-constant fit is intended to be compared against Eq. 1.1 to estimate sea ice/snow densities. The coefficients are reported in Table 3.1 and the fit/test MREs are reported in Table 3.2. We can see that adding snow depth as a variable only slightly improves the fit MRE (average 20%), but the fits remain poorly-generalized, with a test MRE of 28%, only slightly lower than the 31% test MRE of fitting with F only.

Fitting without a constant allows us to directly compare the fitted coefficients with Eq. 1.1. Using typical values of 910 kg m^{-3} for ice density, 1027 kg m^{-3} for water

density and 323 kg m^{-3} for snow density from Worby et al. (2011), the coefficients for the freeboard F and snow depth D should be 8.8 and 6.0. Similarly, Zwally et al. (2008) used corresponding densities of 915.1 kg m^{-3}, 1023.9 kg m^{-3} and 300 kg m^{-3}, giving a freeboard coefficient of 9.4 and a snow coefficient of 6.7. Our results when fitting over all 4 floes are 10.4 for c_1 and 6.8 for c_2, which are comparable to those inferred from Zwally et al. (2008), although there is considerable variation between the floes (7.9-10.6 for c_1; 3.9-6.3 for c_2, not shown in Table 3.1).

Assuming a density of seawater during PIPERS of 1028 kg m^{-3} (determined from surface salinity measurements at these stations), this gives bounds for the effective densities and standard errors of sea ice and snow as 929.4$\pm$3.5 kg m^{-3} and 356.3$\pm$57.2 kg m^{-3}. The snow density is in line with Sturm et al. (1998), who found mean densities of 350 and 380 kg m^{-3} during autumn/early-winter and winter/spring, respectively, in the Ross Sea, as well as the measured snow densities from PIPERS (245-300 kg m^{-3}). The measured PIPERS snow densities may be biased low because they were measured at level areas, and possibly do not represent snow densities in drifts around ridges well. The errors here are propagated from the standard errors found during the regression; they are therefore representative of the error in estimation of the mean densities over all data and do not represent actual ranges in the ice/snow densities. The ice (effective) density estimates here are averaged over the entire PIPERS dataset (including both deformed and undeformed ice) and thus may not apply to other samples from the Ross Sea in winter, as the effective density is affected by the proportion of ridged ice, which is deliberately overrepresented in our sample. Moreover, it is important to note that under this fitting method, the density estimates are coupled (due to ρ_i appearing in both coefficients in Eq. 1.1) and if the estimate of ρ_s decreases, ρ_i increases. For example, if $\rho_i = 935$ kg m^{-3} (unusually, but not impossibly high for the effective density of ridged ice, which includes some proportion of seawater - see Timco and Frederking (1996)), the best estimate for ρ_s becomes 312 kg m^{-3}, which is closer to the measured 300 kg m^{-3} value from PIPERS.

The fact that introducing snow depth as a variable only slightly improves the generalization of the fit may be because snow depth is itself highly correlated with

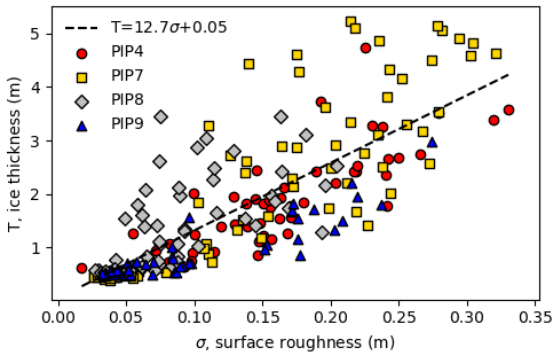

Figure 3-2: Predicting mean ice thickness with just the surface roughness (σ) as the input, with MRE 33%. The best-fit line is also shown, with R^2=0.65.

snow freeboard (e.g. Özsoy-Çiçek et al., 2013). Linear methods of fitting require assuming a constant snow/ice density (or in a one-layer case, a constant 'effective density'), which implies an irreducible error for estimating small-scale SIT. This fails to account for varying ice/snow densities around level/deformed ice. This is discussed further in Section 3.4.1, and motivates the introduction of surface roughness (σ) as an additional variable in our linear fit.

Incorporating surface roughness into the fit

Given that we expect effective density variations for different surface types, we expect SIT estimates to improve with the addition of surface morphology information. The most simple of these is the surface standard deviation, as prior studies have found that this is correlated to the snow depth and the mean thickness(Kwok and Maksym, 2014; Tin and Jeffries, 2001b). Our data also show a reasonable relationship between SIT and surface σ, though it is weaker than fits to the freeboard (Fig. 3-2). Adding the roughness as a third variable to the fit gives an average fit MRE of 18% and an average test MRE of 24%. This is not much of an improvement, and it is possible that σ is too simplistic a metric to improve the fit, or that it is itself highly correlated with F and therefore offers little additional information.

There is no particular reason to expect the surface σ to be linearly combined with the snow depth and snow freeboard, even if it makes dimensional sense. Instead, we

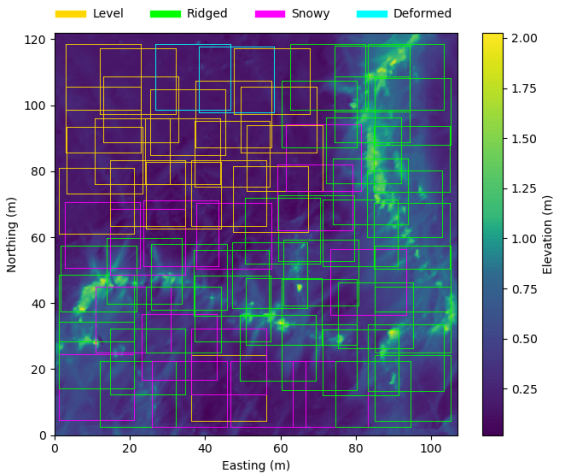

Figure 3-3: An example lidar scan from a station (PIP7) with the manually classified segments. Snow features are clearly visible emanating from the L-shaped deformation. Deformed (blue) surfaces were excluded from the analysis.

can try using the roughness as a regime selector. To do this, the lidar windows were classified manually into snowy surface, level surface, ridged surface and deformed surface categories (Fig. 3-3). If it had both a ridge and snow, it was classified as ridged. 'Level' surfaces were distinguished as those windows with no visible snow/ice features in the majority of the window. 'Snowy' surfaces were those that contained a snow feature (e.g. a dune or drift) in the window. 'Deformed' was intended as a transitional category for images that had no clear ridge but were generally rough - this comprised, typically, $\sim$5% of an image and was excluded from analysis. We acknowledge that this classification can be arbitrary, and use this method only to show that different surface types should be treated differently, but a manual classification does not help much: this motivates the use of a deep neural network in the next section. The snowy, level and ridged categories were individually fitted to see if there were any differences in the coefficients; these are also reported in Table 3.1.

We then used a two-regime model over all four floes, so that ice thicknesses for the low-roughness surfaces are estimated using the 'level' coefficients, and high-roughness surfaces using the 'ridged' coefficients. This resulted in MREs of 16-21% assuming 20-50% of the surface is deformed. This is slightly better than for fitting the 'all' category in Table 3.1 (20% MRE), suggesting that distinguishing topographic regimes improves

thickness estimates. However, this fit has issues with generalizing to other floes. If the fit for the rough/level coefficients is done using only 3 floes and then applied to the remaining (test) floe (using a surface roughness threshold determined from that floe, and again assuming 20-50% of the surface is deformed), the test MREs averaged over all possible choices of test floe are considerably higher (24% when fitting). This does not improve much on the generalization from the two-variable linear fit, where the test MRE was 28%.

3.3.2 ConvNet results

The (irreducibly) poor generalization of linear fits, likely due a locally-varying proportion of snow/ice amongst different surface types, motivates the use of more complex algorithms that can account for the surface structure. For this, we use a ConvNet with training/validation/test datasets as described in Section 3.2.2.

We tried networks with 2, 3 and 4 convolutional layers and 1 or 2 fully connected layers with a variety of filter sizes and found the one shown in Fig. 3-1, with a total of 5 hidden layers, had the best results. The filter sizes were chosen to try and capture feature sizes of <20 m, as discussed in Section 3.2.2. The first layer has a size of 4 m, the second is 8.4 m, and the third is 8.8 m (corresponding to windows of 20, 21 and 11 pixels at 0.2, 0.4 and 0.8 m resolution). For the first two layers, a stride of 2 was used to reduce the dimensionality of the data. The implementation was done using PyTorch with an NVIDIA Quadro K620 GPU and took around 10 minutes.

The input windows were randomly flipped and rotated in integer multiples of 90° to help improve model generalization. Dropout, which randomly deactivates certain weights with some probability p, were added after the first and second convolutional layers ($p = 0.4$) to reduce overfitting (Srivastava et al., 2014).

The selected model for analysis was the best-performing validation error (15.5%) at epoch 881 (Fig. 3-4).

The lowest validation error was 15%, corresponding to a training error of 11% (Fig. 3-4 and Fig. 3-5a and b). The mean test error (on the excluded floe) was 20%. Although the linear models have a similar fit error, they do not generalize as well to

the test set, and the resulting thickness distribution is visibly different to the real test distribution (Fig. 3-5c).

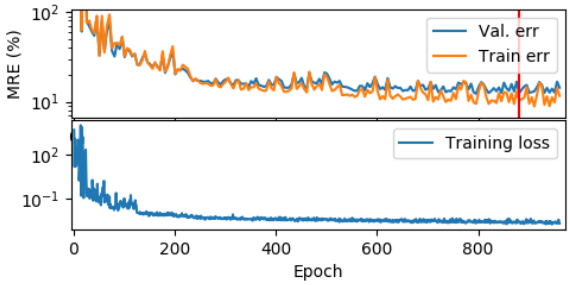

Figure 3-4: The training and validation MREs for our trained ConvNet (top) and the training loss (bottom), showing that the ConvNet has converged without overfitting.

This shows better generalization than the linear models (test MREs from 28-47%). Although the best-performing linear models have only slightly higher test MREs (24% for the 3-variable fit in Section 3.3.1) than our ConvNet (20%), the range of errors is much greater, with test MREs of 18-29%, whereas the ConvNet has remarkably consistent test MREs of 18-20%. Furthermore, it is important to remember that achieving these comparably low MREs with linear models requires snow depth as a variable, which is generally not available. These fits also typically include a negative constant (Table 3.1), which means $T < 0$ for $F = D = 0$ which is clearly unphysical and limits the application of these models to areas of low snow freeboard. The fits to snow freeboard only, which is using the same input data as the ConvNet, have considerably higher test MREs (23-39%, see Table 3.2). For sake of comparison to models that use RMS error such as Özsoy-Çiçek et al. (2013), the validation RMS error for our survey-averaged mean thickness values is 2 cm, which is lower than the RMS error of 11-15 cm from Özsoy-Çiçek et al. (2013). Our fit uses 3 surveys from 3 different floes as an input, which means the fit is likely lower in error than Özsoy-Çiçek et al. (2013), which uses 23 floes. However, we would also expect poorer generalization for our test set from using only 3 surveys. Although our test RMS error for the mean survey thickness (3 cm) cannot be directly compared, it is reasonable to surmise that our ConvNet achieves better generalization than a linear fit. Note that the RMS error is not linked to the surface RMS roughness, which is just the standard

deviation of the snow freeboard.

As shown in Fig. 3-5c, the ConvNet does seem to be capturing the thickness distribution of the test floe, even if the individual window mean estimates have some scatter. In contrast, the linear models have considerably different thickness distributions (Fig. 3-5, red points/lines) despite having similar fit MREs (Table 3.2). The ConvNet also successfully reproduces the spatial variability of the SIT distribution better than the linear fit (Fig. 3-6). Note, because of the small size of the dataset, there is significant oversampling in the ConvNet prediction of the floe SIT distribution. The primary difference between the ConvNet and linear fit for this floe is a large overestimation of level ice thickness. This demonstrates the inability of the linear fit to account for variations of effective densities and/or snow/ice freeboard ratios. The ConvNet prediction can have some large local errors. In this case chiefly on the flanks of the ridge, where steep freeboard or thickness gradients may affect performance. Comparisons for other floes (not shown) are qualitatively similar, though the spatial distribution of fit errors varies among floes. The key result of the ConvNet is in the significantly reduced error in the local (20 m scale) mean thickness (MRE of 15-20%), which also gives a low, $\sim$10% error of the average scan-wide thickness. Moreover, this high accuracy also carries over to test sets from the same region/season. In contrast, linear models, which do not generalize well to new datasets, have a considerable bias (Fig. 3-6), despite having an ostensibly good fit. Analysis of why the ConvNet may be performing better than linear fits is given in Section 3.4.2.

3.4 Discussion

3.4.1 Possible causes for poor linear fit

Our linear regression results for fitting $T = c_1F + c_0$ have markedly different coefficients from drill line data from the same region/season (Özsoy-Çiçek et al., 2013). Here we discuss possible reasons for their differences. The first difference is that our value for $c_1 = 7.67$ (Table 3.1) is much higher. This is almost certainly because our

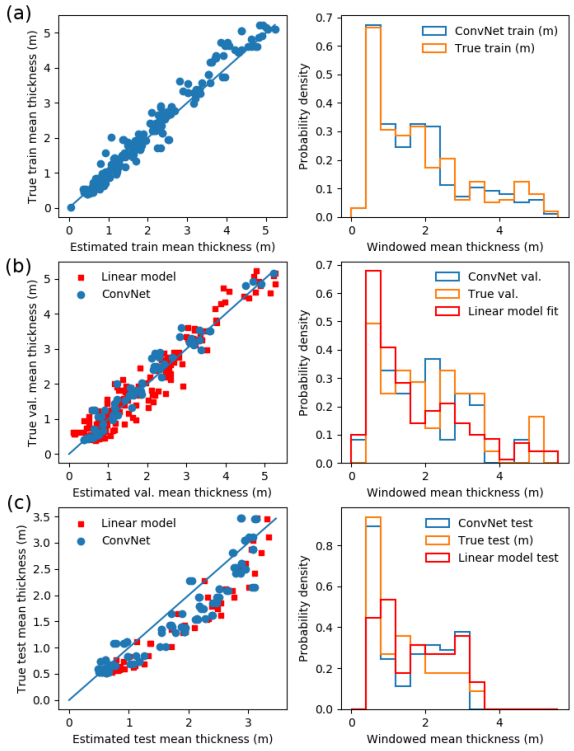

Figure 3-5: ConvNet results, with (a) the learned ConvNet model applied to the training data (80% of randomly sampled 20m x 20m windows from PIP4, PIP7, PIP9), with MRE 12%, (b) the learned ConvNet model applied to the validation data (remaining 20% of the randomly sampled 20m x 20m windows from PIP4, PIP7, PIP9) with MRE 16% as well as a linear model (with snow freeboard + constant) fitted to PIP4, PIP7, PIP9 with MRE 25%, (c) the learned ConvNet model and fitted linear model applied to randomly sampled 20m x 20m windows from PIP8, as a check against learning self-similarity, with MRE 20% (ConvNet) and 32% (linear model)). In each case, the left panel shows a scatter plot with the predicted and true thicknesses, and the right panel shows the resulting thickness distribution. Our results suggest slight overfitting, as the test error is higher than the training error, but the learned model still generalizes fairly well, with MREs much lower than linear models, even when including an unphysical intercept to improve the fit (Table 3.2).

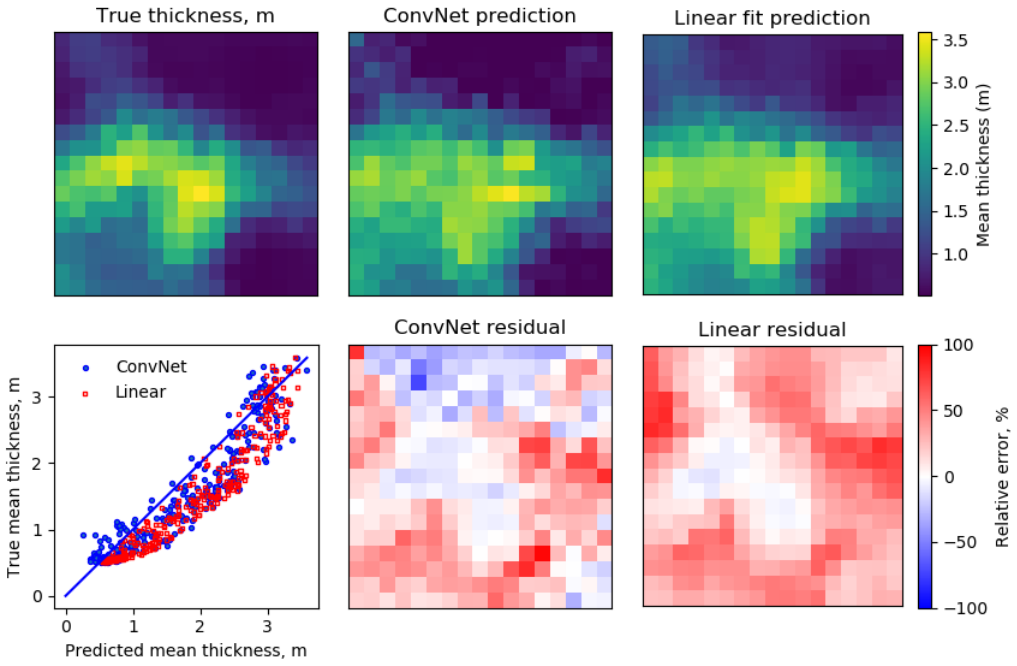

Figure 3-6: Ice thickness profile of the test set (PIP8), using the linear fit ($T = c_1F + c_0$) and ConvNet model, both done with PIP4, 7 and 9 as inputs. The input windows are 20 m x 20 m, with a stride of 5 m in each direction, so there is a considerable oversampling. The mean residual for the linear model (35 cm) is much higher than for the ConvNet (19 cm), which means the resulting mean thickness has almost twice the REM (24% vs. 13%). The scatterplot clearly shows the linear model (using 20m windows as well, with coefficients from Table 3.1) predictions are consistently biased high, which is also apparent in the linear model residual.

dataset includes much more deformed ice, as we deliberately sampled deformed areas on floes. This may be because our dataset includes much more deformed ice, as we deliberately sampled deformed areas on floes. At one extreme, where the snow load is large such that the snow depth = snow freeboard assumption is approximately valid (set $F = D$ in Equation 1.1), which for our data occurs for level thin ice where there is some snow load, Eq. 1.1 would simplify to $T = 2.7F$ (using density values from Zwally et al. (2008)). In contrast, when the topography is sufficiently rough, there is considerable ice freeboard, which may even exceed snow depth. If we assume the snow is negligible ($D = 0$), which may be the case at the sail peak, Eq. 1.1 becomes $T = 9.4F$. These values become lower and upper bounds for fitting c_1 in $T = c_1F$ (without the constant c_0). The best fit value for c_1 is 5.8 when fitting to the full dataset (Fig. 3-7), which falls between these two extremes of snow-only F and ice-only freeboard F. Our coefficient is also comparable to Goebell (2011), who found a coefficient of 5.23 from first-year Weddell ice. Much as in Goebell (2011), our dataset includes considerable deformed ice which has a non-zero ice freeboard, and so the coefficient of F is higher than 2.7. We can estimate the ratio of snow to ice by comparing this with the hydrostatic equation: for example, if we assume typical snow/ice densities of 300 kg m^{-3}/920 kg m^{-3}, this implies that snow, on average, comprises 54% of the measured snow freeboard. Using these values, Eq. 1.1 simplifies to $T = 5.8F$, as in Fig. 3-7. In further support of this, our dataset has mean snow depths for the four surveys ranging from 16-26 cm, and mean snow freeboards ranging from 24-37 cm, implying considerable non-zero mean ice freeboards.

The high scatter of our fit also suggests that the snow/ice ratio is varying locally, as can be expected around level/deformed ice. If the proportion of ice to snow were constant, then the best-fit line, for whatever slope, would have no scatter. This is not the case in Fig. 3-7, and indeed the standard deviation of ice freeboard across all windows was 7.9 cm (mean: 9.0 cm). This means that assuming a constant snow/ice density or a constant snow/ice proportion is not justified, and hence it is likely that simple statistical models break down when looking at deformation on a small scale, or when large-scale snow deposition and ice development conditions vary. This mirrors

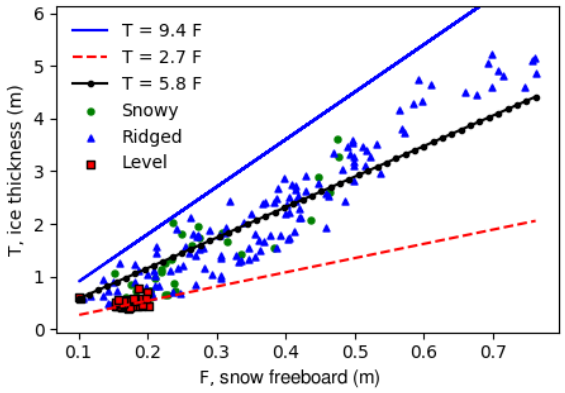

Figure 3-7: The SIT (T) as a function of measured snow freeboard (F). As expected, all points lie between the two extreme regimes (no ice freeboard and no snow freeboard). The level surfaces mostly have no ice freeboard, as expected, though there is some scatter that suggests a varying component of ice freeboard. The best fit line for all windows from Table 3.1 is shown in black. Assuming mean snow and ice densities of 300 and 920 kg m^{-3}, this implies a mean proportion of 55% snow and 45% ice in the snow freeboard. Again, the scatter around the best fit line indicates that this proportion is changing. Some points for the level category fall below the $T = 2.7F$ line, suggesting that snow densities in these areas are <300 kg m^{-3} (or effective ice density <915 kg m^{-3}.)

the conclusions in Kern et al. (2016), who found that linear regressions could not capture locally- and regionally-varying snow/ice proportions. Even when including regime-dependent fits (Sect. 3.3.1, Fig. 3-2), this does not improve the test errors because this is likely too simplistic (even within a ridge, the ratio of snow/ice is likely varying). An important point regarding σ is that it does not actually account for the surface morphology very well, as any permutation of elevations within the window will give the same σ. This means that the 'shape', or 'structure' of the surface is not truly accounted for. This motivates more complex metrics for surface roughness (Section 3.3.2).

Unlike our approach, the fits in Özsoy-Çiçek et al. (2013) and Xie et al. (2011) use large-scale, survey-averaged data. Their coefficients for c_1, 2.4-3.5 and 2.8 for Ross Sea and Bellingshausen Sea data respectively, are near the theoretical value of 2.7 assuming no ice freeboard. This suggests that at large scales for some seasons/regions, it may be reasonable to assume that the mean ice freeboard is zero, but this is not the case at smaller scales. It is also possible that drill lines have undersampled ridged

ice due to sampling constraints, or (in our case) sample heavily deformed areas that are not typically sampled in situ. Thus, empirical fits should be used with caution.

The second major difference is that our intercept is negative, whereas those from Özsoy-Çiçek et al. (2013) and Xie et al. (2011) are all positive. In our case, it is possible to interpret our negative intercept as a result of fitting a linear model across two roughness regimes. From above, the two regime extremes (no-ice vs. no-snow contribution to snow freeboard) give $T = 2.7F$ and $T = 9.4F$ as limiting cases. In general, we expect the proportion of ice freeboard to gradually increase as F increases from thinner, level ice to thicker, deformed ice. Although snow also accumulates around deformed ice, there may also be local windows at parts of the ridge with no snow (e.g. the sail). This means that we expect a gradual transition from $T = 2.7F$ to $T = 9.4F$ as F increases. Fitting one line through these two clusters of points would result in a coefficient for F between 2.7 and 9.4 and a negative intercept, which we find in almost all our cases. The one exception is the fit for the level category, which is essentially a null fit (as over 90% of the thickness values are clustered around 0.5 $\pm$ 0.05 m). In contrast, the coefficients for F from Özsoy-Çiçek et al. (2013); Xie et al. (2011) are all $\sim$3, because these studies average over multiple floes and have a sufficiently small proportion of deformed surface area to assume a negligible ice freeboard as discussed above. In their case, their intercept would be positive, as their ice thickness estimates would be otherwise underestimated due to some of the snow freeboard being ice instead of snow.

When fitting a linear/ConvNet model to snow freeboard data, we cannot know whether there are negative ice freeboards; as such, these methods account for it only implicitly, with a linear fit effectively assuming that a similar percentage of freeboards will be negative. This may contribute to errors when trying to apply a specific linear fit to a new dataset. A ConvNet could conceivably do better here, in that significant negative freeboard is likely to matter most when there is deep snow, which might have recognizable surface morphology, although this is quite speculative.

3.4.2 Plausible physical sources of learned ConvNet metrics

The ConvNet performs better than the best linear models both in fit and test MREs. However, the ConvNet trained with our dataset is very limited in applicability to only datasets from the same region/season. When we applied our trained ConvNet to lidar inputs from a different expedition (SIPEX-II, Maksym et al., in prep) from a different season/region, the MRE is 69%, and the REM is 51%. This suggests that other seasons/regions may have different relationships between the surface morphology and SIT, which is not surprising given that snow accumulates throughout winter. The SIPEX-II data was collected during spring in coastal East Antarctic in an area of very thick, late-season ice with very deep snow with large snow drift features of length scales >20 m (which would not be resolved by the ConvNet filters here). It is also possible that datasets from spring, such as SIPEX-II, will not be as easy to train networks on because the significantly higher amounts of snow may obscure the deformed surface. Although this points out a limitation of this method, which restricts any trained ConvNet to a narrow temporal/spatial range, it also adds weight to the idea that the ConvNet is learning relevant morphological features. A ConvNet trained on Arctic data would likely learn different features (e.g. melt ponds and hummocks), although additional filters may be needed to distinguish multi-year and first-year floes.

We also tried different inputs, such as using 10 m x 10 m windows, which had training/validation/test errors of 9%/18%/25%, and using 20 m x 20 m inputs with half the resolution (i.e. 0.4 m), which had errors of 7%/13%/25%. The smaller window case has a slightly higher validation error than the above ConvNet, and the coarser-resolution input has a slightly lower validation error than the above ConvNet, but both cases have slightly higher test errors. Larger windows, which are more likely to capture surface features, are likely to improve the fit, but our dataset is too small to test this as larger window sizes would mean fewer training inputs. However, it is promising that the validation errors are lower at a coarser resolution. This suggests that this method may indeed extend to coarser, larger datasets like those

from airborne laser altimetry from OIB.

We also tried training to predict the mean snow depth given the lidar inputs, with training/validation/test errors of 15%/17%/18%, which is very similar to the thickness prediction. This is not entirely surprising as, if hydrostatic balance is valid, being able to predict the mean thickness given some snow freeboard measurements naturally gives the mean snow depth via Eq. 1.1. However, using these snow depths to predict the SIT (using the fitted coefficients for F and D from Table 3.1) gives a MRE of 43%. This is likely because the ice and snow densities are varying, but using Eq. 1.1 requires setting fixed values for these. This also suggests that a ConvNet trained to predict SIT directly would have better performance than one that predicted snow depth and used that to estimate SIT.

Although the ConvNet achieved a much lower test error than the linear fits, the inner workings of a ConvNet are not as clear to interpret. We can try to analyze the learned features by passing the full set of lidar windows through the ConvNet to see if the final layer activations resemble any kind of metric. The below analysis of features is very qualitative, as it is inherently very difficult to characterize what a ConvNet is learning.

One helpful way to gain insight on what the ConvNet is learning is to inspect the filters. Filters in early layers tend to detect basic features like edges (analogous to a Gabor filter, for example), with later layers corresponding to more complex features like lines, shapes, or objects (Zeiler and Fergus, 2014). We see similar behavior in our filters; typical filters learned in our model are shown in Fig. 3-8. Early filters highlight basic features like edges when convolved with the input array, while later filters show more complex features. These complex features are hard to interpret, but are clearly converged and not just random arrays. For example, a 'blob' feature could be a snow dune filter, while filters with a clear linear gradient could correspond to the edge of ridges. The filters in the final layer are around $\sim$8 m in size. This may be too small to resolve the entire width of the ridges in our dataset, but would be enough to identify areas near ridges. With a larger windowed lidar scan, such as those from OIB with scan width $\sim$250 m (Yi et al., 2015), we expect better feature

identification, as the entire width of a ridge can be resolved within a filter.

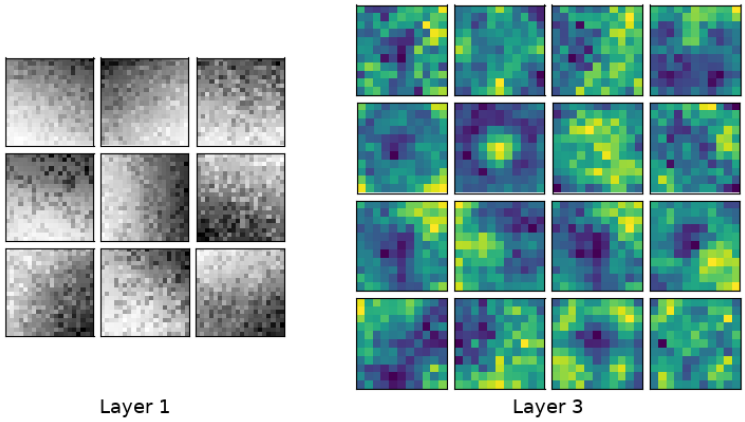

Figure 3-8: Typical weights learned in the first and last convolutional layers. Weights learned from the third layer are shown using the same colormap as the snow freeboard in Fig. 3-3 to facilitate comparison. Darker colors indicate lower weights, but the actual values are not important. The filters in layer 1 correspond to edge detectors e.g. Sobel filters, and the filters in layer 3 may be higher-order morphological features like 'bumps' (snow dunes) and linear, strand-like features (ridges). The filter size of the first layer corresponds to 4.0 m (20 pixels at 0.2 m resolution) and the third layer is 8.8m (11 pixels at 0.8 m resolution). The resolution is halved at each layer due to the stride of 2 (see Fig. 3-1)

.

The learned weights for the final (8 x 1) hidden layer and their activations (when each input window is fed forward through the ConvNet) are shown in Fig. 3-9a, grouped by category (level, ridged, snowy). These should correspond to (unspecified) metrics, which are linearly combined with the weights shown in Fig. 3-9b. It is clear that level surfaces are distinguished from ridged and snowy surfaces, but ridged and snowy surfaces show considerable overlap with each other. While it is not possible to determine with full certainty what each of the 8 features corresponds to, we can correlate these features to metrics that we may expect to be important for estimating the ice thickness and see which ones match. Doing this analysis, for ridged surfaces, features #0, #3 and #6 had a strong correlation ($|R| > 0.95$) to the mean snow

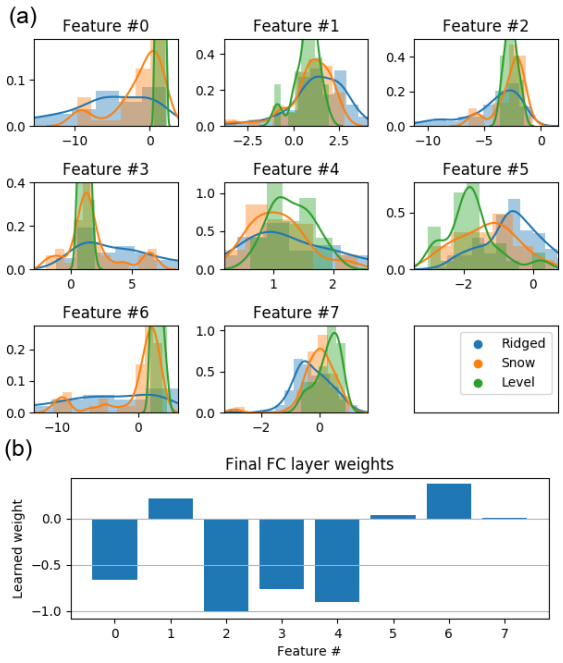

Figure 3-9: (a) Distribution of the final (8 x 1) layer activations for the level, ridged and snow categories from Fig. 3-3, and (b) the learned weights for the final fully-connected hidden layer. To generate the final thickness estimate, the activations in (a) are multiplied with the weights in (b), then summed.

freeboard (Fig. 3-10d); for snowy surfaces, these three features had a slightly weaker correlation ($0.88 < |R| < 0.96$) to the mean snow freeboard; and for level surfaces, features #1 and #5 had a slight correlation ($|R| = 0.67$ and 0.80 respectively) to the mean snow freeboard (Fig. 3-10a). However, features that correlated to the ridged surface mean snow freeboard did not correlate to the level surface mean snow freeboard, and vice versa (Fig. 3-10b and c). This suggests that the mean snow freeboard for level surfaces is treated differently (e.g. given a different effective density) than other categories.

For ridged surfaces, in addition to the mean snow freeboard, the RMS roughness was also important, with features #2 and #4 weakly correlating ($|R| = 0.61$) to the standard deviation of the window. The standard deviation had a slightly weaker correlation ($|R| = 0.55$) for level surfaces, and virtually none at all for snowy surfaces ($|R| < 0.20$). Another measure of roughness is the rugosity (the ratio of 'true' surface area over geometric surface area, see Brock et al. (2004)). This was most important for

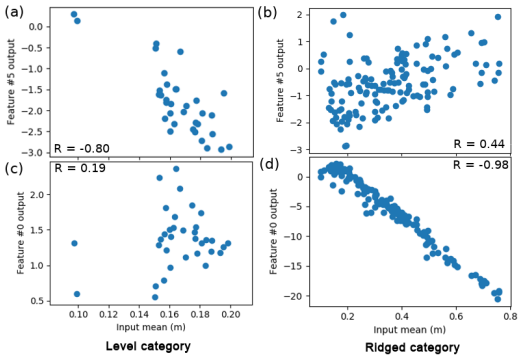

Figure 3-10: Scatter plot showing correlations between features and real-life metrics. Here, features #0 and #5 correlate strongly to the mean elevations of the level and ridged surfaces respectively, but not the other way around. This suggests that the level and ridged surfaces are treated differently, implying a different effective density of the surface freeboard. The correlation for the level category is not as strong; without the two points near $x = 0.1$, $|R| = 0.64$, so this feature is possibly a combination of the mean elevation and something else.

the snowy category, with $|R| = 0.57$ for feature #7, compared to $|R| = 0.53$ for feature #6 for ridged surfaces and $|R| = 0.22$ for feature #2 for level surfaces. As we found before, these features were much more strongly correlated to the mean elevation and standard deviation respectively for their respective surface category. This was not the case for feature #7 for snowy surfaces, which had a similar correlation ($|R| = 0.54$) to the mean elevation and a much weaker correlation ($|R| = 0.35$) to the surface σ. To summarize, for all categories, the mean snow freeboard is important (though weighted differently, as different filters are activating for different categories). For both level and ridged surfaces, the RMS roughness is important, and for snowy surfaces, the rugosity is also important. All the above analysis suggests that there are important regime differences for estimating SIT. It should be noted that these statistical metrics suggested above, with the exception of rugosity, do not account for structure (any permutation of the same numbers has the same mean/σ), which limits the usefulness of this approach to interpreting the ConvNet.

This is by no means an exhaustive list, but it suggests that the ConvNet is learning useful differences between different surface types. However, as suggested by the considerable overlap in the distributions in Fig. 3-5, these categories may also not be

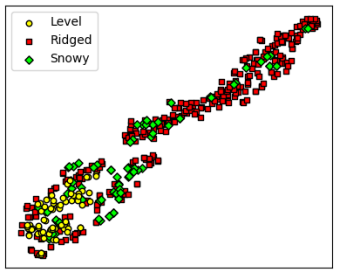

Figure 3-11: The t-SNE diagram for the encoded input, using the first fully-connected layer (feature vector of size 64) (Maaten and Hinton, 2008). The level and ridged categories are most clearly clustered, although the snowy category may also be a cluster. There is some overlap between the snowy/ridged clusters, which may reflect how ridges are often alongside snow features. It is also possible that the ridged category contains multiple different clusters. This result suggests that the manually-determined surface categories shown in Figs. 3-3 and 3-9 are pertinent, but perhaps not the most relevant, for estimating SIT given different surface conditions.

the most relevant classifications. Alternatively, a t-distributed Stochastic Neighbor Embedding (see Maaten and Hinton (2008)), which is an effective cluster visualization tool, shows that ridged and level surfaces are clearly distinguishable, but there is considerable overlap between the snowy and ridged categories (Fig. 3-11). However, the ridged category is quite dispersed, and may even consist of different classes of deformation which should not be grouped all together. Nevertheless, it is apparent that at the very least, the level and non-level categories are meaningfully distinguished. With more data and larger scan sizes (e.g. from OIB), a deep learning neural network suitable for unsupervised clustering (e.g. an autoencoder) could identify natural clusterings with their associated features (Baldi, 2012).

To emphasize the importance of the mean elevation, we also tried training the same ConvNet architecture with demeaned elevation as the input. Our ConvNet architecture is able to achieve a lowest validation error of 25% (training error 10%), but test MRE is relatively high (40%). The test error is worse than the linear model, and has twice the test MRE of our ConvNet with snow freeboard (test MRE: 20%).

We also trained the ConvNet to predict the mean snow depth, with comparable training/validation/test errors of 15%/17%/18% when using raw lidar input, and errors of 15%/22%/45% when using demeaned lidar input, which suggests the same

analyses hold for snow depth prediction. As the snow depth is largely correlated with the snow freeboard (e.g. Özsoy-Çiçek et al., 2013), with the exception of ridged areas, it is not surprising that the demeaned input is not as good a predictor of the snow depth. However, when metrics obtained from the demeaned snow freeboard (such as roughness) are combined with the mean snow freeboard, snow depth estimates (as well as SIT estimates) are improved. This may mean that aside from the mean snow freeboard, surface lidar scans may contain other information (e.g. morphology) capable of improving both SIT and snow depth predictions. This is promising for applications to larger datasets such as OIB or ICESat-2.

Another approach to analyze these learned weights is to look at the sign of the weight and the typical values of the activations in Fig. 3-9. Feature #0 has a negative weight for which the ridged category (and to a lesser extent, snowy) has the largest (most negative) feature values; this leads to adding extra thickness, primarily for the ridged ice category. This perhaps accounts for a higher percentage of ice freeboard in the snow freeboard measurement than for the level and snowy categories. Indeed, most of the level category have values near 0 for this feature. This could therefore be interpreted as a 'deformation correction' of some sort, or increasing the effective density of the ridged surface (perhaps due to a higher proportion of ice). This is also the case for features #3 and #6, which is not surprising as these three features all had strong correlations to the mean elevation for the ridged/snowy categories.

Features #5 and #7 both show some distinguishing of the different surface types, although the weights are so small for these features (Fig. 3-9b) that they are likely not significantly changing the SIT estimate and we do not speculate what these may account for.

The inner workings of ConvNets are not easily interpreted, but the analysis here suggests that the ConvNet is responding in physically realistic ways to the surface morphology. It may be possible to use these physical metrics to construct an analytical approximation to the model, but due to the nonlinearities in the ConvNet as well as the considerable scatter between the features and our guessed metrics, this will not be as accurate as simply passing the input through the ConvNet. Our approach

to choosing the ConvNet hyperparameters was not exhaustive, and a grid-search approach to optimize the number of layers and filter sizes could be done to potentially improve our prediction accuracy.

3.5 Conclusions

Statistical models for SIT estimation suffer from a lack of generalization when applied to new datasets, leading to high relative errors of up to 50%. This is problematic if attempting to detect interannual variability or trends in ice thickness for a region. Deep learning techniques offer considerably improved accuracy and generalization in estimating Antarctic SIT with comparable morphology. Our ConvNet has comparable accuracy to a linear fit (15% MRE vs. 20% MRE) but it has much better generalization to a test floe (20% MRE vs. 28% MRE for applying the best linear fit). This linear fit uses additional snow depth data not included in the ConvNet; without this data, the linear fit has an even higher test MRE of 31%.

We find that even for level surfaces, there is a considerable varying ice freeboard component that creates an irreducible error in simple statistical models, but can be accommodated as a morphological feature in a ConvNet. Our error in estimating the local SIT is <20% (RMS error of ∼7 cm) and the resulting mean survey-wide SIT also has lower errors (RMS error: 2-3 cm) than empirical methods (11-15 cm, see Özsoy-Çiçek et al. (2013)).

In applying any model to a new dataset, it is assumed that the relationships from the fitted dataset hold for the new dataset. We already showed that linear fits do not hold for different datasets (even from the same region/season), with the MRE increasing substantially, likely due to differing snow/ice proportions in the snow freeboard. This is true even when applying relationships from some PIPERS floes on other PIPERS floes. In addition to different surveys having different freeboards, ice/snow densities may also be differently distributed between surveys. Our ConvNet has errors of 12-20% when estimating both the local and survey-wide thicknesses of a test dataset, which is only slightly higher than the validation errors of 7-15%.

This suggests that the morphological relationships learned in the ConvNet also hold for other floes of comparable climatology, which in turn suggests that deformation morphology may be consistent within the same region/season.

Although our survey consists of high-resolution lidar, snow and AUV data, we really only need high-resolution lidar data. Lidar surveys are much easier to conduct than AUV surveys, and so a viable method for obtaining more data for future studies is to use a high-resolution lidar scan, combined with coarser measurements of mean SIT (e.g. with electromagnetic methods, as in Haas (1998)). Snow depth measurements are not needed with this method. This should greatly reduce the logistical difficulties to extend these methods to more regions/seasons.

Another possible strength of our proposed ConvNet is that it could account for a varying ice/snow density, with greater complexity and accuracy than an empirical, regime-based method. Although recent works like Li et al. (2018) have attempted to vary effective surface densities using empirical fits, these are not effective at higher resolutions, where snow/ice proportions may vary locally. Although the workings of ConvNets are somewhat opaque, we have shown that our ConvNet takes into account the spatial structures of the deformation, and given plausible justifications for why the snowy, level and ridged surfaces are treated differently. The learned filters suggest that morphological elements are important for SIT estimation.

Although our ConvNet would be greatly improved with more training data, it is promising that local SIT can be accurately predicted given only snow freeboard measurements. More extensive lidar/AUV/snow measurements from different regions/seasons would improve the ConvNet generalization.The window size of 20 m x 20 m used here may also be valid, with some modifications, to work on OIB lidar data, as the learned features at $\sim$8 m resolution are also resolved by OIB lidar data (resolution 1-3 m).

We have shown that surface morphological information can be used to improve prediction of sea ice thickness using machine learning techniques. This provides a proof-of-concept for exploring such techniques to similarly improve sea ice thickness prediction (particularly at smaller scales) for airborne or satellite datasets of snow

surface topography. While the ConvNet technique presented here is not directly applicable to linear lidar data such as from ICESat-2, related methods that exploit sea ice morphological information might help improve sea ice thickness retrieval at smaller scales from ICESat-2. Alternatively, using a larger training set, it may be possible to use deep learning-based methods to more readily identify relevant metrics for predicting SIT that may be measured/inferred from low-resolution, coarser data like ICESat-2 or Operation IceBridge.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Regional Snow Depth Predictions

This work was published in *Remote Sensing* (Mei and Maksym, 2020) and the content is reproduced here, with minor formatting edits and additional analysis in Sect 4.3.2.

Abstract

The snow depth on Antarctic sea ice is critical to estimating the sea ice thickness distribution from laser altimetry data, such as from Operation IceBridge or ICESat-2. Snow redistributed by wind collects around areas of deformed ice and forms a wide variety of features on sea ice; the morphology of these features may provide some indication of the mean snow depth. Here, we apply a textural segmentation algorithm to classify and group similar textures to infer the distribution of snow using snow surface freeboard measurements from Operation IceBridge campaigns over the Weddell Sea. We find that texturally-similar regions have similar snow/ice ratios, even when they have different absolute snow depth measurements. This allows for the extrapolation of nadir-looking snow radar data using two-dimensional surface altimetry scans, providing a two-dimensional estimate of the snow depth with ∼22% error. We show that at the floe scale (∼180 m), snow depth can be directly estimated from the snow surface with ∼ 20% error using deep learning techniques, and that the learned filters are comparable to standard textural analysis techniques. This error drops to ∼14% when averaged over 1.5 km scales. These results suggest that surface

morphological information can improve remotely-sensed estimates of snow depth, and hence sea ice thickness, as compared to current methods. Such methods may be useful for reducing uncertainty in Antarctic sea ice thickness estimates from ICESat-2.

4.1 Objectives

As described in Chapter 3, determining the regional and interannual variability in sea ice thickness is important to improve our understanding of the climate, e.g. via accurately simulating energy balances in general circulation models. In particular, the biggest obstacle to obtaining estimates of sea ice thickness over a large range is the difficulty of measuring SIT, which is typically done via sparsely-conducted surveys with autonomous underwater vehicles or via sparsely-conducted *in situ* sampling (drilling). In contrast, there are many more measurements of the snow surface, e.g. with airborne lidar (from Operation IceBridge) or satellite (e.g. ICESat-2). The OIB lidar data, in particular, also comes with snow depth data, which can in theory be used in the absence of direct SIT measurements to estimate SIT via Eq. 1.1.

The previous chapter showed that the lidar scans could be used to predict SIT with much better generalization than standard linear regressions. In this chapter, we seek to apply this concept to the much larger OIB dataset. Because there is no SIT data with OIB, we seek to predict snow depth instead, and then discuss the implications for predicting SIT using Eq. 1.1.

There is one important difference between the snow depth data and lidar data of OIB. Although the measurements are taken concurrently, the spatial resolutions are not trivially comparable. The single biggest obstacle is that the snow radar samples across a line (along-track), whereas the lidar data spans a 2D swath of width $\sim$250 m. This means that the snow depths may not be a fair sample of the surrounding lidar; in addition, the snow radar has sampling issues of its own, such as not being able to resolve snow depths < 8 cm. This means that the snow depths need to be adjusted in order to be representative of the adjacent lidar. This motivates the development of an textural segmentation and extrapolation algorithm, presented in Sect. 4.2.1, to extend

the snow depth measurements to be representative of the surrounding lidar span. We can then use the window-averaged snow depths that match the lidar windows (from Fig. 2-11 in Sect. 2.3.2) as training labels for a convolutional neural network. We discuss the possible bias of this extrapolation and compare its magnitude to the sampling bias of the snow radar in Sect. 4.4.1. Using the extrapolated snow depths as the ground truth, we compare the errors of the ConvNet estimates to linear fits (Sect. 4.3), and test the generalization of these different fits to other flights in the same region. We examine the filters to identify plausible metrics that the ConvNet is learning (Sect. 4.4.2), and then discuss the implications for SIT estimates (Sect. 4.4.3).

4.2 Methods

4.2.1 Textural segmentation of snow surface

Our textural segmentation approach is based on the assumption that distinct sea ice types within a local region will have similar morphology, and will have experienced similar snow deposition conditions. For example, thicker ice of a similar age with similar ridge distribution that are nearby each other will likely have accumulated similar amounts of snowfall, and winds will have redistributed and shaped the snow cover on these floes in similar ways. Thinner young ice nearby would have a different growth and deformation regime, and accumulated a different snow depth distribution. A good example would be for a large floe that has subsequently broken into many smaller floes, with younger ice forming in the open water between the older floes. Similar assumptions were used in a simpler earlier study to extrapolate *in-situ* snow depths to a larger area (Worby et al., 2008).

Each lidar input (180 x 180 at 1 m resolution) is first normalized by the maximum snow freeboard in the window, to become an integer between 0-255. We use square windows so that these may also be used with deep learning methods described in Sect. 4.2.2. This essentially turns the lidar scans into grayscale imagery. From

this, the local entropy is calculated, using a circular structuring element of size 10. However, it is still important to consider the characteristics of the snow freeboard values before normalization. For the lidar elevation measurements, the mean and standard deviation (computed on the non-normalized data, to preserve the effect of variations in freeboard among segments) have already been mentioned as relevant statistics: extending these to higher orders gives the skew and kurtosis. We use the linearized versions of these statistics, called the L-skew and L-kurtosis, which are more robust to outliers (Hosking, 1990). Outliers, in this scenario, would be extreme snow freeboard measurements, such as the sails of pressure ridges or extremely high snow depths. The skew is a measure of the asymmetry of the distribution, which is linked to how texturally uniform (or not) the snow surface is. The kurtosis is a measure of the 'tailedness' of a distribution, which may link to the proportion of deformation within a window.

We convolve a set of 20 Gabor filters with each window in order to distinguish different textural areas. The filters are oriented at 45, 90, 135 and 180 degrees, corresponding to wavelengths 2.8-45 m, with Gaussian kernels of size 11, standard deviation 7, zero phase offset and uniform spatial aspect. These hyperparameters were selected by manual experimentation. The filtered images are then passed through a nonlinear transducer (essentially a sigmoidal activation function using hyperbolic tangent) to accentuate high-activation areas, following Jain and Farrokhnia (1990), and Gaussian-smoothed (kernel size 15). Filtered images with a low variance (below 0.0001) are discarded. The resulting feature vector consists of the filtered images and x and y coordinates of each pixel. Similar textures should have similar activations, and so the feature vector can be clustered using k-means clustering with a fixed number of clusters (here, we chose 6 to ensure that we are able to capture all the possible surface types within any 180 m window, which is unlikely to exceed 6). We then look for contiguous segments (some of which may have the same cluster label) by using a standard open-source contour-finding method from Suzuki and Abe (1985) implemented in OpenCV. Then, for each contour (segment), the neighboring segments are found, and their mean entropy and L-kurtosis are computed. These metrics were

chosen by manually classifying 2000 images and using a decision tree to identify the most relevant metrics for distinguishing snow features from deformed ice. If the mean entropies of a segment and its neighbor are within 2.5%, and/or if the L-kurtosis are within 2%, then they are merged. Then, all segments are assigned a unique label. Small windows (less than 1% of the lidar window size) are erased and their pixels are 'absorbed' by their nearest neighbors. This is repeated until there are no more segments that can be merged. This process is illustrated in Fig. 4-1 and summarized in the pseudo-code in Algorithm 4.1. We can now identify the segments that have snow depth measurements.

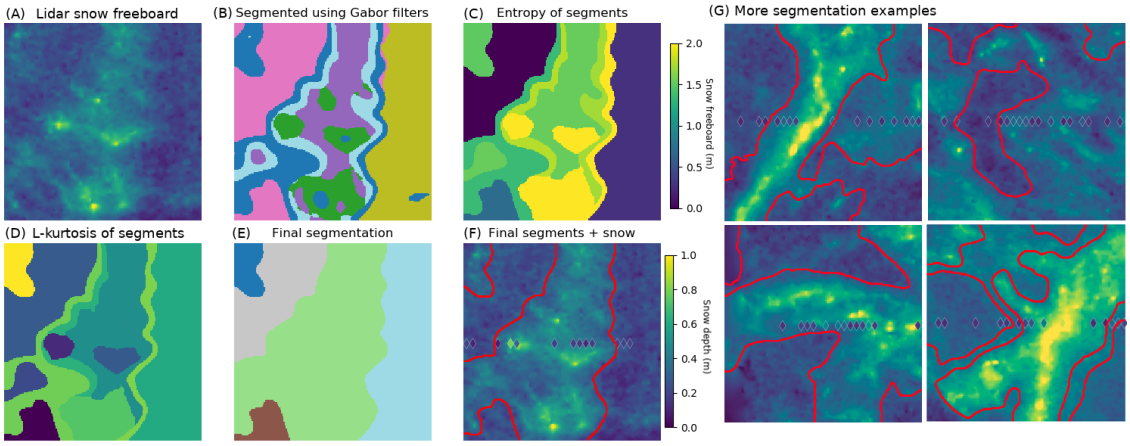

Figure 4-1: Example of the textural segmentation algorithm. The lidar (A) is normalized, essentially making it a grayscale image. This is segmented using Gabor filters (B). The image entropy (C) and L-kurtosis (D) of each segment is shown; 'similar' ones are merged together recursively until a final segmentation (E) is obtained. The segments, along with the snow depth measurements (diamonds) are shown in (F). We deliberately choose to over-segment in (B), as it is easy to merge small segments. More examples of the segmentation algorithm are shown in (G).

For each snow depth D within some segment, we work out the mean elevation of the surrounding 7 m x 7 m lidar window (even if this window crosses into another segment), and take this mean value as F. We chose 7 m in order to match the snow depth footprint size. We then work out the F/D ratio. This essentially is a more complex version of Steer et al. (2016), who used different empirical linear fits to different ice freeboard regimes. Then, all the ratios from each snow point are combined, and the maximum, minimum and average ratios are stored. The 'average'

Algorithm 4.1: Segment the snow surface and extrapolate the snow depths

Normalize each input such that the lidar window is now an 'image' of 8-bit images, with a maximum value of 255.
Calculate the local entropy, using a circular structuring element of size 10.
Convolve a set of 20 Gabor filters through the lidar window.
Pass these convolved outputs through a nonlinear transducer (e.g. tanh).
Gaussian-smooth the results (kernel size 15).
Discard filtered images that have a low variance (below 10^{-4}).
Compile the resulting feature vector of the filtered images + the x-y coordinates of the lidar window (total layers: 22).
Apply a clustering scheme to identify segments (e.g. k-means, with k=6).
Identify contiguous segments in the resulting clusters, using a contour-finding method (e.g. OpenCV).
Count number of segments N.
while N not constant: **do**
  For each segment, work out the mean entropy and the mean L-kurtosis.
  Identify adjacent segments.
  if difference in entropy is within 2% **OR** difference in L-kurtosis is within 2.5%: **then**
    Merge these segments together.
  end if
  Count number of segments.
end while
for each segment with no snow depths: **do**
  Work out the mean, standard deviation, L-kurtosis and mean entropy (4 metrics) for that (target) segment
  Work out same 4 metrics for all segments within ±30 km
  Evaluate the geometric mean of the difference between the 4 above metrics with respect to the target segment (see Appendix A for worked example).
  For all 'matched' segments (geometric mean below the similarity threshold), work out the (harmonic) mean F/D ratio of all the F/D in all matched segments.
  Apply this mean F/D to the F of the target segment, to generate an estimated D.
end for

ratio here is defined as the harmonic mean, to prevent high F/D ratios (from ridges) from skewing the (arithmetic) mean high. The harmonic mean is generally preferred over the arithmetic mean when comparing ratios: this will be examined in Section 4.3.1.

To work out the extrapolated snow depth estimate for a given segment, we first compute the mean entropy, mean L-kurtosis, mean elevation and standard deviation for all segments. Then, for the target segment, we work out the average absolute difference in these four metrics for all other nearby (within a ± 10 km range) segments. For this average, we use the geometric mean to account for the different scales of these metrics. All segments that are within a similarity threshold are identified. Their mean F/D ratios are compiled, and the weighted harmonic mean ratio is computed (weighted by similarity and also number of snow measurements in that segment. This ensures that segments with more snow measurements, and/or higher similarity, are weighted higher). The mean elevation of the target segment is divided by this (harmonically-)meaned ratio to determine the 'true' mean snow elevation of that segment. If there are at least 9 snow points amongst the identified nearby 'similar' segments, then the extrapolation is deemed a 'successful completion'. An illustrated example is given in Appendix A.

In order to test the accuracy of this method, we first take only those segments that contain snow radar measurements. For each segment, we compute the estimated mean snow depth using the extrapolation algorithm applied to that segment (and excluding that segment itself from the list of possible matches). We compare this estimate to the 'true' segment mean snow depth, which is computed as the segment mean F scaled by the harmonic mean of all F/D ratios taken at all snow depth measurement locations in that segment. We call this the 'true' mean snow depth, in contrast to the 'raw' mean snow depth which is just the (potentially biased) mean of all snow measurements in that segment. A variety of similarity thresholds were chosen; higher thresholds lead to more successful completions but higher errors. For all thresholds, using the weighted raw mean led to higher error than copying the F/D ratio instead. The proportion of successful completions is heavily affected by the chosen similarity threshold: higher

thresholds allow more dissimilarity between 'matched' segments, which increases the completion rate. We chose a similarity threshold of 0.03, increasing in increments of 0.005 until 0.05, and keeping the threshold once at least nine snow points were being used for the extrapolation. This resulted in a 97% completion rate, a mean/median absolute error of 11.0 cm/6.6 cm, and a mean/median relative error of 39.1%/23.2%. It is important to note that these errors are an upper bound on the actual error, as the mean of the raw snow depths may not be an accurate estimate of the actual mean snow depth in that segment. This is particularly true for segments that have high snow depth variance but few radar observations, such as around large, deformed areas that have few snow depth measurements. If we only check the error for segments that have 9 or more snow depth measurements,which increases the likelihood that the sampled mean snow depth provides a reasonable estimate of the actual mean snow depth in that target segment, then the mean/median relative errors drop to 22.5%/16.4% and the mean/median absolute errors to 8.5 cm/5.6 cm.

Note that the harmonic mean F/D ratio is not the same as taking the ratio of the mean elevation and mean snow depth within a given floe segment. This is because the snow depths may be biased (in particular when the snow cover is too low to return), and also because the snow samples all fall on a fixed line, which may not be necessarily representative of the segment. In particular, we know that mean snow depths are biased high because the peak-finding algorithm does not return a snow depth when the snow is too thin (<7 cm). Taking the mean ratio attempts to address this bias, but still requires the assumption that the measured snow depths have a similar F/D distribution to the unmeasured snow depths. For comparison, we also apply the extrapolation algorithm to estimating the raw mean snow depth. The algorithm is the same, though instead of taking the weighted harmonic mean of the F/D ratios, we take a weighted arithmetic mean of the raw mean snow depths, again weighted by the number of samples per segment and the similarity. This gave slightly higher errors (mean/median: 25.5%/18.6%) than using the F/D ratio, again taking the error of segments with at least nine snow measurements only.

4.2.2 ConvNet for learning mean snow depth

The previous sections presented a technique to estimate snow depths over a larger region given a small number of measurements based on snow surface texture. However, we now attempt to use deep learning to see if snow depth can be predicted directly from this snow surface texture only. To do this, we apply a convolutional neural network (ConvNet), as this is optimized to identify spatial features. This builds on our previous work which demonstrated this was effective in predicting snow and ice thickness at small scales (Chap. 3). As this is a regression problem, we used a mean squared error loss function. The inputs consist of 180 x 180 lidar windows. There are 4 convolutional layers with filter sizes 14, 10, 10 and 14, each with stride 2 (in lieu of max-pooling) and the SELU activation function (Klambauer et al., 2017), with two fully-connected layers of 512 and 64. The true mean snow depths used for learning are the window-averaged ones as determined from the prior analysis. This provides a more robust mean snow depth estimate than the radar observations on their own when there are few snow depth observations, or when they may be biased by sampling only certain snow and ice regimes within the window. The learning rate was 8×10^{-4}, which gradually decreased to 7.2×10^{-5}. The optimizer used was Adam (Kingma and Ba, 2014) with weight decay of 10^{-5}. 80% of the windows for the 2010 flight, randomly sampled, were used as the training set, and the remaining 20% were used as the validation set. The 2016 flight (the test set) provides a separate data set with different statistics that can be used to evaluate the efficacy of a ConvNet developed for one region or season against an independent dataset. As such, it can provide a measure of the ConvNet's ability to generalize learned features to new datasets, which may allow it to detect variability in snow depth for other datasets, including those that lack snow depth observations.

4.3 Results

The relationship between D and F does not necessarily appear to be one-to-one. In Fig. 4-2, we see the risk of using a mean snow value as a function of F: at

high freeboards, the snow depth is likely to be either equal or nearly equal to snow freeboard *or* independent to snow freeboard, and using a mean value for local snow depth predictions will therefore always be wrong. For larger scale averaged snow measurements, an average measurement may be acceptable, but strictly speaking, this average should be weighted by the occurrence of the two different regimes ($D = F$ and D is independent of F). This is naturally very influenced by sampling bias. The bimodality of the higher-F data also motivates the use of deep learning or other methods that can distinguish the different D values that may be associated with a high F.

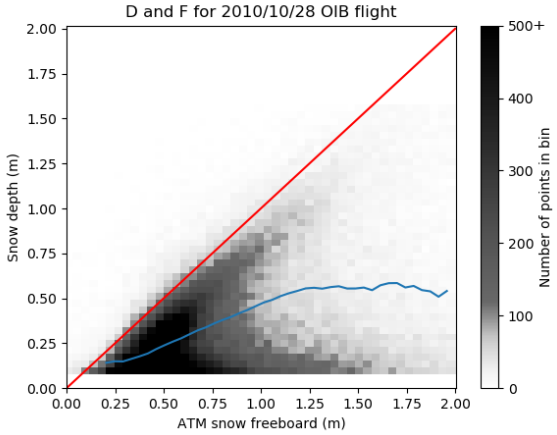

Figure 4-2: A histogram showing the point density for different F and D values. F is binned at 3.3 cm and D is binned at 3.7 cm. The $F = D$ line is shown in red, and the mean value of D for each F bin is shown with the blue line. This can be used as a interpolating empirical function for determining D, i.e. $f(F)$=D (e.g. Kwok and Maksym, 2014). At higher F values, the mean D may give a biased estimate of the snow depth as the true distribution is bimodal.

4.3.1 Extrapolated snow depths

The mean of all snow depth measurements for the 2010 dataset, assuming an average snow density of 300 kg m^{-3}, is 37.7 cm (σ = 29.1 cm). The mean F corresponding to these points is 83.7 cm (σ = 49.4 cm). However, the overall mean F for the entire set of lead-referenced snow freeboards (not just those with D measurements) is 72.6 cm (σ = 53.6). This is suggestive of a sampling bias, as will be explored in Section 4.4.1.

Assuming that the F/D ratio is the same for these two samples, this suggests that the true unbiased mean snow depth is closer to 32.7 cm. Using an empirical functional fit $f(F)=D$ (for the mean snow depth for some given F as shown in Fig. 4-2), $f(72.6$ cm) = 34.3 cm. The mean snow depth of the set of snow depth measurements that are contained within lidar windows is 38.2 cm. This is higher than the overall mean, likely because the lidar windows exclude any image with too much open water, where thin ice and thin snow are more likely to be found. If the (harmonic) mean F/D ratio for all corresponding snow points is applied to the mean F (76.0 cm) of the corresponding segments, the mean snow depth is 34.2 cm. This is slightly lower than taking the arithmetic mean of D (38.2 cm), in part because the snow depths do not sample thin snow (< 7 cm) which likely have lower F. If applying the best empirical functional fit for snow depth (following the methodology of Kwok and Maksym (2014)), the mean snow depth corresponding to $F = 76.0$ cm is 36.0 cm. Both the F/D ratio and empirical function suggest that the mean of the snow depth measurements is biased high, by 2-4 cm (around 10%). See Section 4.4.1 for further discussion of the biases in the snow data.

We find that both copying D and copying F/D lead to good accuracy in the overall survey-wide mean snow depth, within a few mm. Copying F/D has lower mean relative errors (22.5%) than copying D directly (25.5%), and also lower mean absolute errors (F/D: 8.5 cm; D: 9.4 cm). We therefore use F/D ratios for extrapolating the snow depths to those off-nadir segments that have no snow depths. As an additional advantage, using the F/D ratio accounts for the sampling bias for D just discussed. We have scaled the mean F/D ratio for each segment by 1/0.97, to correct for the tendency of the harmonic mean to slightly underestimate the true mean as shown in Fig. 4-3. Note that our algorithm only looks for similar segments within 10 km, as we found that the error would increase along with a higher completion rate if we looked for similar segments in the full flight (mean error: 25.2%). We also tried looking for textural matches in the 2016 dataset to extrapolate the 2010 dataset, as a test of whether textural similarity generalized across datasets. This had slightly higher errors (mean: 28.4%), suggesting the existence of common textural features for the

same region/season for different years.

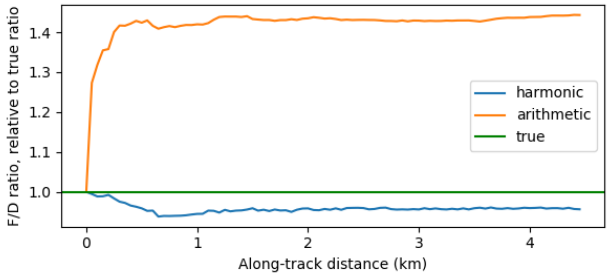

Figure 4-3: The convergence of the harmonic and arithmetic means, vs. the 'true' mean $\frac{\text{mean}F}{\text{mean}D}$, averaged over 58 segments that spanned at least 4.5 km. There is a tendency for the harmonic mean to slightly underestimate the true mean ratio by 3%.

4.3.2 ConvNet results

Results are summarized in Fig. 4-4. The ConvNet had a lower training, validation and testing MRE than a corresponding linear regression fit to the F and D data. This is likely due to the ConvNet learning more advanced metrics and implicitly learning different F/D ratios, whereas a linear regression assumes one overall mean F/D ratio. We also use the forward Kullback-Leibler (K-L) divergence $D_{KL}(\text{True}||\text{Prediction})$ between the true and prediction distributions, an (asymmetric) measure of the difference in distribution, as a measure of the goodness-of-fit of the predicted distribution of snow depth (Kullback and Leibler, 1951). The K-L divergence is defined as $D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} (\neq D_{KL}(Q||P))$ for all x in the probability space of $\mathbf{P}$. When $Q(x)$ is zero, the K-L divergence tends to infinity - we therefore add a small offset of 0.001 to the denominator to avoid this. Because snow depths are more likely to be low than high, it would be easy to get a low MRE just by (for example) uniformly predicting the (thin) modal snow depth, but this kind of result would not be useful to expand our records of Antarctic snow depth. In general, knowing the distribution of snow depth or SIT is necessary to estimate, for example, energy fluxes, which are non-linear with respect to ice thickness and therefore cannot be estimated with just the mean thickness (Leppäranta, 1993; Schramm et al., 1997).

More analysis regarding the ConvNet performance is given in Section 4.4.

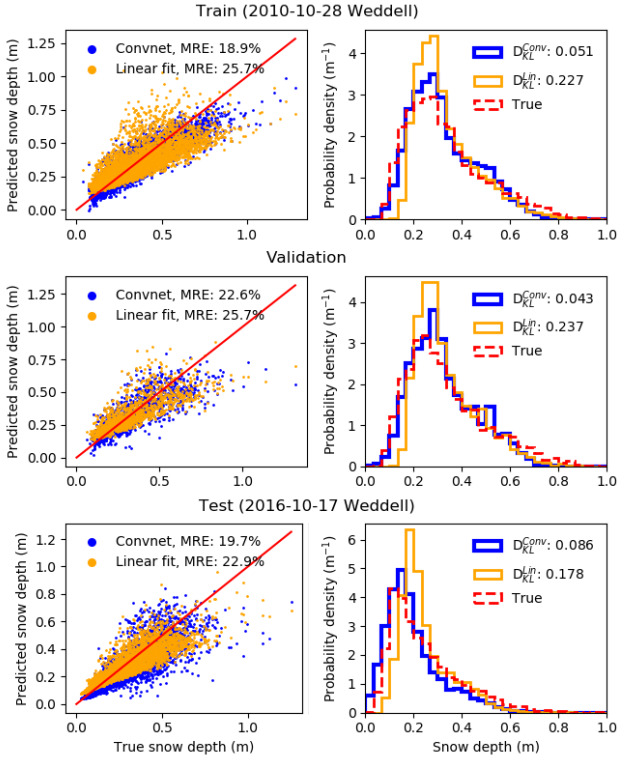

Figure 4-4: Training, validation and test results for predicting snow depth (averaged over a 180 x 180 m window) with a ConvNet. The training/validation sets are randomly sampled from the 2010/10/28 dataset, and the test set is the 2016/10/17 dataset. The linear fit is fitted to the training set only, and then applied to the validation and test sets. The resulting snow depth distributions for each model is shown in the right panels, along with their forward K-L divergences from the true distribution. For comparison with other studies that use RMS error, our training/validation/testing RMS errors are 4.6/5.9/4.4 cm.

Figure 4-5 shows the ConvNet and linear model results, as well as the raw (measured), window-averaged and extrapolated snow depths. When there are many raw points of a similar snow depth (e.g. near x = 2.5 km), the surface is likely fairly uniform and the radar-sampled snow depths are likely an unbiased sample of the true snow depth of the lidar window. The extrapolated snow depth (red) and raw mean (green) therefore are in good agreement. In contrast, near ridges, the measured snow depths have a greater vertical variability (e.g. the peak near x = 1.1 km has a snow depth of 0.9 m). For this example, the raw mean snow depth (green) exceeding the raw mean snow freeboard (black) is suggestive of a sampling bias, and in any case

indicates that the mean of the raw snow depth measurements should not be taken as the true mean snow depth of the window. The snow depth samples are likely biased by the deep snow around the ridge, which may only be a minority of the surface, and so the raw mean snow depth (green) is higher than the extrapolated snow depth (red). Possible biases are discussed in greater detail in Section 4.4.1. Another point to note is that the linear model for snow depth necessarily assumes that similar (mean) freeboards have similar (mean) snow depths. In Fig. 4-5, because the mean freeboards vary narrowly between 0.26-0.48 m, the linear estimate of the windowed snow depth also has a narrow range, of 0.15-0.24 m. In contrast, the raw and extrapolated snow means both have a higher range, of 0.11-0.34 m and 0.08-0.28 m respectively, which is matched by the ConvNet estimate of 0.08-0.27 m. The linear estimate has on average higher errors for predicting the snow depth at each window along this flight track (32% vs 23% for the ConvNet). Similarly, the linear estimate has worse performance when estimating the extrapolated mean snow depth (15.5 cm) along the entire 5 km flight track (linear: 18.9 m, 22% error; ConvNet: 17.3 cm, 12% error).

4.4 Discussion

4.4.1 Effectiveness of segment texture-matching

Snow depths corresponding to successfully matched segments are typically lower than the overall average. This is shown in Fig. 4-6, which shows the distribution of the snow depth and freeboard for completed and non-completed segments for the 2010W flight. However, of the segments that are completed, the estimated mean from textural matching is very close, within 0.01 cm of the true mean. This is not the case if using the empirical function from Fig. 4-2, which predicts a mean snow depth of 36.0 cm, which is 2.4 cm higher than the true mean. This is likely because, as shown in Fig. 4-2, the relationship between F and D is not bijective, and at higher F values, the value for D may be bimodal: high snow freeboards may be thick snow dunes *or* deformed ice with little snow. If these regimes are texturally distinguishable, then

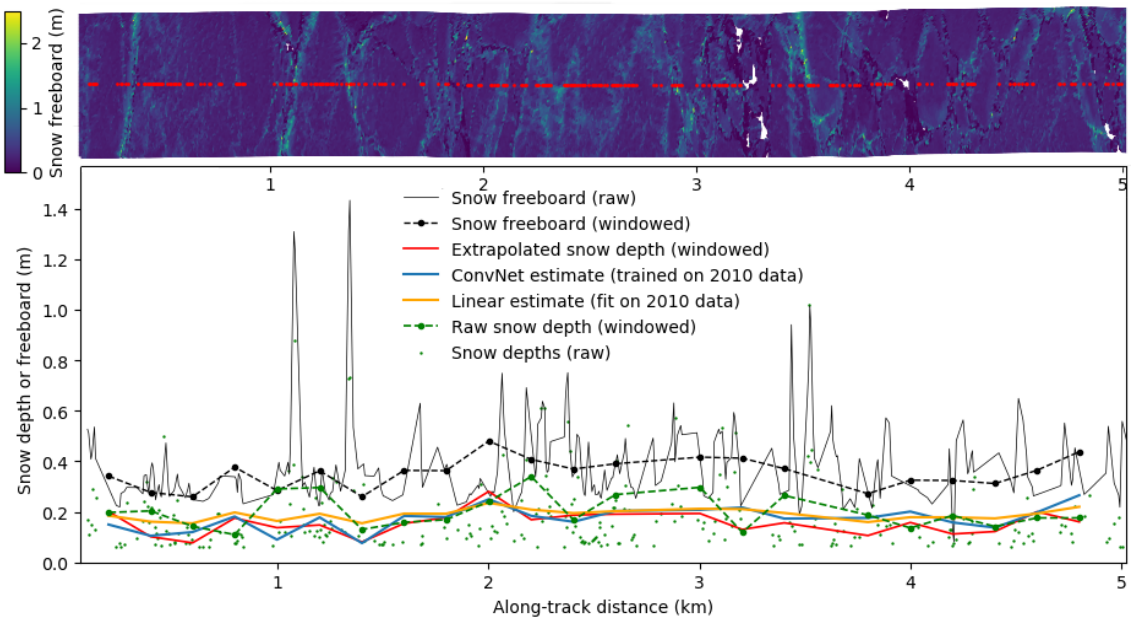

Figure 4-5: An example flight segment from the 2016 dataset, showing the snow depths and freeboards, along with the extrapolated snow depths (for the entire width of the lidar scan, instead of just the snow radar footprint), and the predicted snow depths using a linear model and a ConvNet (in both cases, trained/fitted on the 2010 dataset). Note that the raw snow freeboard (solid black line) are for the 9 m window of the snow radar footprint, whereas the mean snow freeboard (dotted black line) is for the corresponding 180 m lidar window. For the linear model, the mean relative error of this segment is 32%; for the ConvNet, it is 23%. The (extrapolated) mean snow depth for the 5 km segment is 15.5 cm; the ConvNet predicts 17.3 cm (12% error) and the linear fit predicts 18.9 cm (22% error). This is in line with our findings from Fig. 4-7. The mean of the raw snow depths (green) is 17.3 cm, which is coincidentally the same value as our ConvNet average, though it is likely biased high due to the 2-4 cm sampling bias for large-scale snow depths that was first mentioned in Section 4.3.1.

the textural matching would lead to the right regime being picked, as opposed to always applying the average snow depth. One could use a weighted average of the two regimes in Fig. 4-2 to work out the true survey mean snow depth, but this is highly subject to sampling bias (as thin snow has fewer returns). High-F segments are less likely to be texturally matched than low-F segments, perhaps because there are fewer such samples in general, and perhaps also because extremely rough ice reduces the number of successful snow radar returns (e.g. Kwok et al., 2011). This may lead to our estimated mean snow depth being biased slightly lower. However, as shown

in Fig. 4-6, our extrapolated snow distribution compares very well to the true snow distribution, and the estimate of the mean snow depth is only 0.01 cm lower. We can therefore be confident that the extrapolation algorithm, in and of itself, does not induce a significant bias.

There is also a sampling bias that results from the sampling of the snow depths due to thin snow being excluded from the radar return. Our algorithm attempts to address this sampling bias, but it is not known if using the F/D ratio instead accounts for the entire sampling bias correctly. This is also shown in Fig. 4-6 (green and blue lines). By comparing the distributions of the completions that do and do not have snow depths, we can see if segments that have an observed snow depth have different statistical properties to those that do not (which includes both segments that, if the radar flew over them, would have a snow depth and those that would not). This is effectively comparing whether the snow depth distribution where the radar was able to resolve the snow depth is different from the snow depth distribution in locations where it cannot. Their mean snow depth estimates are 33.6 and 29.0 cm respectively, suggesting a sampling bias of $\sim$+5 cm. This is in line with our finding from Section 4.3.1, which found that the mean snow depth was biased low by 2-4 cm because the mean F was slightly larger for snow depth sampling points than for the rest of the lidar scan and noting that larger F tends to be associated with larger D.

There is another sampling bias, called the 'subsampling' bias here, that results from our algorithm only taking snow depth estimates from those snow depths that are contained within the lidar window. In particular, this means that snow depths that are near open water are excluded if there are insufficient lidar points to form a lidar window. For this, we should compare the raw mean of all the snow radar returns, which is 37.9 cm, to the raw mean of all the snow depth points that are within a window, which is 38.2 cm. As expected, by excluding some snow points that are near open water (which tend to be associated with thin ice, and hence are generally lower snow depths), the mean snow depth is slightly higher by 0.3 cm. This 'subsampling' bias is much smaller than the sampling bias in the previous paragraph by an order of magnitude.

To summarize, there are thus *three* biases: the *extrapolation bias* from the algorithm, which is 0.1 cm; the *sampling bias* from the OIB radar, which is 5 cm, and a *subsampling bias* from the lidar windowing of 0.3 cm. Using a similar approach for the 2016W flight, the extrapolation bias is 0.2 cm, OIB sampling bias is 7.6 cm and the lidar windowing subsampling bias is 0.1 cm. In other words, here we find that the sampling bias from the OIB snow radar selectively returning snow depths dominates the biases caused by discarding lidar windows with too much open water and by textural extrapolation of the snow depths. Note that we could try, for example, a different k for our k-means approach, and also to tweak our Gabor filter sizes, as this would change our segmentation results slightly, and we can see if this could reduce our extrapolation bias even further.

For the completions (27921 segments), 50% (13971 segments) have a snow depth, but for the non-completions (2049 segments), 35% (720 segments) had snow depths. As seen in Fig. 4-6, the non-completions have a much higher mean F than the completions. This is in line with the non-completions having a higher average F (and D), as the rougher surfaces are fewer in number (i.e. harder to match), have greater variability (i.e. also harder to match), and are less likely to have a snow depth return (Farrell et al., 2012). This is consistent with our earlier finding that the extrapolated snow depths may be slightly biased low. Thus, there are two sampling bias in the snow radar data has two sources: it is less likely to sample very deep snow in rough deformed ice, but also less likely to sample very thin snow. These biases will partially offset each other. However, the proportion of high snow freeboard is relatively low compared to the proportion of thin snow for this dataset, and so the net result is the raw mean snow depth is biased high, as we find. The bias from OIB may still be considerably less than *in situ* sampling. Our deepest snow depths are deeper than typically measured *in situ* on selected floes (e.g. Massom et al., 1997; Arndt and Paul, 2018), in line with findings from Kwok and Kacimi (2018).

The segment-matching algorithm has a trade-off between completion rate and accuracy. Higher completion rates imply extrapolating with lower-quality matches, which increase the uncertainty in the extrapolation. We noticed that 'local' matches

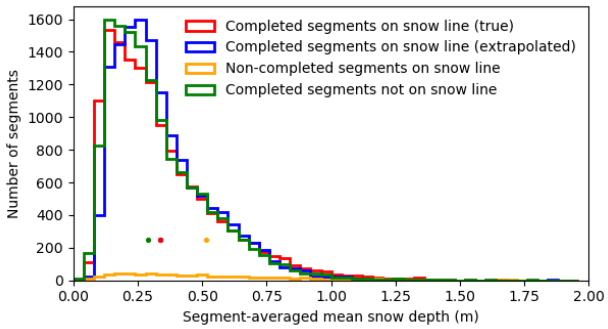

Figure 4-6: The distribution of true (red) and extrapolated (blue) mean snow depths D for successful and unsuccessful (orange) textural matches, for all segments that contain snow depth data (i.e. on the snow line), and the snow depth distribution of all successful extrapolations of those segments that did not have snow depth measurements (green) for the 2010W flight. The mean segment D for all segments on the snow line that were successfully extrapolated (red dot) is 33.59 cm; the extrapolated mean (blue dot) is 33.60 cm; for non-completions (orange dot) it is 51.70 cm. The extrapolated mean snow depth for all segments that have no snow depth measurements (green dot) is 28.96 cm.

(i.e. closer segments) give lower error estimates, but at the cost of a lower completion rate. The median relative error of using a $\pm$ 5 km search grid for the textural matches is 23.6%; for $\pm$10 km is 23.7%; using the full flight is 26.9 %. This is consistent with our assumption that nearby ice floes with similar morphology will have experienced similar snow deposition regimes. However, using a different flight entirely (e.g., extrapolating the 2010 dataset using the 2016 dataset) gave a MRE of 28.2 % and a completion rate of 99.8%. This implies that textural features show some similarity across different years in the same region, although there is enough variation that the errors from using data from different years are slightly higher. This may mean that relationships between surface texture and snow depth exhibit universal or regional behavior. As such, this technique might be effective broadly around the Antarctic, although this needs to be confirmed with data from different regions that experience different snow and ice regimes, such as the Bellingshausen and Amundsen Seas.

Our average F/D ratio of 2.0-2.2 is much higher than (although technically within error of) an equivalent Weddell Sea dataset from Polarstern 1988 (presented in Özsoy-Çiçek et al. (2011)), which had $F/D = 1.12(\pm 0.9)$. However, their data is from an

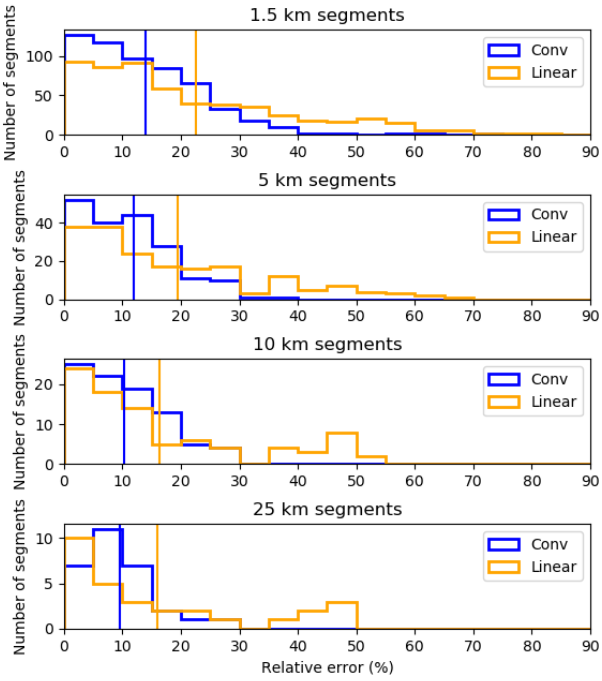

Figure 4-7: Relative error distribution of estimating the mean snow depths, at various length scales, using the linear/ConvNet models fitted to the training set (2010 OIB dataset), and applied to the test set (2016 OIB dataset). The vertical lines show the mean relative error for the corresponding model. The ConvNet is consistently better than the linear fit, though the difference becomes less prominent as the segment size increases. The mean relative and absolute error for the ConvNet with 1.5 km segments are 14.0% and 2.9 cm.

in-situ drill line, which naturally is biased towards thinner ice (lower F) for typical ice conditions during this season (spring). This further suggests that drill line data may not be fully representative of the range of sea ice conditions needed to derive satellite ice thickness estimation algorithms. However, we note that there are no co-located *in-situ* snow depths on Antarctic sea ice to validate the OIB radar data, and it is possible that the radar underestimates true snow depth, although limited comparison does not suggest this is likely (Kwok and Maksym, 2014).

4.4.2 ConvNet results

It is clear that the ConvNet performs well and is able to predict, with good generalization on the test set, mean snow depths. The mean test error of 20% is unlikely to be

able to be reduced further, as the input data (extrapolated snow depths) themselves had a mean error of 22.5% (per segment - this is likely to be reduced when averaging over multiple lidar windows).

To give us an idea of what the trained ConvNet is basing its prediction on, we look at examples of the learned filters for the first three convolutional layers (Fig. 4-8). The first two layers appear to correspond to common filters in computer vision for edge detection and textural analysis (e.g. Freeman and Adelson, 1991), which would detect linear types such as ridges or boundaries between morphological regimes. The filters in the third layer are likely to be textural components (patterns), possibly snow dunes on level ice, or rubble fields.

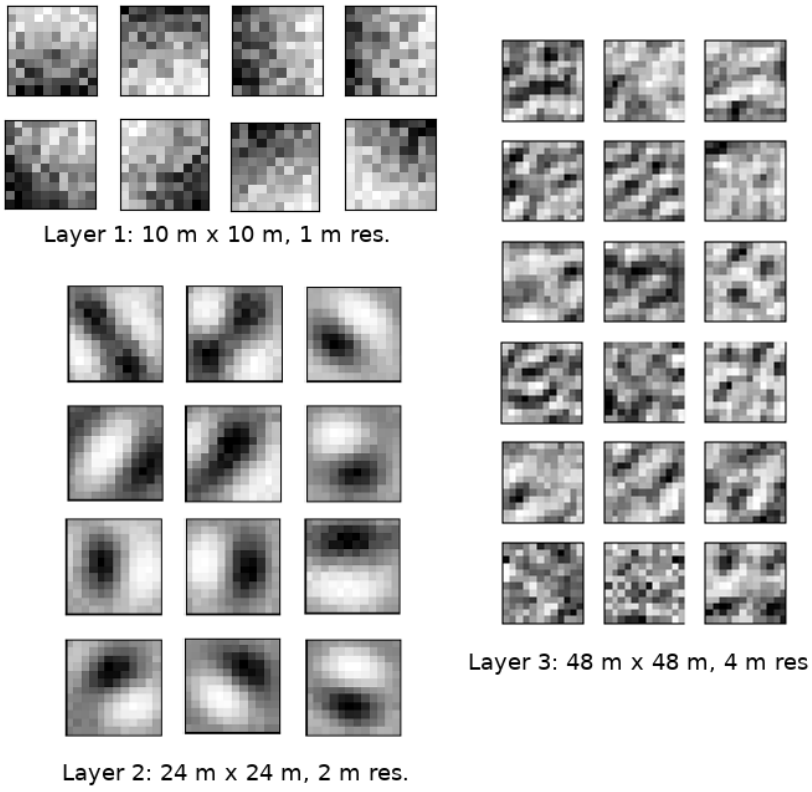

Figure 4-8: Learned weights for the first three convolutional layers. The first layer has basic gradients (with some noise), corresponding to edge detection. The second layer looks very similar to steerable pyramid kernels for G_1 and G_2 in Freeman and Adelson (1991), which correspond to the first and second derivatives of a Gaussian function. The third layer is presumably complex textural components, which are harder to interpret.

One strength of the ConvNet over linear fits is its ability to predict the snow depth accurately at lower length scales. Fig. 4-7 shows the error distributions when taking mean snow depths over segments of 1.5, 5, 10 and 25 km scales. In all cases, the total length of segments spanned by the segments exceeded 1000 km. As the length scale increases, the linear fit approaches the ConvNet, and both errors decrease. This is expected as the linear fit to the training data includes the many different surface types, and at such large length scales the variability in snow depths given some snow freeboard is low enough that linear fits are comparable. But at the smaller scales, the ConvNet performs much better, with fewer large errors. This is encouraging for the possibility of using a similar technique to improve snow depth and sea ice thickness estimation at smaller scales from ICESat-2.

4.4.3 Implications for SIT estimates

Ultimately, an improved estimate of snow depth will permit an improved estimate of SIT. Here, we evaluate whether our technique to estimate snow depth might be sufficient to accurately estimate Antarctic SIT without an independent estimate of snow depth. The first-order uncertainty in estimating sea ice thickness (ϵ_T) given laser altimetry (snow freeboard, F) and snow depth measurements (D), following Giles et al. (2007), is given by:

$$\begin{aligned}\epsilon_T^2 =&\epsilon_F^2(\frac{\rho_w}{\rho_w-\rho_i})^2+\epsilon_D^2(\frac{\rho_s-\rho_w}{\rho_w-\rho_i})^2+\epsilon_{\rho_s}^2(\frac{D}{\rho_w-\rho_i})^2+\\ &+\epsilon_{\rho_w}^2(\frac{F-D}{\rho_w-\rho_i}-\frac{F\rho_w+D(\rho_s-\rho_w)}{(\rho_w-\rho_i)^2})^2\\ &+\epsilon_{\rho_i}^2(\frac{F\rho_w}{(\rho_w-\rho_i)^2}+\frac{D(\rho_s-\rho_w)}{(\rho_w-\rho_i)^2})^2\end{aligned} \tag{4.1}$$

Using the mean F and D values from our 1.5 km segments of 0.44 m and 0.22 m respectively, typical densities of $\rho_w = 1024$ kg m^{-3}, $\rho_i = 915$kg m^{-3}, ρ_s=300 kg m^{-3}, this gives:

$$\begin{aligned}\epsilon_T^2 =& 88.3\epsilon_F^2 + 44.1\epsilon_D^2 + 4.07 \times 10^{-6}\epsilon_{\rho_s}^2 \\ &+ 2.62 \times 10^{-3}\epsilon_{\rho_w}^2 + 6.01 \times 10^{-4}\epsilon_{\rho_i}^2\end{aligned} \tag{4.2}$$

Our mean relative error of 14.0% for snow depth of the 1.5 km segments gives $\epsilon_D = 0.033$ m. We take a typical estimate of the seawater density variability of ± 1 kg m^{-3}. We take $\epsilon_{\rho_s} = 50$ kg m^{-3}, following Kwok and Maksym (2014), which was based on variability in large-scale average snow depths observed in different regions and years (Massom et al., 2001). The uncertainty from snow freeboard is likely due to the lead height accuracy, as well as the range resolution of the lidar. Over a 1.5 km segment, there are sufficient points that the range resolution of the lidar itself contributes little uncertainty. However, the lead height uncertainty (typically 3 cm, (Kwok and Maksym, 2014)) also contributes to the uncertainty. If there are two leads, each with RMS error 3 cm, then assuming the errors are distributed normally, the uncertainty of their average is $\frac{3}{\sqrt{2}} = 2.1$ cm. As the local sea surface height is a weighted average of the nearby leads (in our case, inverse-distance weighting), the resulting variance is a weighted sum of each variance (for simplicity, we assume all leads have RMS error 3 cm), with the weights summed in quadrature. Working this out for all our lidar windows gives a weighted average variance of $0.54\sigma^2$, which gives a snow freeboard uncertainty of $\epsilon_F = \sqrt{0.54} \times 3.0 = 1.6$ cm. The density of sea ice is a large source of uncertainty, due to the variable porosity of sea ice (and that the pores may be filled with either seawater, as is typical for first-year ice, or air, as is typical for multi-year ice), and variable (and unknown) macroporosity (space between rubble blocks) of ridges. We take the uncertainty in sea ice density as $\epsilon_{\rho_i} = 20$ kg m^{-3} for first year ice following Maksym and Markus (2008) (note, this would be considerably higher if we consider that the ice may be multiyear, summer sea ice, or ridges may be unconsolidated).

Putting these values into Eq. 4.2, we have:

$$
\begin{aligned}
\epsilon_T^2 =& 88.3 \times 0.016^2 + 44.1 \times 0.033^2 + 4.07 \times 10^{-6} \times 50^2 \\
&+ 2.62 \times 10^{-3} \times 1^2 + 6.01 \times 10^{-4} \times 20^2 \\
=& 0.023 + 0.048 + 0.010 + 0.003 + 0.240 \\
\epsilon_T =& 0.57 \text{ m}
\end{aligned}
\tag{4.3}
$$

From Eq. 1.1, $T = 2.79 \pm 0.57$ m and so our relative uncertainty in estimating SIT is 20%. This uncertainty is dominated the term corresponding to ϵ_{ρ_i}. Excluding this, the contributions from ϵ_{ρ_s}, ϵ_F and ϵ_D are similar, and the contribution from ϵ_{ρ_w} is negligible.

Our average values for F and T compare well to Yi et al. (2011), which looked at the Weddell Sea from October-December (their range for F: 0.33-0.41, T: 2.10-2.59). Our values are slightly higher than theirs, possibly because the OIB flights take place in October, when the F and T values may be slightly higher than in December (summer).

4.5 Conclusions

We have demonstrated the viability of extrapolating snow depth measurements from nadir-looking (1-dimensional) radar datasets from Operation IceBridge, by texturally segmenting the high-resolution lidar scan of the snow freeboard and then matching texturally-similar areas. We find that the ratio of the snow freeboard F and snow depth D, applied to the segment mean F, is a better predictor of the true segment mean snow depth than just copying in D values from these texturally-similar segments. This is likely because the freeboard provides a physical constraint on the snow depth. But it may also be indicative of a sampling bias present in the radar snow depths. We find evidence of a sampling bias in the snow depth data from OIB, which gives an overall bias of around +2-4 cm for the mean snow depth. This is likely because thin snow depths are not resolved by the snow radar, consistent with findings

from Kwok and Maksym (2014). Our extrapolations for the snow depth have ∼22% error at 180 m scale by extrapolating from nearby floes. However, this error is only slightly larger (∼28%) if extrapolating from a completely different dataset. This suggests that there may be regional, or perhaps even generalizable relationships between surface texture and snow depth, although this needs to be evaluated for datasets from different regions and seasons.

Using this data, we show that the snow depth at 180 m scale can be predicted directly from the snow freeboard data using a convolutional neural network (ConvNet). The learned filters appear to correspond to standard textural analysis techniques, which suggests that there is a relationship between snow surface texture and snow depth. The error when applied to a different dataset from a different year is 20%, suggesting that there is, at least for the Weddell Sea, a connection between the texture of the snow surface and the snow depth. The 20% error (at 180 m scale) for our ConvNet may be irreducible as the snow depth data that it is being trained on itself has a similar error (∼22%). Predicting mean snow depths over a larger length scale (1.5 km) gives lower errors for the snow depth estimates (14%), which allows for a lower uncertainty in sea ice thickness estimates at the 1.5 km scale of ∼20%.

We find that the highest contribution to the sea ice thickness uncertainty, using our snow depth estimates from the ConvNet, is from the uncertainty in sea ice density. This suggests that estimating snow depth from sea ice surface texture alone may be sufficiently accurate to constrain SIT, and that improved treatment of sea ice density is now at least as important. However, this result needs further evaluation for other regions, such as the Bellingshausen/Amundsen sea, for which OIB snow depths are significantly deeper.

Ultimately, this work may provide insight into suitable length scales for snow depth analysis and suggestions of relevant metrics to predict snow depth for other high-resolution datasets, including, if the technique can be extended to linear topography data, ICESat-2.

Chapter 5

Regional and Interannual Variations

Abstract

The distribution of snow depth on Antarctic sea ice is critical to estimating the sea ice thickness distribution from laser altimetry data, such as from Operation IceBridge or ICESat-2. Satellite datasets like ICESat-2, have adequate temporal and geographical coverage to estimate trends and/or variability in sea ice thickness, which is crucial to understanding energy balances in the Antarctic. The morphology of the snow surface may influence how the snow is redistributed by wind, in particular around areas of deformed ice. Here, we use a convolutional neural network trained on Weddell Sea data to predict snow depths for other Operation IceBridge flights, both in the Weddell and Bellingshausen/Amundsen Seas. We compare these predictions to empirical linear fits of snow depth to the mean snow freeboard, and find that the convolutional neural network has better generalization, lower error, lower bias and more consistent errors with respect to mean snow freeboard and the proportion of surface deformation than empirical linear fits. This stays true over a range of length scales ranging from 0.2 to 25 km. We note that despite interannual/regional variations in sea ice conditions that may cause different distributions of snow freeboard, snow depth and surface deformation, the convolutional neural network has similarly low errors across these different conditions. Moreover, the convolutional neural network is able to resolve interannual/regional variability whereas linear fits cannot. These results suggest that

surface morphological information can be used to estimate snow depth, potentially also from 2-D remotely-sensed lidar data like ICESat-2, which can in turn be used to estimate sea ice thickness.

5.1 Objectives

Following from Sect. 1.1, the importance of resolving interannual and regional variability in sea ice thickness is important to identify whether any thickness trends may be, for example, co-occurring with the decreasing (increasing) trends in SIE in the Bellingshausen/Amundsen (Weddell) Seas. In Chap. 4, we showed that a ConvNet could be trained on one Weddell dataset to predict another year's dataset. In this chapter, we seek to further generalize this to other regions and years, with the ultimate goal of being able to characterize the large scale of variability (which likely exceeds any trends that may exist (e.g. Ludescher et al., 2019)), and the resolvable scale of this variability. Operation IceBridge flights (Sect. 2.2) have now ceased and in any case do not have sufficient spatial coverage to make generalized conclusions about any regional trends in sea ice. However, the recent launch of ICESat-2, which similarly uses a laser altimeter to measure surface elevation, means that relationships between snow surfaces and snow depth (and hence ice thickness, via Eq. 1.1) can be adapted to ICESat-2 (with some accounting for the different spatial resolution, as well as the lack of true 2D scanning). In particular, once ICESat-2 has collected data spanning a large enough time range, trends in sea ice thickness may be resolvable with the development of high-resolution, low-error methods for SIT estimation.

We therefore start by extending the results from Chap. 4 in order to create a ConvNet that can generalize well enough to other flights to resolve interannual/regional variability in snow depth. Using the same snow extrapolation method from Sect. 4.2.1, we extend the analysis of the OIB data to 10 more flights, with a key difference being that we now use flights from both the Weddell and Bell/Am Seas, for a total of 12 flights. These are presented in Sect. 5.2.1, with some discussion of inter-flight differences in Sect. 5.2.2. In this chapter, we again check the ability of the ConvNet

to generalize to other flights and compare this against a linear fit, but with our larger dataset, we can also test whether the ConvNet/linear fits can generalize between different regions and/or different years (Sect. 5.4). We investigate what factors allow the training set to generalize, and in what conditions the convolutional neural network does well/badly in predicting snow depth (Sect. 5.5.1). We examine the biases of the two methods with respect to snow freeboard and the proportion of surface deformation (Sect. 5.5.2-5.5.3). We then discuss the implications for resolving intra-flight, inter-regional and interannual variability (Sect. 5.5.4). We discuss the implications of these biases in snow depth for sea ice thickness estimates (Sect. 5.5.5). Lastly, we apply the ConvNet on a 2018 flight that has no processed snow depth data to demonstrate the viability of the ConvNet to pick out interannual variability and to compare with snow depths obtained from extrapolation (Sect. 5.5.6).

5.2 Data

The data used here is the lead-referenced lidar windows (180 m x 180 m, at 1 m resolution) from Operation IceBridge (e.g. Fig. 2-11), following Sect. 2.3.2, and the OIB snow radar returns processed following Kwok et al. (2011) (Sect. 2.2.2). Because the snow depths are sampled only along a line within each lidar window, they may not be representative of the lidar window's mean snow depth, and so the snow depths need to be extrapolated onto the lidar window (Sect. 5.3.1).

5.2.1 Summary of flight data

All surface elevation (F), snow depth (D) and F/D distributions for the flights used in this study are shown in Figs. 5-1 and 5-2. Reasons for their differences will be discussed in Sect. 5.2.2. The geography of the Antarctic peninsula between the Weddell and Bellingshausen Seas creates a barrier to both prevent sea ice advection and to allow for sea ice deformation. As a result, there is often thick multi-year ice along the eastern coast facing the Weddell Sea.

The flight tracks, ridging frequency/fraction and surface roughness (from Ad-

vanced Scatterometer (ASCAT) 5.3 GHz backscatter at 40^o incidence) are shown in Figs. 5-3 and 5-4. Rough surfaces scatter the microwaves more, and show up as 'brighter' in Figs. 5-3 and 5-4. Using the scatterometry data, the flights were partitioned into either four or five zones of approximate equal track distance and with similar surface roughness using the ASCAT data in Figs. 5-3 and 5-4, in order to test the performance of the ConvNet for different surface types in Sect. 5.4.

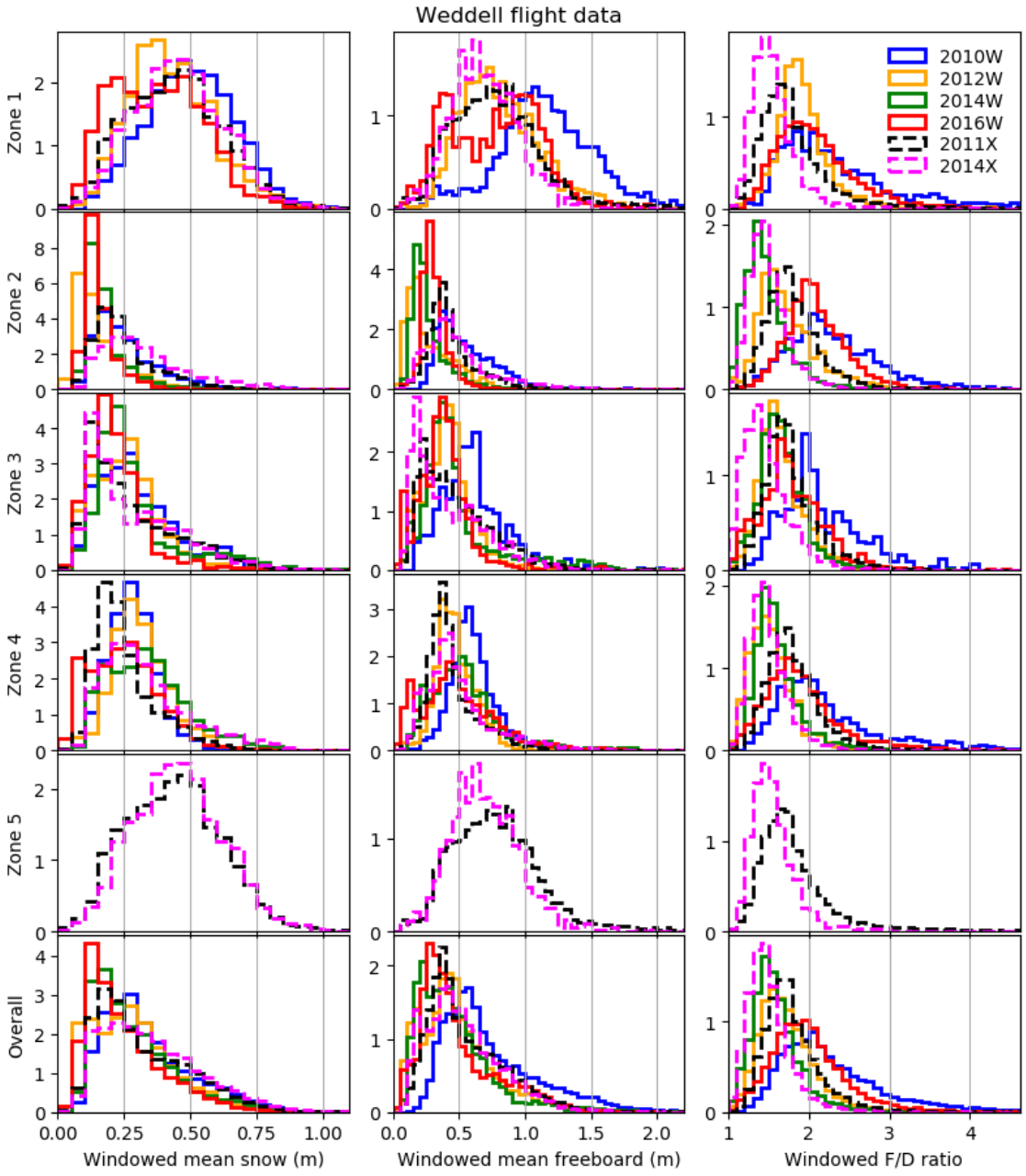

Figure 5-1: Snow (D, binned at 5 cm), snow freeboard (F, binned at 5 cm), and F/D (binned at 0.1) distributions, by zone, for the Weddell Sea flights used in this study. Note that 'W'- and 'X'-type flights (see Fig. 2-8 and Table 2.2) have different zones ('W' have 4 zones and 'X' have 5). The y-axes show probability density; all histograms are normalized. For zone information, see Fig. 5-3.

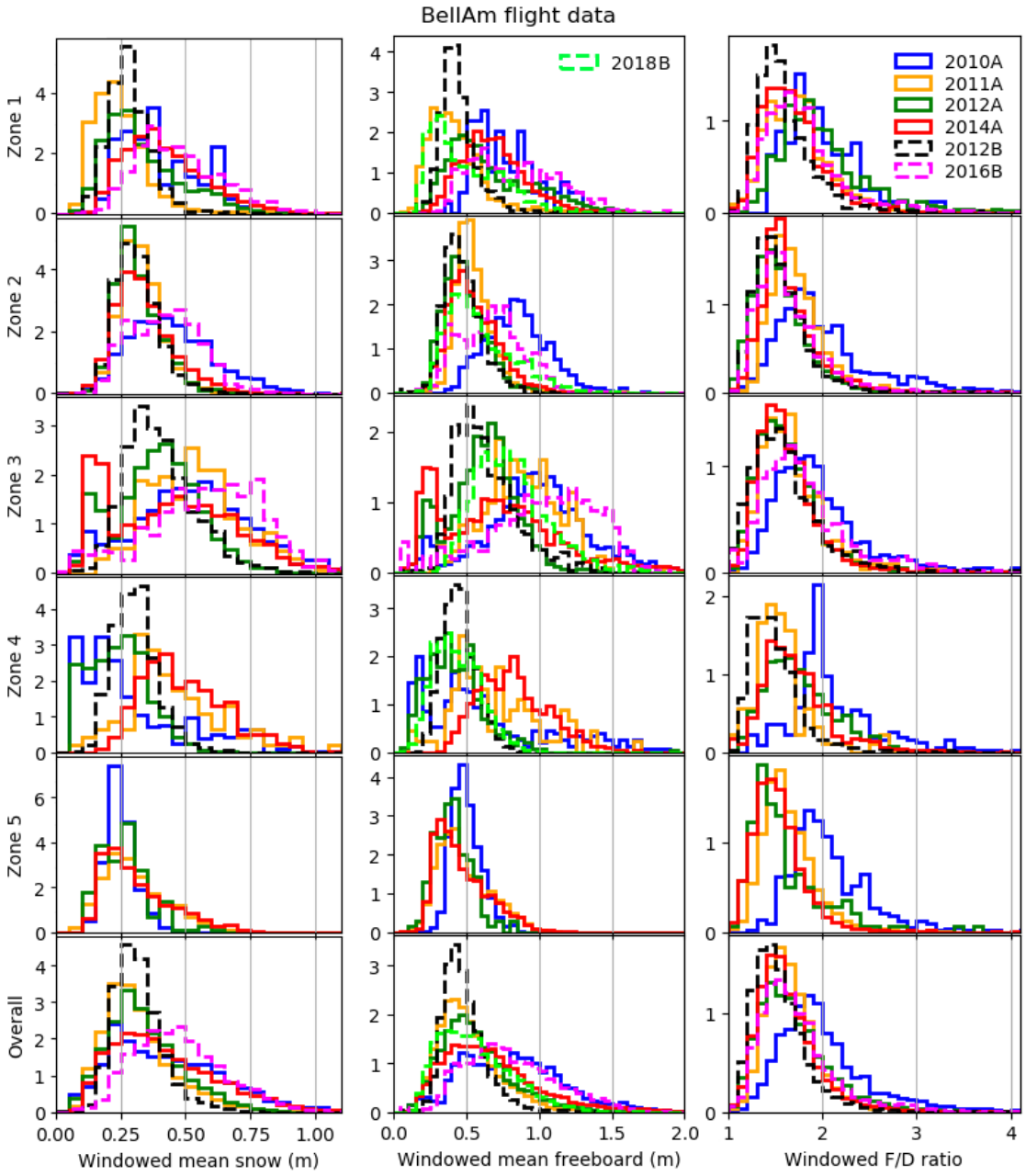

Figure 5-2: The same as Fig. 5-1 but for the Bell/Am data. The modal snow depths in the Bell/Am are generally higher than in the Weddell. 2018B does not have processed snow depths so only F is shown.

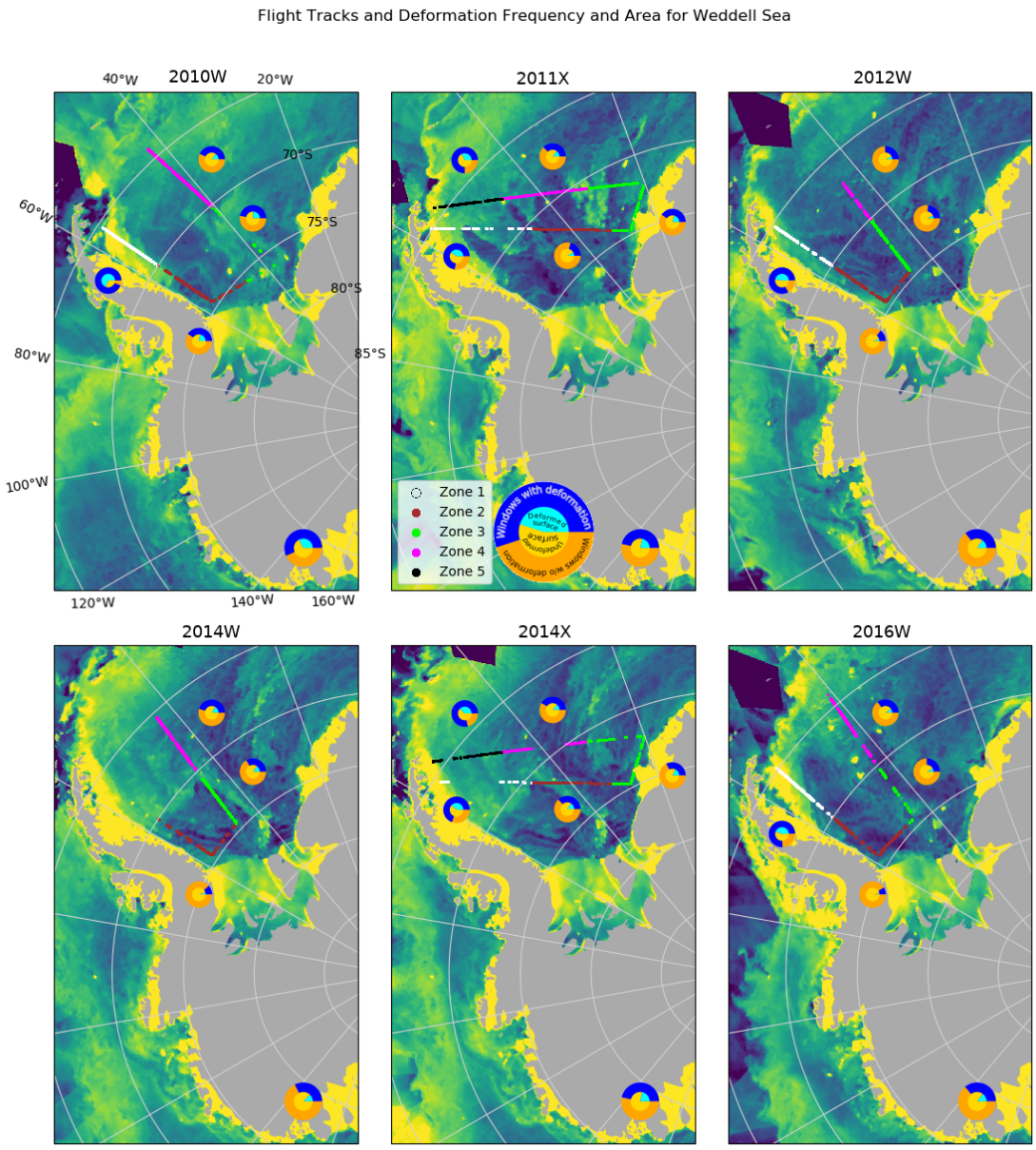

Figure 5-3: Flight tracks for the OIB Weddell Sea data used in this study, along with the deformation frequency (proportion of lidar windows with deformed segments, outer pie chart) and deformed surface area (proportion of lidar surface area that is deformed, inner pie chart). The flights have been sectioned into zones for later analysis. 'W'- and 'X'-type flights (see Fig. 2-8 and Table 2.2) have different zones ('W' have 4 zones and 'X' have 5). Overlaid is the Advanced Scatterometer (ASCAT) backscatter for that date; brighter = higher surface deformation (Lindsley and Long, 2016).

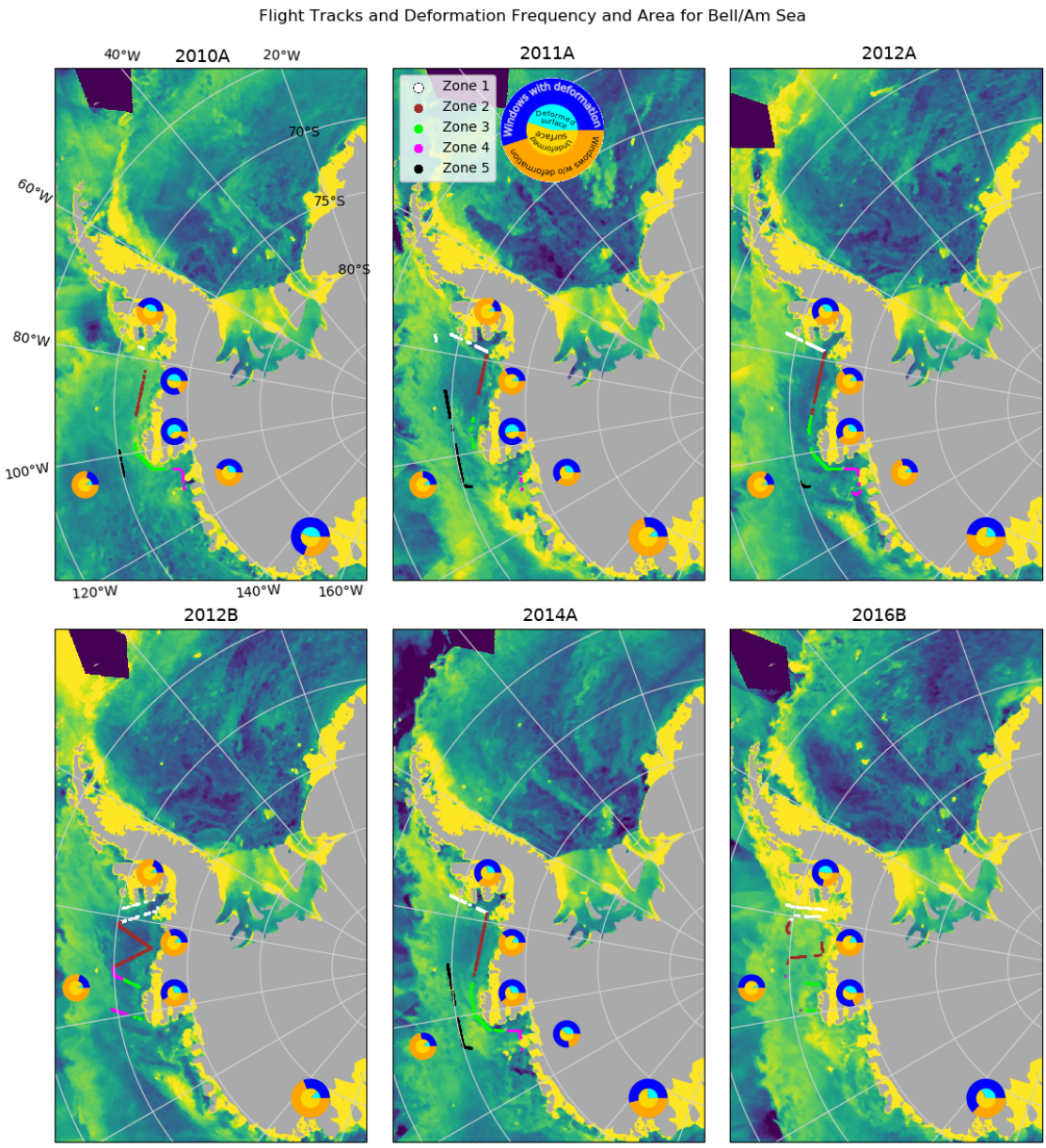

Figure 5-4: Same as Fig. 5-3 but for the OIB Bellingshausen/Amundsen Sea data used in this study. These flight tracks are sectioned into zones for later analysis. Note that 'A'- and 'B'-type flights (see Fig. 2-8 and Table 2.2) have different zones ('A' have 5 zones and 'B' have 4).

5.2.2 Geophysical differences between years

Due to the sampling differences between the flights, not only in the flight tracks but also in the number of successful snow radar and lidar returns, it is hard to directly use the differences in observed snow depth to infer any interannual/regional variations. As discussed in Sect. 1.1, there is considerable interannual variability in Antarctic SIE that may be greater than any observable trends (Ludescher et al., 2019). Looking at October data specifically, the Bell/Am SIE shows much larger variability than the Weddell data (Parkinson, 2019): this is also reflected in our snow depth distributions, which tend to be very similar between Weddell datasets (Fig. 5-1), whereas the Bell/Am snow depths show more variability between years (Fig. 5-2). It is important to see if our prediction method for snow depth is able to resolve these differences, whether large or small, which may arise from multiple causes. For example, for either the Bell/Am or Weddell, mean F may vary anywhere from 14-30 cm and mean D anywhere from 0-15 cm between different years. However, there is generally more variability for the Bell/Am data: the median interannual variability for the mean F (D) of any two random years is 6.9 cm (3.8 cm) for the Weddell and 15 cm (8.8 cm) for the Bell/Am.

Figs. 5-4 and 5-3 show scatterometer data for each of the flights with the flight tracks overlaid. Scatterometer data is an approximation for the current state of deformation, as rough surfaces give more isotropic backscatter responses (although 'roughness' can also be affected by, say, varying surface porosity) and hence show up as brighter in Figs. 5-4 and 5-3. In addition, we use the NSIDC Daily Sea Ice Motion Vectors (Tschudi et al., 2016) to make some observations about possible causes of the differences in surface deformation (Fig. 5-5). For each year, the two-month average drift velocity before the corresponding flight date, and also show the two-month RMS average speed, so that locations with high drift speeds but highly varying directions (leading to a low two-month average drift velocity) can be distinguished from quiescent zones. For years like 2012 where the Weddell and Bell/Am flights are 4 weeks apart, the two-month averages for each region are calculated separately and

stitched together. The observations in this section are qualitative and are intended as plausible descriptions of interannual/regional variability that we can see if the ConvNet can identify. Furthermore, seeing the variation in surface types may be helpful for constructing a diverse training set. For all zones in all flights, there is generally more variability in F and F/D than in D, so it may be helpful to see how/if certain types of topography (i.e. distributions of F) affect the quality of the ConvNet performance.

In general, the transport in the Weddell Sea is dominated by the Weddell Gyre, leading to cyclonic (clockwise) motion, i.e. northward motion along the peninsula, westward motion along the south near the coast, and eastward motion near the northern latitudes. There is considerable deformation along the coast due to compression, and the ice in the northwest Weddell (near the peninsula tip) can often survive the summer melt (Comiso and Nishio, 2008). The southwest Weddell, in contrast, is an area of net ice production, which is exported northwards (Kimura and Wakatsuchi, 2011). The Weddell Sea has generally shown an increasing decadal trend in SIE from 1978-2010, generally due to increasing southerly (northward) winds (Holland, 2014). However, record lows in SIE since 2016, partly attributed to intense storms and the reappearance of the Maud Rise polynya in 2016-17, have made this trend statistically insignificant (Ludescher et al., 2019) and possibly even negative (Turner et al., 2020). There is some evidence of increasing snowfall in the Weddell Sea, though there is only a weak correlation between snowfall and areas of increased SIE (Turner et al., 2015).

In contrast, the Bell/Am sector is the only one to experience a consistent decreasing decadal trend (Ludescher et al., 2019). Stammerjohn et al. (2012) found that this could be plausibly linked to enhanced poleward winds accelerating ice retreat and delaying ice advance. In particular, between 1979-2010, sea ice in the Bell/Am sector started retreating 38 days earlier and advancing 60 days later, leading to a 3-month longer ice-free season. It is also known that the ocean temperature has increased in this sector (Schmidtko et al., 2014). These factors have likely contributed to the decreasing amount of multi-year ice that was once found embayed in Pine Island Bay (Stammerjohn et al., 2015). The geographical boundary of the Antarctic peninsula

separating the Bellingshausen and Weddell Seas prevents sea ice advection and may also contribute to the anomalies that occur in the Bell/Am Seas (Van Den Broeke, 2000).

2010

For 2010A, we see a lot of sea ice compression from westerly and northerly winds from mid-August onward, leading to not much northward export until early October (Fig. 5-5). The overall two-month mean shows strong compression near 90^{o}W 72^{o}S in Zone 2 in particular. This may have caused this flight to have much more deformation, in particular in Zones 2-3, than other years. These increased northerly winds likely led to higher precipitation, and hence higher snow depths overall. All zones (in particular 2, 3 and 5) for 2010A in Fig. 5-2 have above-average F and F/D, consistent with high deformation.

For 2010W, there is modest circulation in the two-month average (Fig. 5-5), with particularly low drift in Zone 1-2. Zone 1 is the only zone for the entire Weddell dataset that has more than 50% deformation by area, which is also visible in the scatterometry data. In 2010, there were no cyclones between July and November in the Weddell Sea (Phillips, 2020), which may have led to more quiescent conditions and hence more thermodynamic thickening of ice (i.e. high D, higher F, and higher F/D, in particular in Zone 1).

2011

For 2011A, the drift motion flips between compression from August-September, then strong eastward advection until October, then mild compression until the flight date. Zone 1 and 2 have strong advection to the northeast. Overall, this may have led to below-average deformation (high ice export) in Zones 1-2 and above-average deformation in Zone 3. Consistent with this, Fig. 5-2 shows below-average F in Zone 1 and above-average F in Zone 3 for 2011A. Zones 4 and 5 are fairly average compared to other years.

For 2011X (note this has different zones to 2010W), the two-month average shows

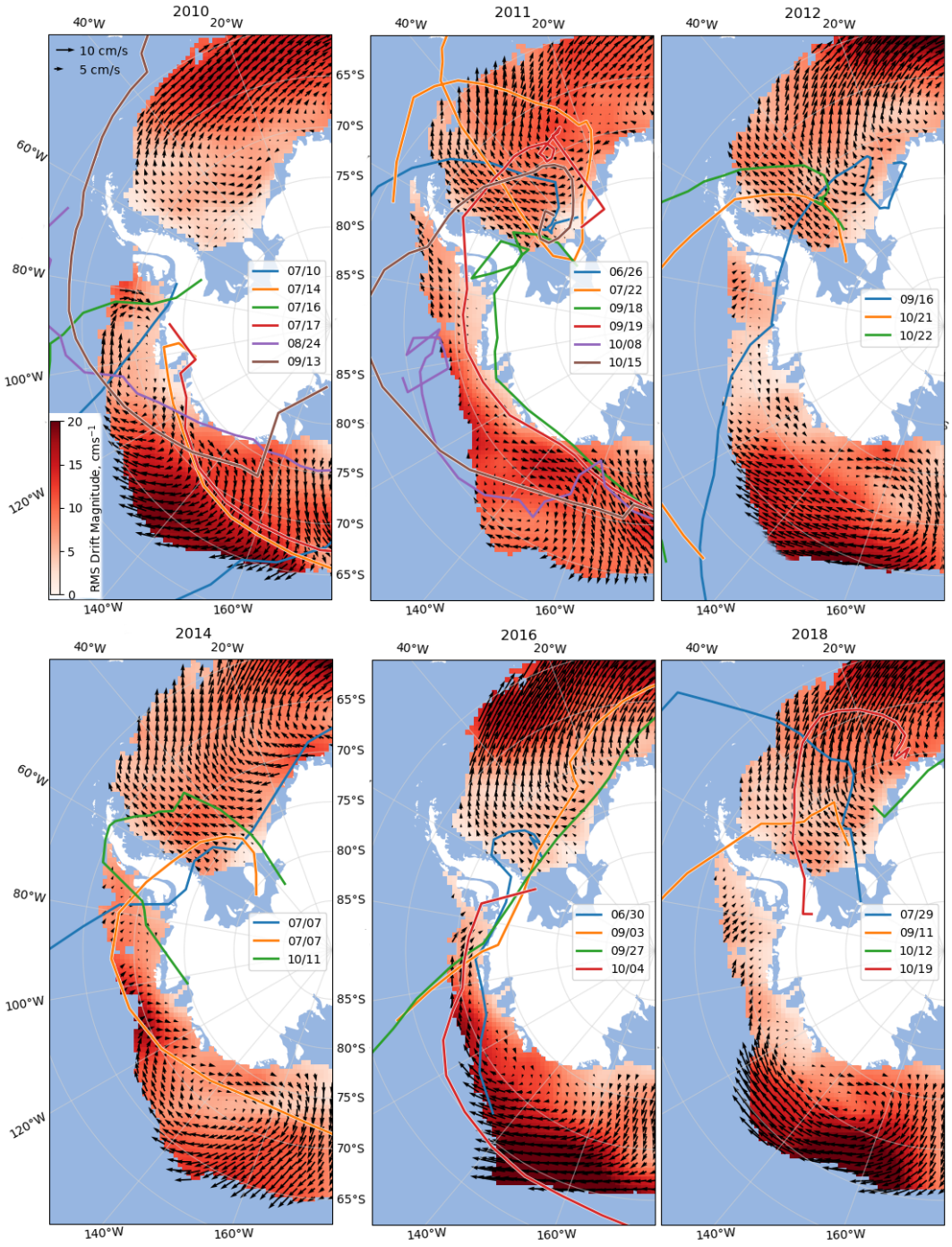

Figure 5-5: Sea ice drift motion for years with OIB flight data. The arrows represent the two-month mean drift velocity and the colormap represents the two-month RMS drift magnitudes, for the two months preceding each flight. All storm tracks between July and October that have a recorded depth of at least 12 hPa and a duration of at least 5 days are shown. Full details are given in the main text. Data are taken from Tschudi et al. (2016) and Phillips (2020).

strong north(east)erly export (Fig. 5-5), with stronger advection at higher latitudes. This likely led to deformation that was then advected to the north(east), corresponding to zones 1 and 5 in 2011X. This is similar to the other flight with this flight path (2014X). 2011X tends to have slightly higher F/D, in particular in Zones 2-4. This may be because there were a lot more storms in the Weddell in 2011 which caused high RMS drift speeds with alternating directions, in particular in Zone 3, which led to increased deformation.

2012

For 2012A, as with 2011A, the drift flips between northward export and compression along the peninsula near Zone 1. There is an unusual lack of northerly drift in Zone 3 (Fig. 5-5), with particularly low RMS drift speeds there. This likely contributes to a much lower deformation frequency there than other years. Zone 4 also had relatively quiescent conditions that may contribute to its low deformation frequency. Interestingly, in 2012B Zone 1, there is much less deformation than Zone 1 in 2012A, perhaps due to the strong advection along the peninsula in the week before the flight date which may have caused the offshore areas to be relatively undeformed. This is consistent with Fig. 5-1, where 2012B has lower F/D, consistent with a lack of deformation, in Zone 1 compared to 2012A. Other zones are consistent with other years.

For 2012W, there is strong overall northward export in Zones 1-2, unlike for 2010W. This is suggestive of high ice production, and hence less deformation (low F/D) in Zone 2. Strong compressive motion in early October (Fig. 5-5) along the peninsula may have contributed to high deformation in Zone 1, though the higher northward export, enhanced by two cyclones in late October, leads to the deformed ice that normally collects near the tip of the peninsula being advected to lower latitudes (visible around 40^{o}W 60^{o}S in Fig. 5-3). This consistent northward motion may also have caused Zones 3 and 4 to be relatively undeformed.

2014

For 2014A, there is some strong compression into/along the peninsula near Zone 1, along with strong ice export to the north-east for the entire month of September, leading to a small net velocity but high RMS drift speed in Fig. 5-5. As a result, unlike other years such as 2010 and 2011, there is more deformation in Zone 1 than in Zone 2. The strong compression into/around Thurston Island (Zone 3) also contributes to high deformation there, with less deformation off-shore (Zone 5). This is a fairly typical year overall, with the distributions of F, D and F/D for all zones are fairly typical compared to other 'A' flights (Fig. 5-2).

In contrast, 2014W and 2014X in the Weddell have slightly lower F and F/D values. In 2014W, there is consistent moderate-to-strong northward export in the two months preceding the flight (Fig. 5-5), which was likely enhanced by a cyclone passing through zone 1 in the week immediately preceding the flight. This likely reduced the amount of deformed ice in Zone 1; coupled with the consistent northward export in Zone 2, these would have contributed to low deformation rates and low F/D. In Fig. 5-1, there is particularly low F values for Zone 2, suggesting there was a lot of new, thin, level ice. The 2014X flight is largely similar to 2011X, with slightly more quiescent conditions with low RMS drift speeds in Zone 3 (Fig. 5-5). Without the influx of ice from the east Weddell that occurred in 2011X, there may have been less deformation in Zone 3.

2016

2016 is notable for having the lowest (annual) SIE on record, with multiple factors posited as causes, among them an anomalously negative southern annular mode causing rapid sea ice retreat, a strong El-Niño event causing warm surface waters and strong northerly winds hampering new ice production/export (Stuecker et al., 2017; Turner et al., 2017). However, during October, the SIE in the Weddell was actually above-average, although in the Bell/Am it was already near record lows (Phillips, 2020).

For 2016B, these strong northerly winds, potentially enhanced by successive storms that headed directly towards Ellsworth Land (73°S 88°W) in September and October, likely caused above-average deformation via compression along Zones 1-3, with less export to the northeast than other years (Fig. 5-5). These cyclones would have compressed the Bell/Am ice along the coast and contributed to significantly below-average SIE as well (Phillips, 2020).

For 2016W, there is mixed circulation, but with a general north-eastward motion, in particular in the east Weddell. Fig. 5-5 shows strong north-eastward advection in Zones 3 and 4, which may have caused low deformation rates there. Zone 2 had a fairly consistent northward drift that may have similarly led to low deformation and more new ice production (Fig. 5-5). This is consistent with the low F and low D for Zone 2, suggesting relatively young, level ice with a thin snow cover, matching what we see in the scatterometry (Fig. 5-1). In the Weddell sector, there were two zonally-traveling cyclones in September, whose circulation would have caused more northward drift of sea ice (in particular in Zone 2, hence creating more new, thin ice, but also causing more deformation in Zone 1). The increased northward drift may have caused an above-average October SIE (that would become record-low by December). These cyclones passed over a large stretch of land before reaching the ice in the south Weddell, and so likely brought drier conditions with low snowfall. This may account for the below-average snow depths in Zones 1 and 2, which meant a high F/D.

2018

For 2018B, there is a fairly standard overall northeasterly advection, with some compression into the peninsula near Zone 1 (Fig. 5-5). Similar to 2014, there is strong advection from the Ross Sea. This ice deforms as it advects eastward, and is visible as a plume of deformed ice around 73°S 120°W in Fig. 5-16. In the immediate weeks before the flight date, there were relatively quiescent conditions, in particular in Zone 1, which may have led to low deformation there. 2018B has relatively low F compared to other 'B' flights, most prominently for Zone 1, which is suggestive of young, thin

ice without much surface deformation (Fig. 5-2).

Summary

The distribution of F and D can vary due to different geophysical forcings, for different regions or seasons. For both regions, northerly winds will tend to increase the amount of deformation (and hence F/D) by compressing the ice along the coast. Additionally, northerly winds are more likely to be moist, as they pass over considerable amounts of ocean before reaching the ice, and hence may deposit more snow. In contrast, southerly winds (originating from the continent) will export the ice, and generally allow for more new (thin) ice formation; however, this air is more likely to be dry and hence probably will not cause much snowfall. Cyclones can also cause considerable export or compression of sea ice depending on the storm track.

In general, for the Weddell 'W' flights, Zone 1 is highly deformed and generally has both thick snow and thick ice, due to the presence of multiyear ice; Zone 2 largely consists of new ice, which is thin and level with relatively low snow depths; Zone 3 and 4 have a mixture of surface types depending on the gyre-driven advection. For the 'X' type flights, Zones 1 and 5 have generally more deformation as these zones span the general northeastward drift of deformed ice from the NW Weddell. Other zones have a mix of deformed and undeformed surfaces as the flight path goes diagonally across the Weddell Gyre. 2010 was an anomalous year, possibly due to the absence of cyclones, which allowed the ice along the coast to thicken (high F and F/D). 2016 was also an anomalous year, possibly due to multiple cyclones successively passing over the Zone 2 area, bringing dry air from the continent (i.e. low snowfall = low D = high F/D) and leading to higher deformation in Zone 1 (high F/D).

For the Bell/Am 'A' flights, the coastal areas (Zones 1-4) will have large amounts of deformation and also higher snow depths, whereas Zone 5 (off-shore) will have less deformation and also lower snow depths. The 'B' type flights, due to their zig-zag sampling pattern, typically sample a mixture of near-shore deformed ice and off-shore undeformed ice. 2010 was an anomalous year, possibly due to a higher-than-usual number of cyclones near Zones 2/3 in July, which allowed for a longer

duration of thermodynamic growth (hence increased F and F/D). 2016 was also an anomalous year with high amounts of deformation both near-shore and off-shore, possibly due to multiple cyclones that headed directly north into the coast near Zone 2/3 in October/November, possibly with high precipitation, that led to a near-record low SIE and increased compression (hence increased F and D) along the coast.

This means that high F/D may be associated with either (or both) level surfaces, which implies old, thick ice which may have above-average snow depths, but are so thick that the $F/D \gg 1$, or deformed surfaces, which implies high deformation. Notably, the former would not cause a record low SIE, whereas the latter would. However, these would give rise to different surface morphology, so we want to see what prediction methods, if any, could distinguish these.

5.3 Methods

5.3.1 Extrapolation of Snow Depth

The extrapolation of the snow depth from a series of collinear measurements onto the OIB lidar scans is done the same way as in Chap. 4. Full details may be found in Sects. 4.2.1 and 4.3.1. Similar to Sect. 4.4.1, we find that our extrapolation bias is typically < 1 cm and the sampling bias of the snow radar is typically +4-8 cm, in line with findings from Kwok and Maksym (2014). To generate the snow depths for the ConvNet, we extrapolate each flight's snow depths using the snow depths that are recorded for that flight only. The average error when self-extrapolating is 16-19%; when using another randomly-selected flight, the average extrapolation error is 19-25%. This suggests there are minor differences between the relationship between surface texture and snow depth for different flights (at least with respect to the metrics used in the snow extrapolation, namely image entropy and the mean, L-kurtosis and standard deviation of F).

5.3.2 Prediction Methods

ConvNet

The architecture for the ConvNet is similar to the one in Chap. 4, with four convolutional layers of size 10, 12, 12 and 14 respectively (each with a stride of 2) with dropout ($p = 0.5$) and an activation function of SELU (Klambauer et al., 2017) between each of these. The final fully-connected layers were of size 128 and 32 respectively, again with a SELU activation function in between. The inputs were 180 x 180 lidar windows from Sect. 2.3.2 using the extrapolated windowed mean snow depth from Sect. 4.3.1 as the ground truth. A slightly different optimizer (AdamW) was used instead of Adam that accounts for the weight decay correctly (Kingma and Ba, 2014; Loshchilov and Hutter, 2018). The key difference with Mei and Maksym (2020) is that the input windows are normalized to have values between 0 and 1. This leads to the window mean snow freeboard value being lost; as the mean F is so highly correlated with the snow depth, we add this mean value in to the 128-length fully-connected layer (so that it becomes a 129-length layer). The SELU activation function is also applied after this first FC layer. The second FC layer thus allows the network to learn non-linearities with the mean (thus increasing the model complexity vs. the linear fit). The goal of the normalization is to allow the convolutional layers to learn structural features, instead of just learning to predict the mean F. We tried both mean squared error (MSE) and mean absolute percentage error (MAPE) as loss functions and found slightly better performance with MAPE loss. Hence, subsequent mentions of the ConvNet in this chapter refer to these above parameters.

The training set was taken as 80% of the 2010W and 2012W datasets, randomly sampled, and the validation set was the remaining 20% of these data. More details on how this training set was chosen are in Sect. 5.5.1. The trained model was tested on the remaining datasets (2010A, 2011A, 2012A, 2012B, 2014A, 2016B from the Bell/Am Seas, and 2011X, 2014W, 2014X and 2016W datasets from the Weddell Sea, for a total of 10 test sets). We also make predictions for the 2018B dataset which has only lidar data with no (processed) snow data.

Linear fit

The linear fit is an empirical ordinary least squares linear regression based off the method of Özsoy-Çiçek et al. (2013). It is a simple ordinary least-squares linear regression of F vs. D, using the same training set as the ConvNet, and evaluated on the same validation and test sets as the ConvNet. We do not use the empirical relationships from Özsoy-Çiçek et al. (2013) directly due to the noted sampling bias from drilling data (e.g. Kwok and Maksym, 2014; Williams et al., 2015). In particular, there is little drilling data in spring, where there may be a lot of heavily deformed ice. Such deformed ice comprises a substantial portion of the OIB data. Furthermore, there are rarely expeditions to the west Weddell due to the difficulty of navigating heavily deformed ice in spring. For sake of comparison, our best fit regression ($D = 0.38F + 0.07$) is very different from theirs ($D = 0.88F + 0.01$) for the corresponding (Western) Weddell region.

5.4 Results

5.4.1 ConvNet and linear model predictions for snow depth

Full results for the ConvNet and linear model fitted on the 2010W and 2012W flights and tested on the other flights are shown in Figs. 5-6 and 5-7 and summary statistics given in Tables 5.1 and 5.2. In general, we find that the ConvNet is only marginally better than the linear fit in terms of MRE, but much better in terms of the K-L divergence. In Figs. 5-6 and 5-7, we therefore only show the K-L divergences. We find that the linear fit is generally close to (slightly worse than) the ConvNet in predicting the overall mean snow depth, but fails to capture the distribution. As mentioned in Sect. 4.3.2, knowing the overall distribution is necessary to model processes with nonlinear relationships to ice thickness, for which (by definition) knowing just the mean ice thickness is not enough.

In general, the ConvNet out-performs the linear model in all metrics, although certain zones from flights may have either prediction only marginally better than the

other (e.g. for 2014W, the overall MRE is only 0.6% higher for the ConvNet). In particular, for almost all zones in all flights, the K-L divergence D_{KL}(True||Prediction) of the ConvNet fit is lower than the linear fit, indicating a closer prediction to the true distribution. The ConvNet has very similar performance for the Bell/Am test sets as the Weddell test sets, despite being trained on Weddell data only. Further discussion of ConvNet performance is in Sect. 5.5.

Overall, the MREs of the two models are fairly similar, with the ConvNet typically out-performing the linear model by a few percent. This suggests that the per-window MRE may not be as good of a metric to measure model performance, as it does not give any indication whether the overall shape of the distribution is accurate.

The variability in our predictions closely matches that of the true snow depth distribution. From Sect. 5.2.2, the variability in the per-flight mean snow depths for a given region was as much as 15 cm in the Bell/Am (standard deviation: 6.0 cm) and as much as 10 cm in the Weddell (standard deviation: 3.6 cm) between different years. Our ConvNet predicts a very similar variability of 16 cm in the Bell/Am (standard deviation: 5.7 cm) and 10 cm in the Weddell (standard deviation: 3.1 cm).

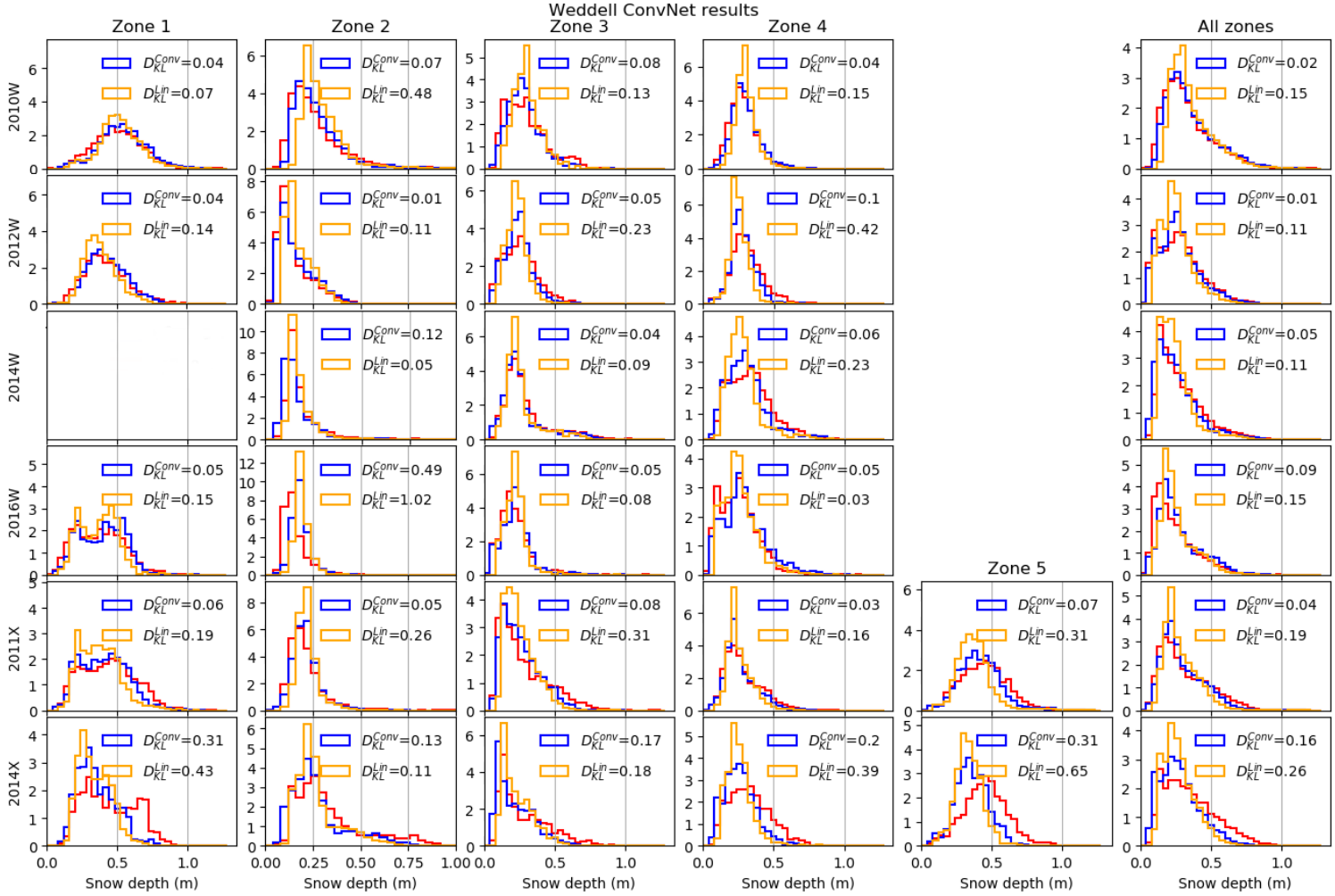

Figure 5-6: Results for applying the ConvNet (trained on 2010W and 2012W, here the top two rows) tested on the remaining Weddell datasets (blue). These are compared to an empirical linear fit (orange) using the mean of the same windows as the training set. The true distribution is shown in red. Also reported is the forward K-L divergence D_{KL}(True||Prediction)

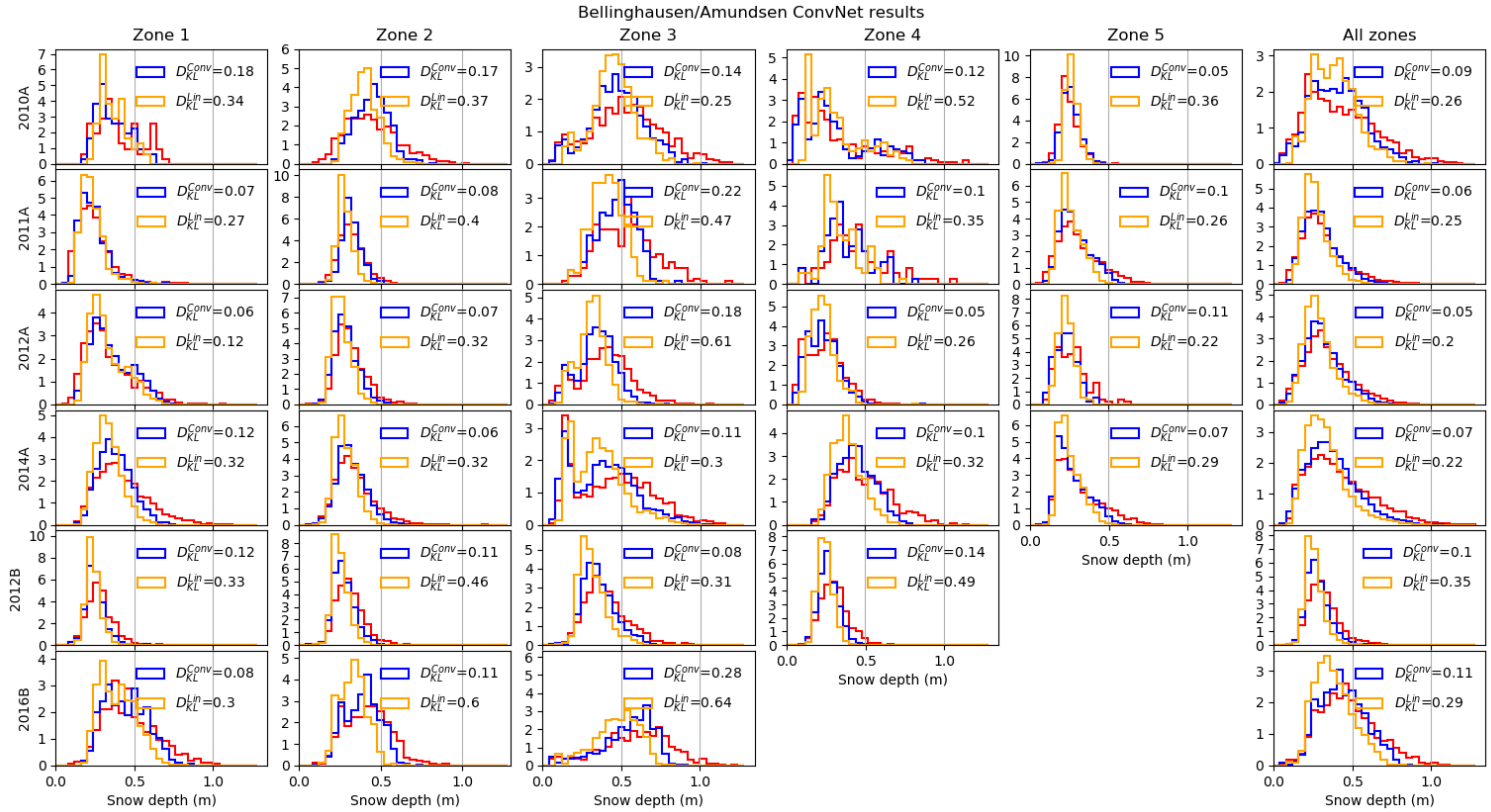

Figure 5-7: The same as in Fig. 5-6 but with the Bell/Am datasets as the test sets. Note that the training set is still 2010W and 2012W. 2016B Zone 4 has been excluded due to lack of data.

Table 5.1: Results for applying the ConvNet and empirical linear regression, both fitted on 2010W and 2012W, on the Weddell test sets. The forward K-L divergence D_{KL}(True||Prediction), the mean relative error (%), the mean residual (%) and the bias in predicting the overall zone mean (%) are also shown. N/A indicates there were insufficient data in the zone.

	Zone	K-L Div.		MRE (%)		Mean res. (%)		Overall res. (%)	
		Conv	Lin	Conv	Lin	Conv	Lin	Conv	Lin
	Train	0.008	0.045	18.6	25.6	-6.2	-10.7	-1.6	0.0
	Val.	0.016	0.046	20.5	25.5	-7.0	-10.8	-2.3	-0.7
2011X	1	0.061	0.195	17.6	21.5	3.2	8.2	7.2	14.5
	2	0.053	0.257	15.0	19.8	-4.7	-11.6	-0.0	-3.6
	3	0.080	0.307	14.8	22.7	1.0	-4.9	7.7	8.5
	4	0.033	0.159	14.5	18.3	2.5	1.1	6.2	8.8
	5	0.066	0.306	17.9	22.2	1.4	8.7	5.2	14.6
	All	0.036	0.186	15.7	20.8	0.7	-0.5	5.7	9.7
2014W	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	2	0.120	0.049	17.2	11.6	12.8	-2.2	13.0	3.5
	3	0.043	0.086	14.6	13.9	7.2	4.3	8.5	10.3
	4	0.064	0.229	15.2	18.0	8.2	11.3	9.2	16.0
	All	0.046	0.107	15.7	15.1	9.4	5.9	9.9	12.4
2014X	1	0.313	0.426	21.1	22.3	17.7	18.9	20.7	23.4
	2	0.127	0.114	19.3	17.7	16.5	13.3	18.3	19.7
	3	0.170	0.177	22.6	17.8	21.0	8.2	22.8	17.8
	4	0.201	0.392	21.6	21.6	19.6	18.0	20.8	23.2
	5	0.314	0.651	21.3	24.8	18.8	22.1	20.5	25.7
	All	0.157	0.258	21.1	20.1	18.8	14.9	20.4	21.7
2016W	1	0.049	0.145	21.0	20.2	-11.0	-5.6	-5.8	2.2
	2	0.494	1.022	29.4	34.9	-26.7	-32.5	-20.2	-23.7
	3	0.052	0.084	17.2	21.3	-8.4	-12.2	-5.4	-2.4
	4	0.051	0.032	23.4	24.6	-19.3	-15.0	-15.4	-4.3
	All	0.092	0.145	23.9	26.2	-18.2	-17.7	-12.1	-5.6
	All	0.028	0.147	18.2	20.7	2.2	0.4	7.0	10.5

Table 5.2: The same as Table 5.1 but for the Bell/Am test sets.

	Zone	K-L Div.		MRE (%)		Mean res. (%)		Overall res. (%)	
		Conv	Lin	Conv	Lin	Conv	Lin	Conv	Lin
	Train	0.008	0.045	18.6	25.6	-6.2	-10.7	-1.6	0.0
	Val.	0.016	0.046	20.5	25.5	-7.0	-10.8	-2.3	-0.7
2010A	1	0.177	0.339	17.4	16.6	5.3	3.5	9.1	8.7
	2	0.170	0.373	22.5	22.5	-9.0	2.1	-0.5	10.3
	3	0.138	0.254	20.4	24.0	4.2	5.1	10.8	14.9
	4	0.116	0.519	20.3	33.7	-6.5	-22.2	1.3	-0.5
	5	0.053	0.357	18.2	16.6	-7.6	-10.2	-4.6	-6.3
	All	0.088	0.261	20.5	23.0	-2.7	-1.4	5.2	10.0
2011A	1	0.067	0.271	17.3	17.3	-2.2	-1.3	2.8	5.5
	2	0.075	0.398	11.9	15.5	-1.7	10.4	0.7	13.1
	3	0.217	0.469	17.8	24.5	8.2	19.3	12.4	23.5
	4	0.103	0.353	17.2	23.6	12.1	20.5	14.8	24.5
	5	0.104	0.264	15.0	18.9	0.4	8.7	5.5	15.5
	All	0.064	0.253	14.5	18.0	0.8	8.4	5.4	14.9
2012A	1	0.057	0.123	21.6	20.6	-12.4	-5.6	-5.4	2.6
	2	0.072	0.321	15.0	19.3	7.6	15.7	9.7	18.4
	3	0.178	0.613	17.3	23.6	8.1	13.8	13.1	21.4
	4	0.055	0.260	16.9	24.4	-0.5	-6.5	5.2	5.4
	5	0.111	0.218	20.9	20.6	9.6	7.3	13.5	13.6
	All	0.054	0.201	17.6	21.4	2.9	7.7	7.5	14.9
2012B	1	0.118	0.330	15.0	18.0	8.3	10.7	11.1	14.6
	2	0.113	0.464	15.0	19.2	9.8	15.5	11.7	18.5
	3	0.077	0.313	17.1	22.7	7.6	17.6	10.9	21.5
	4	0.144	0.489	15.8	20.5	9.8	16.8	12.4	20.3
	All	0.100	0.355	15.5	19.8	9.2	15.4	11.6	18.9
2014A	1	0.115	0.321	16.0	20.4	6.0	15.8	10.6	20.7
	2	0.056	0.323	14.5	18.6	2.6	13.1	6.0	17.2
	3	0.109	0.297	17.9	22.5	9.7	12.8	13.3	21.1
	4	0.099	0.319	17.0	22.2	6.2	17.3	10.2	21.3
	5	0.075	0.291	16.4	19.8	9.7	13.6	13.4	19.9
	All	0.068	0.225	16.7	21.2	7.4	14.7	11.4	20.8
2016B	1	0.084	0.296	17.0	20.8	2.4	12.7	6.2	16.6
	2	0.115	0.596	15.6	21.6	4.8	17.8	8.0	21.3
	3	0.275	0.642	18.7	24.5	1.9	13.0	7.5	20.0
	4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	All	0.107	0.285	16.7	21.9	3.4	15.0	7.3	19.4
	All	0.032	0.160	17.4	20.6	3.0	5.1	7.8	13.9

5.5 Discussion

5.5.1 Choosing a good training set

The ConvNet presented in Chap. 4, which was trained on data from 2010W, only had good generalization to the 2016W dataset, and to a lesser extent 2011X (Fig. 5-8). When applied to another Weddell dataset (for example, 2014W), the ConvNet in fact generalized worse than the linear fit. This was also the case when applied to a Bell/Am dataset (Fig. 5-8). However, we found that adding an additional flight (2012W) to the training set improved the generalization considerably. Reasons for this are discussed below. Our goal here is to create a minimal training set that spans many different surface types, because this allows us to draw conclusions about regional similarities (or lack thereof) between the Weddell and Bell/Am datasets. We could of course get good results from cross-validation and training on 11 datasets at a time and keep 1 as a validation set (i.e. leave-one-out cross-validation), though this would considerably increase training time, and would lose interpretability.

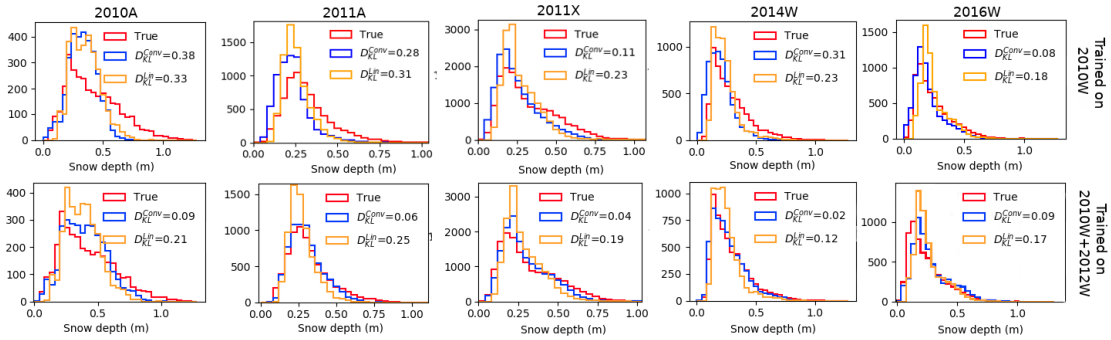

Figure 5-8: Applying the ConvNet previously shown in Chap. 4 to 2010A, 2011A, 2011X, 2014W and 2016W (top panels), as compared to a ConvNet trained on 2010W and 2012W, applied to the same test sets (bottom panels). The K-L divergences from the true snow depth distribution are also reported. All y-axes are the number of windows for that snow depth bin.

From Fig. 5-1, the 2010W flight has higher F/D ratios on average than all other flights, with 2016W close behind. Fig. 5-1 also shows that the 2010W distribution of F is the highest on average, with all other flights having similar distribution; similarly, the D distributions are all fairly similar. 2016W and 2014W cannot be

easily distinguished by their F and D, as both have similar distributions, with 2016W having slightly higher F values. However, they can be distinguished by their F/D, as the 2016W F skew a little higher and presumably correspond to low D and hence give rise to much higher F/D. This suggests that the F/D distribution can offer some insight into the ConvNet prediction success. Looking at a zone-by-zone breakdown for the Weddell predictions, the goodness of fit seems to depend most on Zone 2 (Fig. 5-9). In particular, the 2016W prediction using the 2010W-trained ConvNet is only successful because of its success in predicting zone 2.

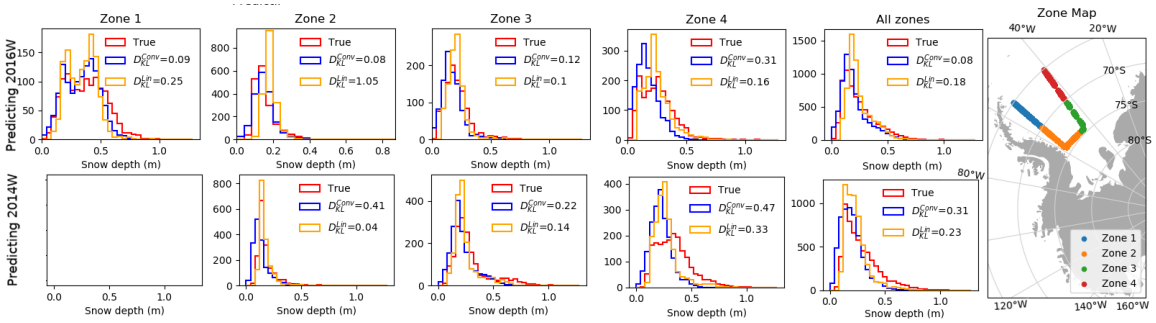

Figure 5-9: The zone-by-zone prediction for 2016W (above) and 2014W (below) using the 2010W-trained ConvNet. Note that Zone 1 of 2014W has no data, as was the case is Fig. 5-6. The snow distribution tends to be underpredicted by the ConvNet, suggesting that the F/D is over-estimated. Here, the linear fit outperforms the ConvNet (trained on 2010W only) in almost all zones. This is the opposite of Fig. 5-6, using a ConvNet trained on 2010W and 2012W.

Fig. 5-3 shows that Zone 2 is typically associated with undeformed surfaces. If the F/D and F distributions are divided based of those with and without surface deformation (Fig. 5-10), then the distinctiveness of 2010W is much more apparent. In particular, the F and F/D for level 2010W windows stands out - this suggests that the level ice in 2010W, with higher F but the same snow depths, had more ice in the snow freeboard (i.e. the snow was on thicker level ice). The deformed surfaces in 2010W also have higher F on average, so this suggests that the whole flight had similar snow depths on thicker ice (hence higher F) than other years. However, as deformed surfaces typically have higher F/D (i.e. less snow within the snow surface), this means that the effect on snow depth predictions for deformed surfaces is less than

for level surfaces.

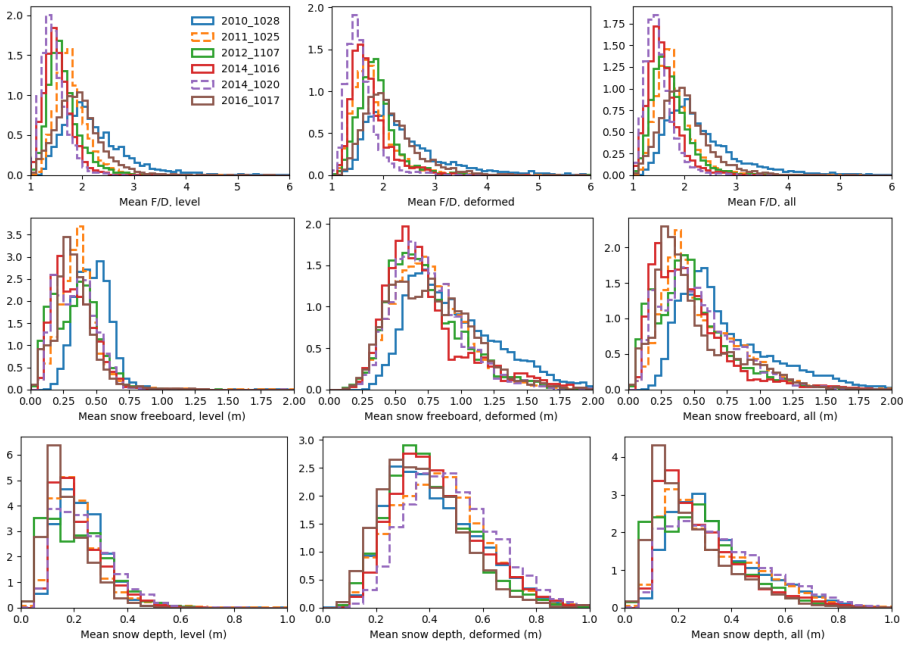

Figure 5-10: The distribution of per-window meaned Weddell F/D (top), F (middle) and D (bottom) for level, deformed and all windows.

Based on this analysis, 2011X, which has the next-highest distribution of level F/D ratios after 2010W and 2016W, would be the next-best prediction. Indeed, it is the only other Weddell test set that has a lower K-L divergence than the linear fit for the 2010W-trained ConvNet. This may imply that the ConvNet has a harder time distinguishing between different types of level surfaces, perhaps because a 'flat' surface could have varying proportions of snow depending on the thickness (in particular, age) of the ice beneath it.

This suggests that a good training set will have both a large variety of deformed topographies and undeformed topographies. 2010W provides the former; 2012W provides the latter. This combined training set improved the generalization of the ConvNet to the other Weddell test sets considerably, and also improved the generalization to the Bell/Am datasets (Fig. 5-8). Other training sets were considered, such as 2010A only, 2014A only, 2012W only, 2016 only, 2010W + 2014W, 2010W + 2012W + 2010A, and so on, but these had worse generalization than 2010W + 2012W. In particular, we found that the Bell/Am datasets could not, in any com-

bination, generalize to the 2010 and 2016 Weddell datasets, although any Bell/Am dataset could predict any other Bell/Am dataset very well. This may be because 2010 and 2016 had such anomalously high F/D for both deformed and level surfaces compared to other years. We also found that adding Bell/Am data to a Weddell training set would improve generalization to Bell/Am data but worsen generalization to other Weddell flights. This suggests that the surface types that are found in the Bell/Am Seas may be a subset of those in the Weddell. One reason may be that the large proportion of heavily-deformed, multi-year ice in the (northwest) Weddell Sea may give rise to more surface types, thus becoming a superset of those types in the Bell/Am Seas.

5.5.2 Multi-kilometer averaging of snow depth

We expect that the linear models, which *de facto* assume a constant snow/ice ratio, will have reduced relative errors as the window size increases. This is simply because, over large scales, the snow/ice ratio varies much less than at local scales. One strength of the ConvNet is its ability to predict snow depth at **all** length scales with little bias. Fig. 5-11 shows that the linear fit tends to under-predict snow depths for low F, and overpredict snow depths for high F, whereas the ConvNet has a consistent (and lower) error for all F. For a linear fit, the relative error in prediction is much higher for high F than for low F: for every increase of 0.5 m in the mean lidar window freeboard, the prediction error using a linear fit increases by 14%, but only by 1.5% using the ConvNet. The overall bias for all test sets is 3% low, i.e. the ConvNet predictions should be scaled by 1.03 to have a mean bias of 0%. This can be broken down into a bias of 3% for 'A'-type flights, 7% for 'B'-type flights, -5% for 'W'-type flights and 6% for 'X'-type flights. This bias is entirely empirically determined and may be linked to differences in the modal level ice F/D ratio for different datasets.

Using these biases to scale the predictions improves the K-L divergence for the ConvNet predictions but not the linear fits (Fig. 5-12. Here, we have separated the different flight types to account for possible sampling differences which may affect this scaling factor. Scaling is most effective for the Bell/Am flights, which have

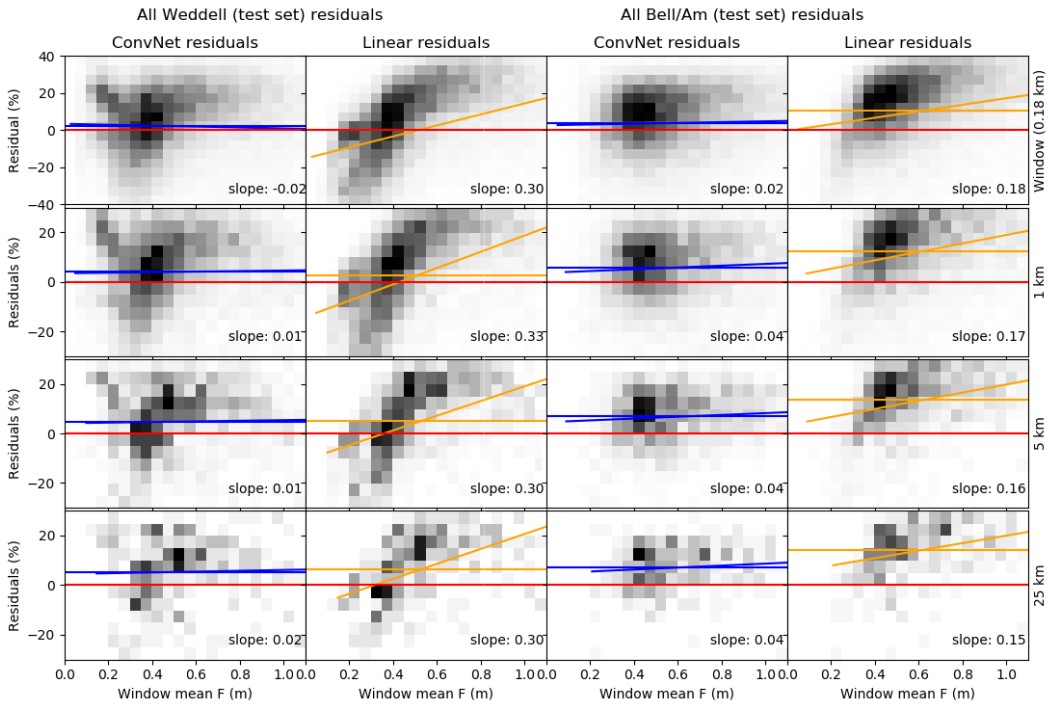

Figure 5-11: Residuals (expressed as percentages) for the 10 flights comprising the set, divided into 4 Weddell and 6 Bell/Am flights. For each window size, the horizontal line shows the mean residual for that window size, and the sloped line shows a least-squares regression for the residual as a function of F. The slope is also given in the bottom-right corner of each panel. For the Weddell, the overall mean bias is very similar at all length scales for the ConvNet and linear models, with the ConvNet slightly closer to 0 net bias for length scales >5 km and the linear model being slightly closer to 0 for length scales <5 km. For the Bell/Am, the ConvNet has a closer mean bias to 0 for all length scales. For all windows, the ConvNet shows a much flatter slope (i.e. the mean residual does not vary much with F) compared to the linear fit.

more consistent biases (Table 5-7) than the Weddell ones (Table 5-6). This suggests that the ConvNet is much better at capturing the *shape* of the distribution (and is perhaps slightly off in estimating the *scale* of the distribution) as compared to the linear fit. This is particular noteworthy as the training set was only from 'W'-type flights. This suggests that the different flights from different regions have similar surface-snow relationships (within some scaling factor, ranging from 0.95-1.07), such that the ConvNet is able to generalize better than the Linear fit. Although typically we are interested in knowing only the mean snow depth in order to predict SIT, being able to accurately reproduce the distribution of snow depths allows for other

statistics, such as the skew or variance of the snow depth estimates, to be accurately estimated. This is only possible with the ConvNet.

The bias correction term may represent interannual/inter-regional variability in F/D ratios associated with a given surface type, e.g. a level surface may have less snow one year, such that there is more ice freeboard, which would reduce the F/D ratio slightly. It is also possible that these biases are caused by the error in F due to variations in the returned lead elevations. As discussed in Sect. 2.3.2, the lead accuracy is typically better than 3 cm. For example, taking our mean F (0.61 m) and D (0.30 m) for level surfaces (defined as surface standard deviation <0.2 m) from the 2010W dataset, a shift of 3 cm in the F changes the F/D ratio from 2.03 to 1.93, which would be equivalent to a scaling factor of ∼5%.

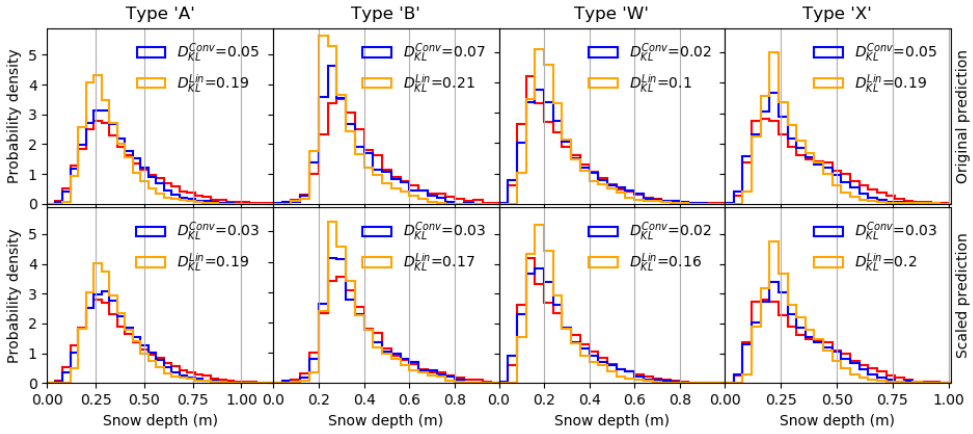

Figure 5-12: The predicted snow depth distribution using the ConvNet and Linear models, before and after applying scaling. The scaling factor for each fit and type of flight is simply 1 + the mean residual of all flights of that type (see Tables 5.1 and 5.2), weighted by number of windows in each flight. The biases are 3% for 'A'-type flights, 7% for 'B'-type flights, -5% for 'W'-type flights and 6% for 'X'-type flights.

One major advantage of the ConvNet over linear fits is that they do not require averaging over large length scales to reduce the error. Of course, as first shown in Fig. 4-7, the MRE reduces somewhat as the window size increases, as the MRE is related to the mean residual in that it is the mean of the absolute value of all residuals. However, the net residual (Fig. 5-11) actually (slightly) increases as window size increases. This is due to the distribution of residuals being asymmetrical with respect

to F, i.e. although high residuals may be found for both high and low F, for both very high $F > 0.8$ m and very low $F < 0.2$ m, bias is more likely to be positive than negative, such that the overall slope is between 0-0.04 for the ConvNet residuals in Fig. 5-11. In contrast, the linear fit clearly shows that snow depths in windows with low F are over-predicted and vice versa.

5.5.3 Effects of deformation on residuals

We expect that the linear models, which *de facto* assume a constant snow/ice ratio, will have increased larger residuals when predicting lidar windows containing deformation, as this assumption is more valid for level surfaces where we may take $F \approx D$. In contrast, one of the ConvNet's strengths is that it has the same errors for surfaces with and without deformation. This is summarized in Fig. 5-13, where it is clear that the ConvNet residual changes very little with respect to mean deformation proportion, whereas the linear fit shows a clear trend for windows with higher deformation to have higher residuals. This is particularly apparent in the Weddell test set: for example, surfaces with 50% more deformation by area will have, in general, 15% more error in the snow depth prediction. This means using a linear estimate of snow depth to predict SIT will give higher uncertainties, as the error in snow depth will be much larger. While this may be reduced by averaging over larger distances as per Sect. 5.5.2, the benefit of the ConvNet is that it will predict snow depth with consistently low bias at all length scales, with or without deformation. This is another demonstration of the ability of the ConvNet to generalize to different surface types in a way that a linear fit cannot.

5.5.4 Resolving intra-flight, inter-annual and inter-regional variability

The variability is, in essence, the differences in distribution of the snow depths, which were shown already in Figs. 5-6 and 5-7. If we are able to resolve this variability, then we can use the predicted snow depths from future measurements of F (that do

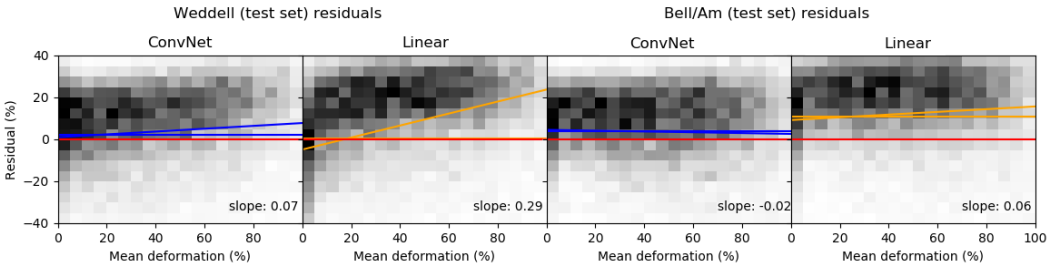

Figure 5-13: Residuals (expressed as percentages) for the 10 flights comprising the test set, separated into the 4 Weddell and 6 Bell/Am flights, binned by percentage deformation within the window. The mean residual (horizontal line) and the least-squares regression fit for the residual as a function of deformation (sloped line) is shown, and the slope is given in the top-right corner of each panel. For both regions, the ConvNet is considerably more consistent than the linear fit. There are disproportionately more windows with 0 deformation than any other single bin, and so the 0-5% deformation column has been scaled to align the highest bin count with the highest bin count for columns with >5% deformation.

not have D) to affirm changes in geophysical forcings or conditions. The ConvNet prediction is generally much closer to the true snow distribution than the linear fit, but it is possible that the linear fits could resolve a scaled version of the variability if the fit is consistently biased. To compare snow depth distributions to each other, whether inter-flight (for interannual or inter-regional variability) or intra-flight, we introduce the Wasserstein Distance (WD) metric for measuring statistical similarity. Sometimes called the Earth-Mover's Distance, which is intuitively the minimal 'amount' of the distribution, visualized as a pile of earth, that needs to be moved (along the x-axis) in order to look like the other distribution (Vasershtein, 1969). It is a true metric, whereas the K-L divergence is technically asymmetric (Sect. 4.3.2), and has the additional benefit of being defined for binned data where some bins may contain no entries. Like the K-L divergence, a perfect fit would have a WD of zero, and a higher value implies a worse fit.

To determine whether the variation in two flights, or in two zones of the same flight, is captured by the ConvNet (or linear fit), we work out the WD between the true snow distribution of the two flights/zones, and compare this to the WD between the (either ConvNet or linear) predicted snow distribution of the same two flights/zones. If the WD between the predicted snow depths of the two flights/zones is the same as

the WD between the true snow depths of the two flights/zones, then we can say that the prediction is resolving the variability. To test **interannual** variability, we take each flight for a given region (either Bell/Am or Weddell), and measure its WD for all other flights in that region. For example, we take 2010W, and work out the true and predicted WDs to 2012W, 2014W, 2016W, 2011X and 2014X. Then we take 2012W and work out its distance to 2014W, etc., for a total of 15 (ignoring duplicates, as WD is symmetric) distances. We do the same for the BellAm flights, giving another 15 distances. We then plot the true vs. predicted WDs, to see whether interannual variability is resolved (Fig. 5-14). We can repeat the same analysis for **inter-regional** variability (taking the true and predicted snow depths from each Weddell flight and computing its WD to the respective snow depths from each BellAm flight). Lastly, for **intra-flight** variability, we take the snow depth distribution of each zone, for a given flight, and work out the WD to the other zones of the same flight.

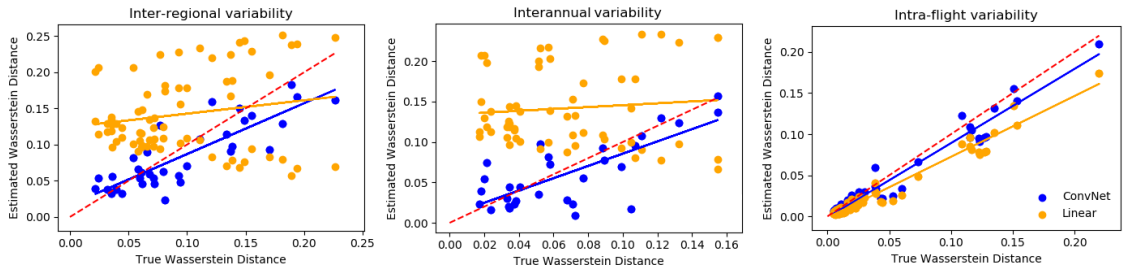

Figure 5-14: Inter-regional, interannual and intra-flight (inter-zonal) variability, for the ConvNet (blue) vs. the linear fit (orange). The 1:1 line for a perfect variability prediction is shown in red. The ConvNet is able to resolve the variability in all three cases, whereas the linear fit can only resolve intra-flight variability. The correlation coefficient for inter-regional variability is $R = 0.85$ (ConvNet) and 0.18 (linear); for interannual it is 0.75 (ConvNet) and 0.10 (linear); for intra-flight it is 0.98 for both. Slopes closer to the 1:1 line imply that the detected variability is closer to the true variability; R values closer to one imply that the variability is being resolved. For the linear inter-regional and interannual variability, the null fit has a p-value > 0.05.

As summarized in Fig. 5-14, the ConvNet can resolve all three types of variability, whereas the linear fit can only resolve intra-flight variability. Even in the intra-flight case, however, the ConvNet variability is closer to the true variability, with a best fit slope of 0.90 vs. 0.74. This is presumably because the ConvNet bias is both lower and more consistent (with respect to deformation, and also F). Intra-flight variability

is easy to resolve because in general, no matter how off the linear fit is, it will still associate higher F with higher D, as the bias in the fit would apply to all data points in that flight. This bias only becomes a problem when comparing flights to each other, as each flight has a different bias, and we cannot predict which bias will be higher or lower. This means that it is risky to use linear fits to measure variability, as a good fit for any single dataset may resolve variability within a single dataset and lead to false conclusions about interannual variability.

The slope for the ConvNet is, in all three cases, closer to one than with using the linear fit, meaning that the ConvNet-predicted variability is higher than the linear-predicted variability. Overall, the ConvNet slightly under-predicts the interannual and inter-regional variability, whereas the linear fit tends to wildly overestimate the interannual and inter-regional variability. The linear fit (given it is within error of a null fit) does not permit any conclusions regarding the variability between two flights/zones/regions, as sometimes it will identify variability when there is actually almost none. This is linked to the poor degree of correlation, R, for the linear fits, which are less than 0.2 for the inter-annual and regional cases, whereas the ConvNet R are between 0.7-0.9 for these cases. Indeed, the variability predicted by the linear fit is within error of a null fit for the inter-regional and interannual cases (Fig. 5-14). This means that not only is the ConvNet more able to resolve variability, the variability that it resolves is also closer to the true variability.

5.5.5 Implications for sea ice estimates

When trying to use snow depth estimates to estimate SIT, it may be more relevant to consider the mean absolute residual for different window sizes, so that we know the typical error in SIT/snow depth for any individual window. The total residual, allowing both positive and negative values, may be near zero (Fig. 5-13), but this only indicates that the overall estimate of mean snow depth for the entire flight is unbiased, regardless of the size of the binned windows used to calculate this overall mean. To determine the typical error at the window level, we need to look at the distribution of the magnitude of residuals for each window size (Fig. 5-15). The correlation between

ConvNet residuals and F tends to decrease, on average, with increasing window size, whereas for the linear fit it tends to increase. This means that at all length scales, the ConvNet has more consistent performance across all possible F; the expected error when estimating snow depth for a lidar window with mean $F = 0.2$ m is the same as if the mean F was 0.7 m instead, whether you are looking at local ($\leq$ 1 km) or satellite-observation (25 km) scales.

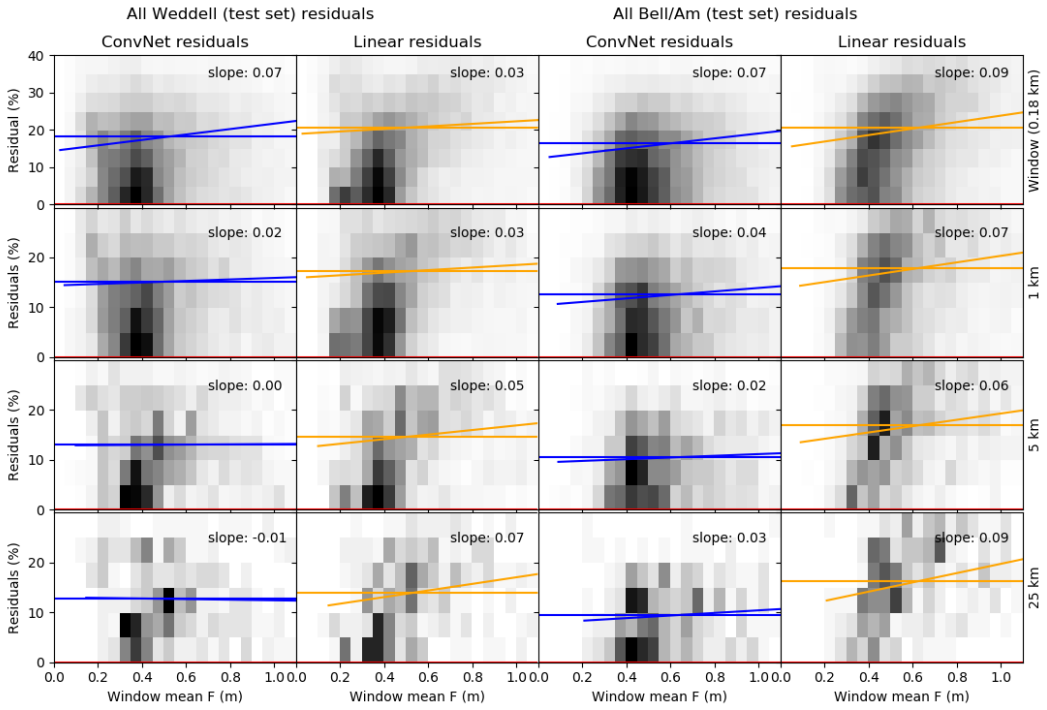

Figure 5-15: Magnitudes of relative residuals (as percentages) for the 10 flights comprising the test set, separated into 4 Weddell (left) and 6 Bell/Am (right) flights, for window sizes ranging from 0.18 km x 0.18 km to 0.18 km x 25 km. The average residual for that window size is shown as a horizontal line, and the overall least-squares regression as a function of F is shown as a sloped line, with the corresponding slope in the bottom-right corner of each panel. For both regions (though particularly for the Bell/Am), the ConvNet shows a lower average error than the linear fit for all window sizes. The ConvNet, in general, also shows a flattening of the slope as window sizes increase, whereas the linear fit shows the opposite trend.

The mean residual is also lower in magnitude for the ConvNet at all length scales. One interesting point is that the ConvNet seems to have slightly lower residuals when predicting Bell/Am data as compared to Weddell data (e.g. at 5 km scales, the mean magnitude of the residual is 11% for the Bell/Am vs. 14% for the Weddell). In

contrast, the linear fit has higher test errors for the Bell/Am (17%) than the Weddell (15%) for the same length scale of 5 km. This is in line with the idea that linear fits are difficult to generalize between regions due to, for example, different frequency of deformation (comparing Fig. 5-3 and Fig. 5-4), and also explored by Özsoy-Çiçek et al. (2013); Li et al. (2018). The ConvNet, in contrast, generalizes much better (and indeed has lower test errors in the Bell/Am despite being trained on Weddell datasets only). This may indicate that features found in the Bell/Am are only a subset of those found in the Weddell. To support this point, we found that a ConvNet trained in Bell/Am data predicted Weddell test sets with far higher error than the Bell/Am test sets. Also, as discussed previously (Sect. 5.5.4), linear fits cannot resolve inter-flight variability despite being able to reproduce mean snow depths (because they cannot reproduce the snow depth distribution), whereas ConvNets can. This means that we cannot just look at point-based error metrics (MRE or RMSE) and we need to also consider whether the distribution itself is accurately predicted.

As previously shown in Sect. 4.4.3, with a mean relative error of 14% for ConvNet-predicted snow depth, using typical values for sea ice and snow parameters, we can achieve a SIT error estimate of 20%, which is dominated by the error in sea ice density. Just taking the error from uncertainty in ice density from Eq. 4.1, we have:

$$\epsilon_T = \epsilon_{\rho_i} \frac{F\rho_w}{(\rho_w - \rho_i)^2} + \frac{D(\rho_s - \rho_w)}{(\rho_w - \rho_i)^2} \tag{5.1}$$

The relative error, which is ϵ_T / T where T is given by Eq. 1.1, we have:

$$\begin{aligned} \frac{\epsilon_T}{T} =& \epsilon_{\rho_i} \frac{\frac{F\rho_w}{(\rho_w-\rho_i)^2} + \frac{D(\rho_s-\rho_w)}{(\rho_w-\rho_i)^2}}{\frac{\rho_w F + (\rho_s-\rho_w)D}{\rho_w-\rho_i}} \\ =& \epsilon_{\rho_i} \frac{1}{\rho_w - \rho_i} \end{aligned} \tag{5.2}$$

This is a constant with respect to F and D, and only depends on the values chosen for ρ_w, ρ_i and estimated for ϵ_{ρ_i}. Using the values from Sect. 4.4.3, this comes out to 18.3%. As sea ice density cannot be better constrained, and in fact may be even more

variable than previously thought (e.g. Hutchings et al., 2015), this is essentially an irreducible error for predicting SIT from F and D. Although we find, in agreement with Kern et al. (2016), that improving snow depth measurement accuracy would certainly improve SIT errors, without improved sea ice density estimates the SIT error cannot be reduced below 18%. However, a method that directly predicts SIT from a surface F, without distinguishing F and D in the snow freeboard, could be more promising as the ConvNet can account for sea ice density variations, essentially as one mixed layer with some effective density, as suggested by Worby et al. (2008); Kern et al. (2016) and attempted in Sect. 3.4.2. As previously argued in Sect. 3.3.2, this suggests that predicting SIT directly from F would have lower errors, as the ConvNet can learn non-constant values for snow and ice density. However, this requires simultaneous snow freeboard and sea ice thickness data, which is harder to obtain, although such datasets do exist (e.g. Haas et al., 2009). Alternatively, we can also improve ConvNet predictions of SIT by having better constraints on the variability in sea ice density.

5.5.6 Application to 2018B flight

As a demonstration of the viability of the ConvNet to generate snow depths for lidar datasets that have no snow depth points, we use the ConvNet to predict the snow depths for the 2018 Bell/Am flight, which does not (yet) have processed snow depth data. We also compare the predictions to those generated by extrapolating another Bell/Am flight that does have snow depth data. On average, extrapolating any Bell/Am flight with any other Bell/Am flight has an overall MRE of 16-19%. When testing the accuracy of using one Bell/Am flight to extrapolate all other Bell/Am flights, the dataset with the lowest average MRE (17.5%) was 2012A. We therefore use this to extrapolate 2018B. Although we cannot know the error of this extrapolation to 2018B, the extrapolations of all other Bell/Am data suggest that the mean relative error of the extrapolated segments is 15-20%. Our predictions for 2018B, sectioned into zones, are shown in Fig. 5-16. The ConvNet and extrapolation agree very well with each other after a scaling of 7% is applied to the ConvNet predictions

(following Sect. 5.5.2). Taking the extrapolated snow depth as the ground truth, the K-L divergence of the ConvNet is 0.006 (before scaling, 0.042), vs. the linear fit (0.076 after scaling and 0.149 before scaling). Although in this case, the scaling improved the K-L divergence of the linear fit, we note that this is generally not the case (Fig. 5-12), and even in this case, the pre-scaling K-L divergence of the ConvNet prediction was lower than the post-scaling K-L divergence of the linear fit. In general, each scaling is flightpath-specific, and so for locations not covered by OIB (which is likely for ICESat-2 data), no scaling can be applied.

Moreover, it should be pointed out that it is striking that the ConvNet, being trained on 2010W/2012W, so closely emulates the textural extrapolation (which was extrapolated using 2012A). The mean relative difference (calculated the same way as the MRE, except the extrapolated snow is not necessarily the ground truth and hence it is not an 'error') is 6.7%, corresponding to a RMS error of 3.1 cm. This suggests that the surface-to-snow relationships within 2010W/2012W that are captured by the ConvNet are the same (within some scaling factor) as the segment-matching algorithm. This suggests that the segment snow properties can be characterized by their F/D ratio, and matched by their standard deviation, image entropy, mean and L-kurtosis.

Another point that may be raised by the similarity between the textural extrapolation and ConvNet results in Fig. 5-16 is that there may be no point to using the ConvNet to predict the snow depth of future lidar datasets, if we can just segment the input lidar and apply the textural extrapolation algorithm. Indeed, Table 5.3 and Fig. 5-17 suggest that the predicted mean snow depth with extrapolation from 2010W+2012W is almost the same as the ConvNet prediction, with a MRE of 12% (vs 11% for the ConvNet, though the ConvNet MRE can be improved to 6% with scaling, following Fig. 5-12). The ConvNet also predicts closer snow depth distributions to the self-extrapolated snow depths, with typical K-L divergences of 0.05-0.09 (0.04-0.07 with scaling) vs. 0.06-0.14 for extrapolation from 2010W+2012W and 0.21-0.25 for the linear fit. However, there are two particular reasons to prefer the ConvNet. Firstly, the ConvNet is being trained on the self-extrapolated snow depth data, using

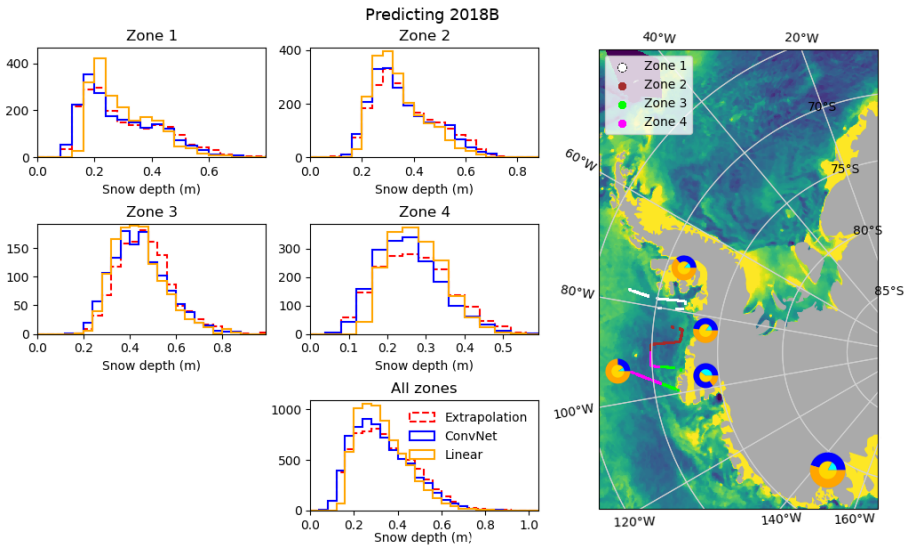

Figure 5-16: Predictions for snow depth for the 2018B flight, using the ConvNet, empirical linear regression and textural matching the segments to 2012A. The ConvNet prediction has been scaled by 1.07 following Sect.5.5.2. The ridging frequency, deformation proportion and scatterometer data are also shown (right).

it as the ground truth. We showed that the extrapolated snow depths themselves had a typical error of 16-19%. In this sense, if the ConvNet were predicting with more accurate mean snow depths for the input lidar windows, then the test error may well decrease. Secondly, the ConvNet method can be generalized using 1-D convolutions for use with ICESat-2 data. In contrast, the extrapolation algorithm requires metrics like image entropy, which are not well-defined for 2-D datasets. 2-D lidar data may also be harder to cluster, as the ICESat-2 track may cross over sinuous deformation features which would not be easily segmented in 2-D.

5.6 Conclusion

This chapter shows that textural information from snow surface freeboards can be used to accurately predict snow depths using a convolutional neural network with 15-20% errors when tested on different datasets from different regions and years. These have much better generalization than linear fits, which cannot use this textural information. In particular, we find that the residuals for the ConvNet are both lower

Table 5.3: Predicting the mean snow depth for an entire flight using self-extrapolation from the existing snow depth measurements in that flight; using extrapolation from the 2010W+2012W superset; using the ConvNet trained on 2010W+2012W [with scaling applied]; using the linear fit. The superset extrapolation and ConvNet predictions are all generally within a few cm of each other, except for 2014X. Scaling (Fig. 5-12) of the ConvNet results improves the Bell/Am matches considerably and the Weddell matches somewhat. Excluding the training set (2010W+2012W), the average error in the mean snow depth is 12% for the superset extrapolation, 11% for the ConvNet prediction (6% with scaling) and 17% for the linear prediction.

Date	Mean flight-wide snow depth (cm)			
	Self-extrapolated	2010W+2012W	ConvNet	Linear
2010W	33.7	37.3	35.4 [33.6]	36.0
2011X	31.8	29.8	30.0 [31.8]	28.7
2012W	28.0	26.3	27.3 [25.9]	25.7
2014W	28.4	25.1	25.6 [24.2]	24.9
2014X	35.2	28.1	28.0 [29.7]	27.6
2016W	24.7	26.8	27.7 [26.3]	26.1
2010A	42.9	40.9	40.7 [41.9]	38.6
2011A	31.9	29.3	30.2 [31.1]	27.1
2012A	34.1	30.9	31.5 [32.4]	29.0
2012B	31.9	27.6	28.2 [30.2]	25.9
2014A	41.8	35.8	37.0 [38.1]	33.1
2016B	47.4	42.2	43.9 [47.0]	38.2

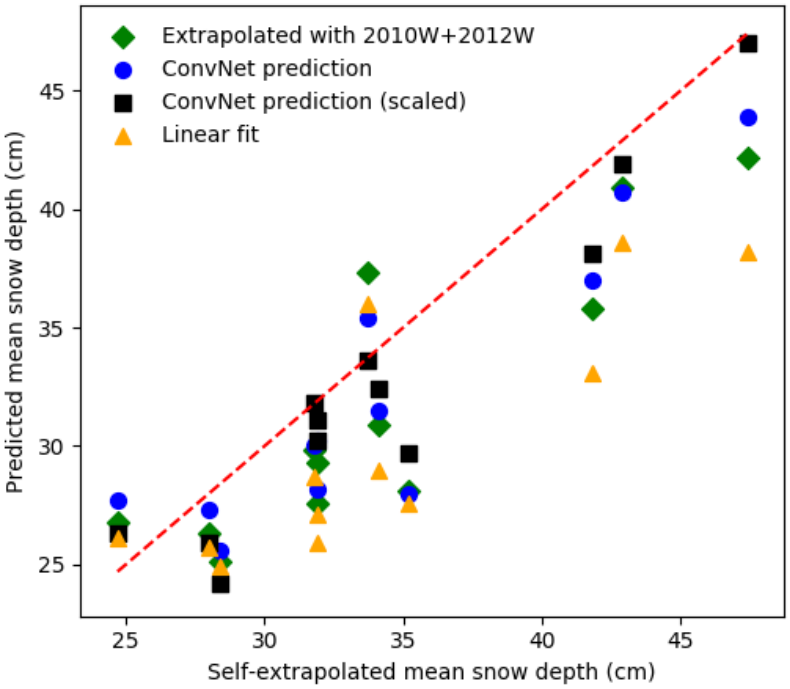

Figure 5-17: Table 5.3 plotted, showing the mean snow depth estimated using a ConvNet trained on 2010W+2012W (both with and without our empirical scaling from Fig. 5-12), compared to extrapolating the segments directly from 2010W+2012W, as a function of the true (self-extrapolated) snow depths.

than the linear fit, as well as less correlated to the window mean F and surface deformation proportion. This means that ConvNets can be used to predict snow depth at both local ($\leq$ 5 km) and satellite ($\sim$ 25 km) scales, a key difference with linear fits.

Moreover, the ConvNet is able to model the distribution of snow depth much more accurately. This is important for modeling processes which have non-linear relationships to the sea ice thickness, and hence cannot be modeled with just the mean thickness. The K-L divergence from the true snow depth distributions is much lower for the ConvNet prediction, typically $\ll 0.1$, and becomes even lower when a small scaling factor (of less than 10%) is used to account for possible differences in surface-to-snow relationships between different regions and/or flight tracks. This means that the ConvNet can more accurately estimate of higher-order statistics like

the variance and skew of the snow depth distribution. Similarly, because the ConvNet models the distributions much more accurately, despite having a low bias, it is able to resolve interannual and inter-regional variability, whereas linear fits cannot.

The importance of creating a representative training set that spans all possible surfaces cannot be understated. In particular, the training set needs to have a wide set of deformed surfaces, as well as a wide variety of level surfaces. In particular, level surfaces that have varying amounts of snow (and hence varying underlying ice thicknesses) need to be represented in the training set in order to ensure good interannual and inter-regional generalization. In particular, the surfaces found in the Weddell may be a superset of those in the Bell/Am, perhaps due to the significantly larger proportion of multi-year ice.

This technique may also be expanded to incorporate other types of lidar data, such as ICESat-2. There were several OIB underflights for ICESat-2 in 2018, which could be used as training sets for ICESat-2 data from subsequent years that do not have OIB underflights. Alternatively, we can artificially generate ICESat-2 lidar data for the rest of the OIB catalog from 2010-2018, using a subsampled along-track mean of the lead-referenced lidar to emulate the 1-D lidar product (with a 10 m footprint) generated by ICESat-2. In addition, the 2013 OIB flight over the Ross Sea operated without a snow radar onboard, and this technique may be used to obtain snow depth estimates for the Ross Sea. The Ross Sea shows similar SIE trends to the Weddell Sea, and the 2013 data show a multimodal mixture of F due to the significant new (thin) ice formation and ridging, but without multi-year ice (Tian et al., 2019). It is thus plausible that the Ross Sea surface features are also a subset of the Weddell Sea ones.

The success of the ConvNet opens the door to 1D ConvNets using the ICESat-2 data to predict snow depth, using slightly less structural information but is a more informed choice than a generic deep neural network. This saves computational complexity (and reduces overfitting). To do this, it is possible to use use OIB lead-referenced lidar datasets to generate quasi-ICESat-2 data, using interpolated, windowed mean F along-track (potentially using multiple lines parallel to each other along-track to

increase training set size).

However, we find that even with a perfect snow depth prediction, an uncertainty in sea ice density of 20 kg m^{-3}, and to a lesser extent the $\sim$20% error in snow depth extrapolation, may lead to an irreducible uncertainty in kilometer-scale SIT predictions of $\sim$18%. This means that methods that directly predict SIT may be preferable, such as that discussed in Chap. 3, although there is much less training data for this (e.g. Haas et al., 2009). Alternatively, more work can be done in investigating the variability of sea ice density (which may also be linked to the surface morphology).

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Final Summary

6.1 Future work

Our approach to choosing the ConvNet architecture was a mixture of intuition and some experimentation. This could be made more exhaustive to make sure that the chosen hyperparameters, such as learning rate, number of layers, filter size etc. are optimal. This may further improve the ConvNet performance, though this would not change any of our conclusions.

In order to increase the interpretability of the ConvNet, we can add predetermined physically-plausible variables to the post-convolution (fully-connected) layers. For example, we added the mean F to the final layer in Chap. 5 as we expected this to be useful for the ConvNet's prediction. We could also add, for example, the surface morphology or the surface rugosity. This would be a lot of guesswork to see which variables improved performance, but it would give us a better idea of which surface properties are actually important for the ConvNet prediction. We can similarly use a variety of simulated deformed surfaces using ridging simulations to see if our predictions of SIT or snow depth are accurate, and how they differ as we change the surface incrementally.

The success of using surface structural features to predict snow depth may be extended to 2-D datasets, such as those from ICESat-2, by using 1-D convolutional filters to predict snow depths and hence SIT. As it is easy to downsample existing 3-D

lidar datsets from OIB to match the ICESat-2 resolution, we can generate artificial ICESat-2 data for 2010-2018 by downsampling the OIB data to match ICESat-2 resolution. We can then train a ConvNet on this artificial ICESat-2 data, to be applied to real ICESat-2 data from 2018 onward. Other regions may also be predictable by this method, though a lack of OIB flights with snow depths in the Ross Sea and East Antarctica regions makes this harder to verify. However, it is plausible that the Weddell, having considerable multi-year ice, has a larger variety of surfaces than other regions (hence it was able to generalize to the Bell/Am, but not always vice versa). Similarly, as there are no OIB flights in other seasons, it is not clear whether our trained ConvNet could be used to predict other seasons. The types of features may be different in different seasons, though there are likely some overlapping features (e.g. there is new sea ice formation in both spring and winter in the S. Weddell) which may have similar snow/ice proportions. Ideally, there would be some more OIB underflights of ICESat-2 that collect snow depth data for different regions/seasons. This would allow for greater generalization of our ConvNet performance.

It may also be possible to generalize this method to Arctic data. The Arctic has a lot of multi-year ice that would likely behave similarly to our Weddell dataset. However, the occurrence of melt ponds, which do not feature in our Antarctic dataset, may add an additional complication to surface morphology. This may add information (e.g. it may give details about the age of the snow, and hence ice), and identifying different types of melt ponds may itself be an interesting application of a ConvNet. Otherwise, the smoothing effect of snow, in particular for multi-year ice, may give rise to similarly-flat surfaces that have varying snow proportions (which would be harder to train a ConvNet from). Moreover, due to the different geography (i.e. that the Arctic is a basin surrounded by continents), the ridging structures may look different, with potentially different ice densities too. This suggests that we expect an Antarctic-trained ConvNet to apply to the Arctic; instead, we can train a ConvNet directly on Arctic data.

We found that more work needs to be done in constraining the sea ice density, as using snow depth and freeboard to predict SIT is now dominated by, and hence

limited by, uncertainties in sea ice density. This suggests that direct predictions of SIT from snow freeboard, using ice thickness + high-resolution surface lidar scans like in Haas et al. (2009) may have more success, as the sea ice density (and snow density) is implicitly accounted for during the deep learning process. As the variability in snow depths can be resolved via convolutional neural networks, the variability in SIT is likely also resolvable, in contrast to current methods using linear fits. This suggests that more measurements of SIT and F need to be collected to further decrease SIT prediction errors of convolutional neural networks. Alternatively, we can collect more measurements of sea ice density, so we can understand how it varies for different surface types (as it is likely that sea ice density itself depends on the surface morphology).

6.2 Final conclusions

Sea ice thickness and its interannual/regional variations have long been difficult to assess due to the paucity of measurements. Because sea ice thickness cannot be easily measured by remote means, often snow freeboard, sometimes in tandem with snow depth, is used to estimate the SIT. However, snow freeboard remains far easier to measure remotely than snow depth, and as a result there is considerably more snow freeboard data than snow depth data. Errors in snow depth estimates have long plagued SIT estimates, sometimes to the tune of 50%. Despite the existence of a few simultaneous and collocated datasets with both snow depth and snow freeboard, the distribution of snow within the snow freeboard is not well understood, limiting efforts to estimate snow depth, which limits efforts to estimate SIT, which ultimately limits our ability to resolve interannual/regional changes in SIT.

Deep learning techniques have recently been applied to solve sea ice problems to great success. Remote-sensing datasets, such as those collected by satellite or Operation IceBridge, are typically vast and current studies have not yet used these datasets to their full potential. Computer techniques can be used to automate tasks like finding local sea surface heights, as well as tasks that are less clear to the human eye,

like segmenting snow surfaces into areas of similar texture. Deep learning techniques are also more complex than typical linear regressions used in SIT and snow depth estimation.

In Chap. 3, we showed that high-resolution (sub-meter) SIT could be accurately predicted using a convolutional neural network to less than 20% error (i.e. better than $\sim$50% errors for linear fits, following Kern and Spreen (2015)) using a high-resolution snow depth, snow freeboard and ice draft measurements taken during the PIPERS expedition to the Ross Sea. The sparsity of other such datasets means that we were unable to establish whether the relationship between snow freeboard and SIT could be translated into other regions or seasons. We found that the learned convolutional filters appeared to correspond to edge detectors and basic snow features, and that the ConvNet appeared to be accounting for differences in snow/ice ratios, snow density and ice density simultaneously as one effective density. In contrast, empirical linear fits typically fail to generalize to other floes, even within the same region, as they implicitly assume a constant proportion of snow and ice in the snow freeboard, which is not true. We also found that adding the surface standard deviation only marginally improves the linear fit, suggesting that the more complex metrics for surface roughness learned by our ConvNet are necessary to reduce the error in SIT prediction.

In Chap. 4, we showed that the snow freeboard measurements from Operation IceBridge, a far larger dataset, could be used to estimate snow depths. Firstly, we showed that snow depths could be matched, using the average snow freeboard to snow ratio, to texturally similar segments using standard deviation, mean snow freeboard, image entropy and L-kurtosis, with snow depth prediction errors of $<$20%, which could be then used to extrapolate the 2-D snow depth measurements onto the 3-D lidar scans. Then, we trained a ConvNet to predict snow depths, and found that the learned filters appeared to correspond to edge detectors and also steerable pyramid kernels, suggesting that the ConvNet was learning something similar to textural metrics used in computer vision. Our 14% error in snow depth estimates at a length scale of 2.5 km translated to an uncertainty in SIT of around 20%. We found that our reduced error in snow depth, especially when averaged over multi-kilometer scales, meant that

the error in snow depth was no longer the largest component of SIT uncertainty, but rather the sea ice density was.

In Chap. 5, we showed that a ConvNet trained on snow freeboard and snow depth values from the Weddell Sea was able to generalize to the Bellingshausen and Amundsen Seas (and that a linear fit could not). We showed that the ConvNet snow depth predictions had biases that were almost constant with respect to snow freeboard and deformation proportion, unlike linear fits, which meant that snow depths at the sub-kilometer scale could be predicted with the same (low) bias no matter the surface characteristics. We then applied the ConvNet (trained on only Weddell data) to a 2018 Bellingshausen Sea flight data lacking snow depth measurements, and showed that the ConvNet predicted very similar results to matching the segmented textures to another Bellingshausen Sea flight, suggesting that textural features and their relationship to snow depth in the Bellingshausen/Amundsen Seas are consistent (perhaps a subset) of those in the Weddell Sea. Although linear fits are comparable to ConvNets when estimating large (25 km) scale means, their prediction of small-scale means is highly biased, which means they cannot predict the snow depth distribution, and hence are unable to resolve the large amounts of interannual and inter-regional variability that characterize Antarctic sea ice, whereas the ConvNet can. The ConvNet also shows a consistent (and low) bias with respect to the mean surface elevation and deformation amount, whereas the linear fit has much higher bias for higher surface elevations and higher deformation amounts. We found that the ConvNet needed to have a large 'library' of surface types to be trained on: in particular, we found that the training set had to span a wide variety of undeformed (level) surfaces with varying amounts of snow cover. This 'library' can be also used to directly match texturally-segmented snow surfaces, which may also be used as an indicator of how similar snow surfaces between different datasets are (and hence whether the ConvNet could be generalized to a new dataset of lidar surfaces).

ConvNets, when properly constructed with suitable training sets, are a powerful way to identify relationships between surface morphology and either SIT or snow depth. Given the large amount of surface elevation data from sources such as Oper-

ation IceBridge and ICESat-2, our ConvNet offers a superior way to estimate snow depths and sea ice thicknesses, instead of linear fits that cannot account for varying snow proportions in the snow freeboard (equivalently, varying effective densities of the ice). By providing lower-error and lower-bias estimates of snow depth and SIT, our ConvNet is able to resolve regional and interannual differences, and provides an alternative method to generating snow depth/SIT estimates from future laser altimetry datasets such as ICESat-2.

Appendix A

Example of Segment Matching

This appendix gives a demonstration of the snow depth extrapolation algorithm from Sect. 4.2.1.

Figure A-1 shows an example of the segment-matching algorithm in action. For this example, we only use a subset of nearby windows for clarity. All relevant metrics are for all segments shown are in Table A.1. For example, for segment 1e, the only other segments it is matched to (here, we use a similarity threshold of 0.04 so that we can just have 2 segment matches) that contain snow depth measurements are 3c and 5e. The set of metrics $M = \{F, \sigma, \text{entropy, L-kurtosis}\}$ for 1e is $\{0.4349, 0.096, 3.841, 0.152\}$. Taking the differences between these values, M_{1e} and their corresponding values for segments 3c (M_{3c}) and 5e (M_{5e}) gives $|M_{3c} - M_{1e}| = \{0.007, 0.006, 0.071, 0.009\}$ and $|M_{5e} - M_{1e}| = \{0.173, 0.003, 0.022, 0.073\}$. Adding 0.001 to each value (to prevent taking a geometric mean with a zero) and then taking their geometric means to give the similarity metric S gives $S_{3c} = 0.0142$ and $S_{5e} = 0.0330$. The number of snow depth measurements for each segment are $N_{3c} = 3$ and $N_{5e} = 5$. We cannot just use a weighted arithmetic mean of the snow depths due to a known sampling bias from the snow depth radar (Sect. 4.4.1). Instead, we will work out the average F/D ratio and apply that to the mean F of 1a.

The resulting estimate for the mean F/D ratio for 1e is the weighted harmonic mean of the ratios for segments 3c and 5e, weighted by N/S (normalized). We assume more points = more confidence in the textural match, and lower S = higher similarity

= more confidence in the textural match. The (unnormalized) weights are therefore $N_{3c}/S_{3c} = 3/0.142 = 212$ and $N_{5e}/S_{5e} = 5/0.0330 = 151$. To normalize, we simply divide the weights by their sum, to give a final weight of 0.58 for segment 3c and 0.42 for segment 5e. The mean ratio is determined by taking the weighted harmonic mean of the F/D ratios corresponding to segments 3c and 5e, i.e., $\frac{1}{\frac{0.58}{4.571}+\frac{0.42}{5.126}} = 4.79$. The estimated snow for this segment is therefore $F_{1e}/4.79 = 0.091$ m.

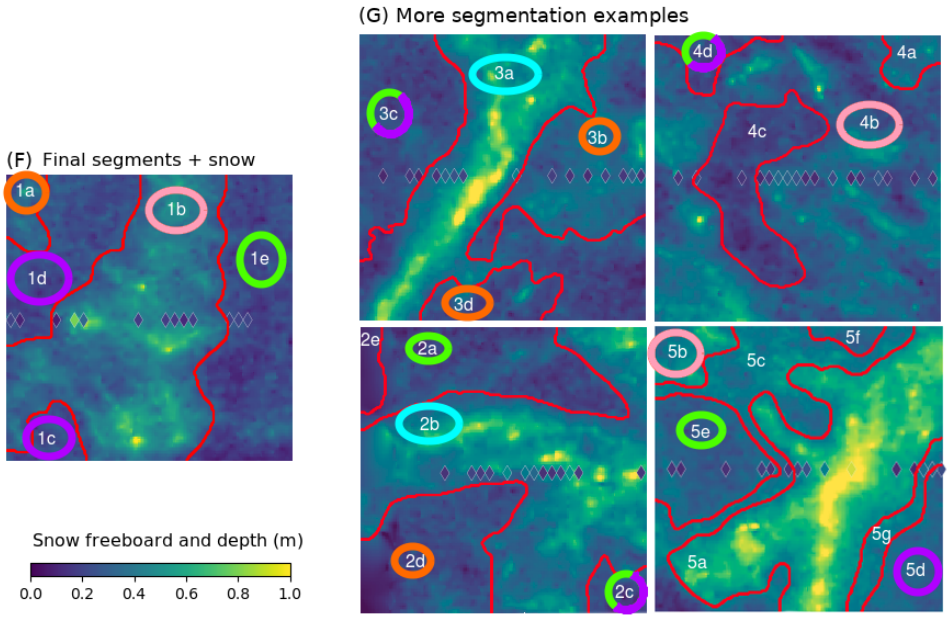

Figure A-1: Example segment matching, for a similarity threshold of 0.04. Segments are color-coded to show their textural matches: 1a is matched to 2b, 3b, 3d; 1b is matched to 4b, 5b; 1c is matched to 1d, 2c, 3c, 4d, 5d; 1e is matched to 2a, 2c, 3c, 4d, 5e and 2b is matched to 3a.

The true (extrapolated) segment mean snow depth (scaling F_{1e} by the F/D ratio for that segment) should be $0.434/2.290 = 0.190$ m, so the overall prediction error for this segment is 52% (or 9.9 cm). Note that in the full algorithm, instead of taking 4 nearby lidar windows, all lidar windows within ± 10 km are checked for textural matches (potentially up to 120 windows). The similarity threshold is also higher in the real algorithm (0.050), which will also give more matching segments (in this example, it would also match 1e with 1d and 4c). Lastly, we require at least 9 snow depth samples to be used in the calculation, and so this particular example (that uses

a total of 8 snow depth points) would be discarded as a low-quality estimate.

Table A.1: List of metrics for the segments in Figure A-1: the segment ID, the segment area, the number of snow depth samples in that segment (N), the mean of all raw snow depth measurements (D), the mean snow freeboard for the whole segment (F), the standard deviation of the snow freeboard (σ), the mean entropy of the segment, the L-kurtosis of the segment snow freeboard. F/D is the harmonic mean of all snow depth measurements within the segment and the corresponding mean snow freeboard for that snow radar footprint. This means that F/D is not the same as taking the quotient of the F and D columns.

ID	Area (m^2)	N	D (m)	F (m)	σ (m)	Entropy	L-Kurtosis	F/D
1a	1143	0	-	0.6255	0.133	4.357	0.191	-
1b	16454	8	0.2546	0.805	0.209	4.506	0.133	3.72
1c	1090	0	-	0.4833	0.074	4.016	0.092	-
1d	5032	2	0.1500	0.491	0.074	3.732	0.121	3.34
1e	8681	3	0.2033	0.434	0.096	3.841	0.152	2.29
2a	6374	0	-	0.4277	0.113	4.150	0.111	-
2b	18009	13	0.2355	0.887	0.274	4.726	0.097	3.83
2c	1001	0	-	0.4279	0.103	4.088	0.119	-
2d	6276	0	-	0.4160	0.138	3.981	0.193	-
2e	740	0	-	0.2313	0.071	3.799	0.027	-
3a	16027	5	0.2555	0.864	0.409	4.753	0.084	3.00
3b	7698	7	0.1573	0.660	0.154	4.209	0.200	4.31
3c	6451	3	0.1039	0.441	0.090	3.770	0.143	4.57
3d	2224	0	-	0.4831	0.148	4.400	0.187	-
4a	964	0	-	0.6883	0.105	4.117	0.147	-
4b	25523	16	0.1680	0.637	0.197	4.472	0.137	3.44
4c	4784	1	0.1322	0.374	0.086	4.024	0.128	2.75
4d	1129	0	-	0.4086	0.078	3.937	0.118	-
5a	14752	6	0.3305	1.313	0.291	4.944	0.123	4.43
5b	1521	0	-	0.9537	0.203	4.678	0.139	-
5c	7251	1	0.1305	0.779	0.160	4.431	0.246	5.27
5d	2424	1	0.2169	0.636	0.076	3.608	0.097	3.16
5e	4486	5	0.1281	0.607	0.093	3.819	0.079	5.12
5f	548	0	-	0.6733	0.079	3.935	0.056	-
5g	1418	2	0.3406	0.840	0.083	4.265	0.188	2.26

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Lead-finding in the OIB ATM data

This appendix provides a more detailed look at the algorithm used to reference the surface elevation (F) measurements from Operation IceBridge to the local sea level (Sect. 2.3.2).

B.1 Thin (gray) ice and lead elevations

Typically, if an image has (white) ice, (black) water, and thin (gray) ice, the pixel intensities will have three peaks corresponding to these three components. In this case, the minima between these peaks are used as the thresholds for these three components. An ideal example, with three such peaks, is shown in Fig. B-1.

However, not all images will have all three components, and even if it does, it may not have all three peaks. This is generally due to the wider range of grayscale values for gray ice, which can be different shades depending on its thickness. In these cases, the locations of the peaks are used to infer which components exist, and to create a best-guess estimate of the thresholds. Typically, the water peak is around a pixel intensity of 50 (out of 255), and for ice it is somwhere above 100, with thin/gray ice somewhere in between. Depending on the particular range of thicknesses of the gray ice (if there is any in the image), there may be no peak (in particular, if the peak is near one of the other peaks, e.g. Fig. B-2), or multiple peaks, between the water and ice peaks. In these cases, the peak with the highest grayscale intensity is identified,

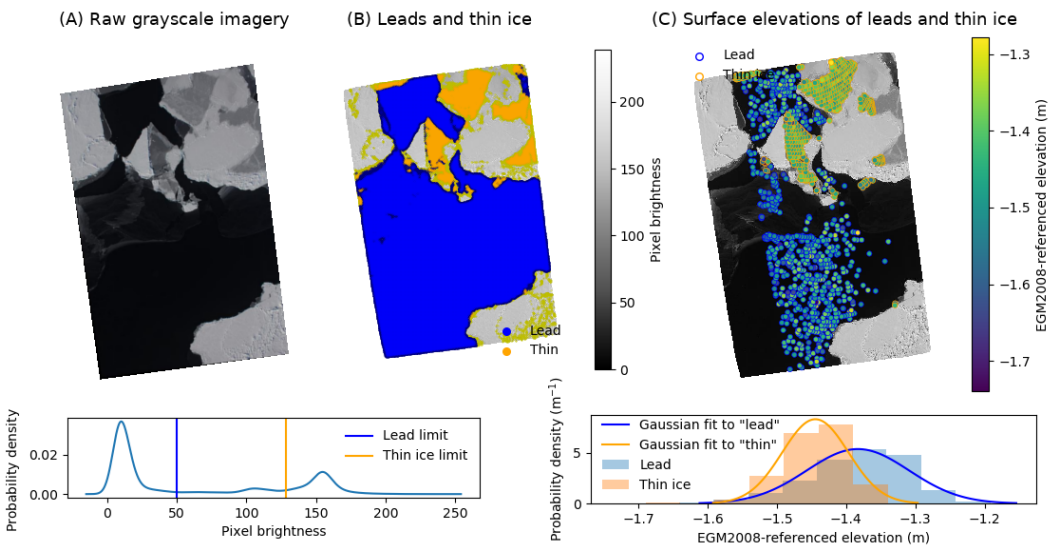

Figure B-1: (A) The raw DMS camera imagery from an OIB flight. (B) The open water and thin/gray ice areas identified, using the peaks in the pixel brightness distribution (bottom). (C) The surface elevation distributions for these regions, binned at 5 cm, with a Gaussian fit added. Note that the 'thin' elevations seem to have a lower mean than the 'lead' elevations.

.

and if it is above 100 (i.e. it is inferred to be white ice), then the half width at half maximum (HWHM) in the direction away from the peak (i.e. towards 0) from this peak is calculated, and the grey ice threshold is set as (white ice peak − 1.5× the HWHM value) (Fig. B-2).

Sometimes, when the sun is low (as OIB flights can take as long as 10 hours), the images will be insufficiently bright. These images are first brightness-corrected by scaling the 99th-percentile pixel brightness to the maximum (255). Another consequence of the low sun is that shadows from ridges may be accidentally flagged as thin/gray ice. Thes shadows can be eliminated by using erosion techniques to 'shrink' the lead and thin-ice regions, as ridge-cast shadows are typically thin and long (like ridges), whereas grey ice has no such constraint. We use a 3 pixel erosion in all directions, which *de facto* sets a lower bound on the minumum size of grey ice that can be picked out to at least a 6 x 6 pixel square. An example of this is shown in Fig. B-3.

However, shadows from clouds have to be manually eliminated as these can have any shape. Icebergs are rarely encountered along the OIB track, and can be filtered

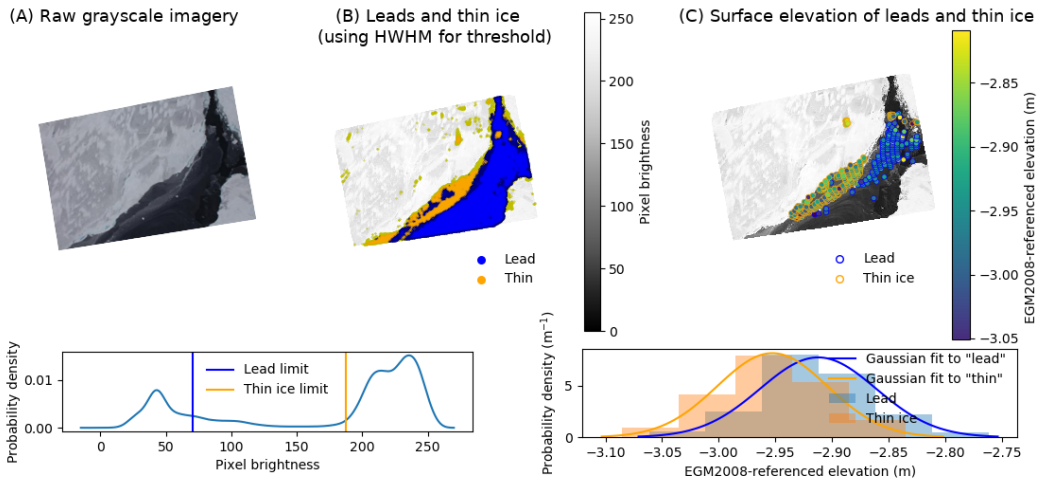

Figure B-2: Same as Fig. B-1, but in the case of there being no clear peak corresponding to thin ice, a threshold is estimated based off the half width half maximum (HWHM) of the ice peak. The gray ice threshold is taken as the pixel intensity that is 1.5× the HWHM of the ice peak.

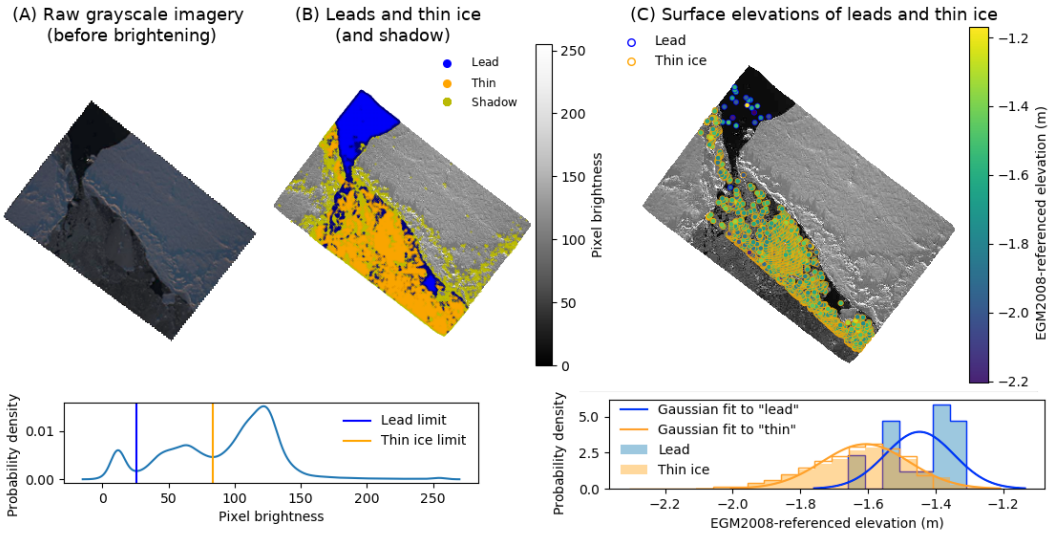

Figure B-3: The same as Fig. B-1, but this time there are shadows in (B) which have been removed by erosion (3 pixels in all directions). This means that (most of) the thin ice is kept, and the shadows are excluded from the thin ice filter. Note that the raw image in (A) was quite dark due to the low sun (which also caused the shadows), and so the image was brightened. Here, as with Fig. B-1, the thin ice has a lower apparent elevation than the leads.

out by setting range limits on the geoid-referenced surface elevations.

B.2 Compiling the final list of leads for lead-referencing

One surprising point in Figs. B-1, B-2 and B-3 is that the lead elevations are sometimes apparently higher than the thin ice elevations. The difference is typically between 5-10 cm, though can be as high as 20 cm. Although this is not always the case, the majority of images that have both lead and thin ice elevations have this situation. The reasons for this are unknown. One possibility is that the lidar return is coming from the ice-water interface as opposed to the air-ice interface for thin ice. Another is that the water, which is close to black and has hence a low reflectivity, has poorer quality returns.

Because we do not know which direction the correction should be applied (i.e. whether the thin ice should be corrected to be higher elevation, or whether the leads should be corrected to lower elevation), we are forced to pick one to ensure the self-consistency of the list of lead elevations. In particular, when there is no thin ice in the image, we must decide whether the lead elevation should be corrected or not. One important point to stress is that, no matter what correction we choose, as long as it is applied universally to all our data, then the viability of the methods discussed in this thesis still holds as this error would be systematic.

For a given image, there will either be both a lead and thin ice elevation, just one of the two, or neither. Images that have neither obviously are not used for compiling this list. If there is both a lead and thin ice elevation, if the lead elevation is below the thin ice elevation then this is taken to be the best estimate. If the thin ice elevation is lower, then we take that as the best estimate and subtract 2 cm from it to account for a typical thin ice freeboard following Kurtz et al. (2012). If there is only a thin ice elevation, we subtract 2 cm from it. If there is only a lead elevation, then we assume that this lead elevation is faulty, and we find the typical offset for other leads in a neighborhood of ± 60 images that have thin ice lower than the lead. This average is subtracted from the lead elevation to essentially estimate what the thin ice return would probably be, if there were thin ice in the image and it were the case that the thin ice was lower than the lead elevation (which is more likely than

not). We then subtract 2 cm as before from this 'thin ice elevation estimate' and use this as the estimate of the lead elevation in the image. Finally, these are flagged as possibly problematic elevations. Note that this assumption is not necessarily true, and is accounted for in the final quality control step in the next paragraph. As an example of this last case, if an image has no thin ice and a lead elevation of -2.50 m, and the average difference of images where the thin ice elevation is below the lead elevation is 10 cm, then we subtract 10 cm from the lead elevation to get a thin ice elevation estimate of -2.60m, and then a further 2 cm to account for the thin ice freeboard and the final lead estimate is hence -2.62 m.

The final quality control step is to check the self-consistency of the lead elevations. To do this, we take each lead estimate in turn, and estimate what the local lead elevation would be if interpolating from all other recorded leads within a ± 5 km area. First, we check if this interpolation can be done without using any of the flagged leads from the previous paragraph (i.e. the leads that have no thin ice in the image). If this is possible (i.e. if there are some leads within 5 km that are not flagged), then we perform the interpolation with without the flagged leads, so we only use the flagged leads if no other leads are available. This interpolation is done using an inverse-distance weighting. We choose 5 km as this corresponds to the first Rossby radius of deformation at polar latitudes, which characterizes the 'natural' scale of eddies and fronts and ensures that the non-linear variation of lead heights due to the rotation of the earth is kept minimal (Chelton et al., 1998). If the difference between the interpolation and the recorded lead elevation exceeds 10 cm, the lead is discarded. This process is iterated until the number of lead stabilizes. In this way, if a flagged lead actually had the true lead elevation (and hence the offset-correction was erroneously applied), then the flagged lead would be discarded as an error. Using this method, we find that our RMS error for the lead referencing is < 5 cm, with a median absolute error of < 3 cm. This means that more than 50% of leads have accuracies of better than 3 cm, on par with Kwok and Kacimi (2018).

We can now use this final compiled list of leads to convert the geoid-referenced surface elevations from the OIB ATM lidar scans to lead-referenced surface elevations

Table B.1: Comparing our along-flight average snow freeboard (F) with those from Wang et al. (2020b). There are minor sampling differences between the two, as our method only includes lidar points that are both within 5 km of a lead (in order to be lead-referenced) as well as within a lidar window that has less than 15% open water (for the lidar window interpolation). This should lead to a slightly positive bias for our mean F as compared to Wang et al. (2020b).

Flight Track	Mean F	
	Wang et al.	Our method
2010A	0.89 m	0.86 m
2011A	0.53 m	0.53 m
2011X	0.54 m	0.56 m
2012A	0.48 m	0.52 m
2014A	0.63 m	0.60 m
2014X	0.57 m	0.53 m
2016B	0.91 m	0.98 m

(F) and make lidar windows as in Fig. 2-11.

A comparison of mean F with Wang et al. (2020b), which uses a reflectivity-based method to detect leads, is shown in Table B.2, showing good agreement.

Bibliography

U. Adolphs. Ice thickness variability, isostatic balance and potential for snow ice formation on ice floes in the south polar Pacific Ocean. *Journal of Geophysical Research: Oceans*, 103(C11):24675–24691, 1998. doi: 10.1029/98jc02414.

H. Akaike. A New Look at the Statistical Model Identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer, 1974. doi: 10.1007/978-1-4612-1694-0-16.

W. Aldenhoff, C. Heuz'e, and L. E. B. Eriksson. Sensitivity of Radar Altimeter Waveform to Changes in Sea Ice Type at Resolution of Synthetic Aperture Radar. *Remote Sensing*, 11(22):2602, Nov. 2019. ISSN 2072-4292. doi: 10.3390/rs11222602.

I. Allison et al. Antarctic sea ice growth and oceanic heat flux. *Sea level, ice and climatic change*, 131:161–170, 1981.

S. Arndt and S. Paul. Variability of Winter Snow Properties on Different Spatial Scales in the Weddell Sea. *Journal of Geophysical Research: Oceans*, 123(12): 8862–8876, 2018. doi: 10.1029/2018JC014447.

P. Baldi. Autoencoders, Unsupervised Learning, and Deep Architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

J. Beck, K. Prazdny, and A. Rosenfeld. A Theory of Textural Segmentation. In *Human and machine vision*, pages 1–38. Elsevier, 1983. doi: 10.1016/B978-0-12-084320-6.50007-4.

A. Behrendt, W. Dierking, E. Fahrbach, and H. Witte. Sea Ice Draft in the Weddell Sea, Measured by Upward Looking Sonars. *Earth Syst. Sci. Data*, 5(1):209–226, 2013. doi: 10.5194/essd-5-209-2013.

A. Braakmann-Folgmann and C. Donlon. Estimating snow depth on arctic sea ice using satellite microwave radiometry and a neural network. *The Cryosphere*, 13(9): 2421–2438, 2019. doi: 10.5194/tc-13-2421-2019.

J. C. Brock, C. W. Wright, T. D. Clayton, and A. Nayegandhi. Lidar Optical Rugosity of Coral Reefs in Biscayne National Park, Florida. *Coral Reefs*, 23(1):48–59, 2004. doi: 10.1007/s00338-003-0365-7.

D. J. Cavalieri and C. L. Parkinson. Arctic sea ice variability and trends, 1979-2010. *The Cryosphere*, 6(4):881, 2012. doi: 10.1029/2007jc004558.

D. B. Chelton, R. A. DeSzoeke, M. G. Schlax, K. El Naggar, and N. Siwertz. Geographical Variability of the First Baroclinic Rossby Radius of Deformation. *Journal of Physical Oceanography*, 28(3):433–460, 1998. doi: 10.1175/1520-0485(1998)028<0433:gvotfb>2.0.co;2.

J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz. Adaptive Perceptual Color-texture Image Segmentation. *IEEE Transactions on Image Processing*, 14 (10):1524–1536, Oct. 2005. doi: 10.1109/TIP.2005.852204.

D. A. Clausi and B. Yue. Comparing Cooccurrence Probabilities and Markov Random Fields for Texture Analysis of SAR Sea Ice Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 42(1):215–228, 2004. doi: 10.1109/TGRS.2003.817218.

A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal Bases of Compactly Supported Wavelets. *Communications on Pure and Applied Mathematics*, 45(5): 485–560, 1992. doi: 10.1002/cpa.3160450502.

J. C. Comiso and F. Nishio. Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data. *Journal of Geophysical Research: Oceans*, 113 (C2), 2008. doi: 10.1029/2007jc004257.

J. C. Comiso, R. A. Gersten, L. V. Stock, J. Turner, G. J. Perez, and K. Cho. Positive Trend in the Antarctic Sea Ice Cover and Associated Changes in Surface Temperature. *Journal of Climate*, 30(6):2251–2267, 03 2017. ISSN 0894-8755. doi: 10.1175/JCLI-D-16-0408.1.

W. Dierking. Laser Profiling of the Ice Surface Topography during the Winter Weddell Gyre Study 1992. *Journal of Geophysical Research: Oceans*, 100(C3):4807–4820, 1995. doi: 10.1029/94jc01938.

R. Dominguez. IceBridge DMS L1B Geolocated and Orthorectified Images, Version 1 [2009-2014] (updated 2015), 2010.

H. Eicken and M. Salganek. *Field Techniques for Sea-ice Research.* University of Alaska Press, 2010.

O.-C. Ekeberg, K. Høyland, and E. Hansen. Ice Ridge Keel Geometry and Shape Derived from One Year of Upward Looking Sonar Data in the Fram Strait. *Cold Regions Science and Technology*, 109:78–86, 2015. doi: 10.1016/j.coldregions.2014.10.003.

G. D. Evangelidis and E. Z. Psarakis. Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, Oct 2008. ISSN 1939-3539. doi: 10.1109/TPAMI.2008.113.

S. L. Farrell, N. Kurtz, L. N. Connor, B. C. Elder, C. Leuschen, T. Markus, D. C. McAdoo, B. Panzer, J. Richter-Menge, and J. G. Sonntag. A First Assessment of IceBridge Snow and Ice Thickness Data Over Arctic Sea Ice. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2098–2111, June 2012. ISSN 1558-0644. doi: 10.1109/TGRS.2011.2170843.

S. Filhol and M. Sturm. Snow Bedforms: A Review, New Data, and a Formation Model. *Journal of Geophysical Research: Earth Surface*, 120(9):1645–1669, 2015. doi: 10.1002/2015JF003529.

S. W. Fons and N. T. Kurtz. Retrieval of Snow Freeboard of Antarctic Sea Ice Using Waveform Fitting of CryoSat-2 Returns. *The Cryosphere*, 13(3):861–878, 2019. doi: 10.5194/tc-13-861-2019.

A. Fraser, T. Toyota, P. Jansen, N. Kimura, J. Lieser, G. Williams, E. Trujillo, K. Leonard, T. Maksym, and R. Massom. Satellite remote sensing of Antarctic sea-ice roughness using scatterometer data. In *International Symposium on Sea Ice in a Changing Environment*, page 69A779, 2014.

W. T. Freeman and E. H. Adelson. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(09):891–906, 1991. ISSN 1939-3539. doi: 10.1109/34.93808.

T. Geldsetzer, J. B. Mead, J. J. Yackel, R. K. Scharien, and S. E. L. Howell. Surface-Based Polarimetric C-Band Scatterometer for Field Measurements of Sea Ice. *IEEE Transactions on Geoscience and Remote Sensing*, 45(11):3405–3416, 2007. doi: 10.1109/tgrs.2007.907043.

K. Giles, S. Laxon, D. Wingham, D. Wallis, W. Krabill, C. Leuschen, D. McAdoo, S. Manizade, and R. Raney. Combined Airborne Laser and Radar Altimeter Measurements Over the Fram Strait in May 2002. *Remote Sensing of Environment*, 111(2):182–194, 2007. ISSN 0034-4257. doi: 10.1016/j.rse.2007.02.037. Remote Sensing of the Cryosphere Special Issue.

S. Goebell. Comparison of Coincident Snow-freeboard and Sea Ice Thickness Profiles Derived from Helicopter-borne Laser Altimetry and Electromagnetic Induction Sounding. *Journal of Geophysical Research: Oceans*, 116(C8), 2011. doi: 10.1029/2009jc006055.

H. Goosse and T. Fichefet. Importance of ice-ocean interactions for the global ocean circulation: A model study. *Journal of Geophysical Research: Oceans*, 104(C10): 23337–23355, 1999. doi: 10.1029/1999jc900215.

M. Gupta, D. G. Barber, R. K. Scharien, and D. Isleifson. Detection and classification of surface roughness in an Arctic marginal sea ice zone. *Hydrological Processes*, 28 (3):599–609, 2014. doi: 10.1002/hyp.9593.

C. Haas. Evaluation of Ship-based Electromagnetic-inductive Thickness Measurements of Summer Sea-ice in the Bellingshausen and Amundsen Seas, Antarctica. *Cold Regions Science and Technology*, 27(1):1–16, 1998. doi: 10.1016/s0165-232x(97)00019-0.

C. Haas, Q. Liu, and T. Martin. Retrieval of Antarctic Sea-ice Pressure Ridge Frequencies from ERS SAR Imagery by Means of in Situ Laser Profiling and Usage of a Neural Network. *International Journal of Remote Sensing*, 20(15-16):3111–3123, 1999. doi: 10.1080/014311699211642.

C. Haas, J. Lobach, S. Hendricks, L. Rabenstein, and A. Pfaffling. Helicopter-borne Measurements of Sea Ice Thickness, Using a Small and Lightweight, Digital EM System. *Journal of Applied Geophysics*, 67(3):234–241, 2009. doi: 10.1016/j.jappgeo.2008.05.005.

S. Harms, E. Fahrbach, and V. H. Strass. Sea Ice Transports in the Weddell Sea. *Journal of Geophysical Research: Oceans*, 106(C5):9057–9073, 2001. doi: 10.1029/1999jc000027.

F. A. Haumann, N. Gruber, M. Münnich, I. Frenger, and S. Kern. Sea-ice Transport Driving Southern Ocean Salinity and Its Recent Trends. *Nature*, 537(7618):89, 2016. doi: 10.1038/nature19101.

M. M. Holland, J. A. Curry, and J. L. Schramm. Modeling the thermodynamics of a sea ice thickness distribution: 2. Sea ice/ocean interactions. *Journal of Geophysical Research: Oceans*, 102(C10):23093–23107, 1997. doi: 10.1029/97JC01296.

M. M. Holland, C. M. Bitz, and A. Weaver. The influence of sea ice physics on simulations of climate change. *Journal of Geophysical Research: Oceans*, 106(C9):19639–19655, 2001. doi: 10.1029/2000jc000651.

M. M. Holland, C. M. Bitz, E. C. Hunke, W. H. Lipscomb, and J. L. Schramm. Influence of the Sea Ice Thickness Distribution on Polar Climate in CCSM3. *Journal of Climate*, 19(11):2398–2414, 2006. doi: 10.1175/jcli3751.1.

P. R. Holland. The seasonality of Antarctic sea ice trends. *Geophysical Research Letters*, 41(12):4230–4237, 2014. doi: 10.1002/2014gl060172.

P. R. Holland and R. Kwok. Wind-driven trends in antarctic sea-ice drift. *Nature Geoscience*, 5(12):872–875, 2012. doi: 10.1038/ngeo1627.

P. R. Holland, N. Bruneau, C. Enright, M. Losch, N. T. Kurtz, and R. Kwok. Modeled Trends in Antarctic Sea Ice Thickness. *Journal of Climate*, 27(10):3784–3801, 2014. doi: 10.1175/jcli-d-13-00301.1.

B. Holt, M. P. Johnson, D. Perkovic-Martin, and B. Panzer. Snow depth on Arctic sea ice derived from radar: In situ comparisons and time series analysis. *Journal of Geophysical Research: Oceans*, 120(6):4260–4287, 2015. doi: 10.1002/

2015JC010815. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC010815.

J. R. M. Hosking. L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society: Series B (methodological)*, 52(1):105–124, 1990. URL http://www.jstor.org/stable/2345653.

J. K. Hutchings, P. Heil, O. Lecomte, R. Stevens, A. Steer, and J. L. Lieser. Comparing Methods of Measuring Sea-ice Density in the East Antarctic. *Annals of Glaciology*, 56(69):77–82, 2015. doi: 10.3189/2015aog69a814.

A. K. Jain and F. Farrokhnia. Unsupervised Texture Segmentation Using Gabor Filters. In *1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings*, pages 14–19, Nov. 1990. doi: 10.1109/ICSMC.1990.142050.

S. Jenouvrier, C. Barbraud, and H. Weimerskirch. Sea Ice Affects the Population Dynamics of Adélie Penguins in Terre Adélie. *Polar Biology*, 29(5):413–423, 2006. doi: 10.1007/s00300-005-0073-6.

L. Kaleschke, F. Girard-Ardhuin, G. Spreen, A. Beitsch, and S. Kern. ASI Algorithm SSMI-SSMIS Sea Ice Concentration Data, Originally Computed at and Provided by Ifremer, Brest, France, Were Obtained As 5-day Median-filtered and Gap-filled Product for 2017/06/02 from the Integrated Climate Date Center (ICDC, icdc.cen.uni-hamburg.de/), 2017.

P. Kanagaratnam, T. Markus, V. Lytle, B. Heavey, P. Jansen, G. Prescott, and S. P. Gogineni. Ultrawideband Radar Measurements of Thickness of Snow Over Sea Ice. *IEEE Transactions on Geoscience and Remote Sensing*, 45(9):2715–2724, 2007. doi: 10.1109/IGARSS.2010.5654342.

S. Kern and G. Spreen. Uncertainties in Antarctic Sea-ice Thickness Retrieval from ICESat. *Annals of Glaciology*, 56(69):107–119, 2015. doi: 10.3189/2015aog69a736.

S. Kern, B. Özsoy-Çiçek, and A. Worby. Antarctic Sea-ice Thickness Retrieval from ICESat: Inter-comparison of Different Approaches. *Remote Sensing*, 8(7):538, 2016. doi: 10.3390/rs8070538.

J. Kim, D. Kim, and B. J. Hwang. Characterization of Arctic Sea Ice Thickness Using High-resolution Spaceborne Polarimetric SAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 50(1):13–22, Jan. 2012. ISSN 1558-0644. doi: 10.1109/TGRS.2011.2160070.

N. Kimura and M. Wakatsuchi. Large-scale processes governing the seasonal variability of the Antarctic sea ice. *Tellus A: Dynamic Meteorology and Oceanography*, 63 (4):828–840, 2011. doi: 10.3402/tellusa.v63i4.15860.

J. King. Climate science: A resolution of the Antarctic paradox. *Nature*, 505(7484): 491–492, 2014. doi: 10.1038/505491a.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *Arxiv Preprint Arxiv:1412.6980*, 2014. URL `https://arxiv.org/abs/1412.6980`.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing Neural Networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 00034851. doi: 10.1214/aoms/1177729694.

N. Kurtz. IceBridge Quick Look Sea Ice Freeboard, Snow Depth, and Thickness Product Manual for 2013. *Boulder, Colorado Usa: NASA DAAC at the National Snow and Ice Data Center*, 2013.

N. Kurtz and T. Markus. Satellite Observations of Antarctic Sea Ice Thickness and Volume. *Journal of Geophysical Research: Oceans*, 117(C8), 2012. doi: 10.1029/2012jc008141.

N. T. Kurtz and S. L. Farrell. Large-scale Surveys of Snow Depth on Arctic Sea Ice from Operation IceBridge. *Geophysical Research Letters*, 38(20), 2011. doi: 10.1029/2011gl049216.

N. T. Kurtz, S. L. Farrell, L. S. Koenig, M. Studinger, J. P. Harbeck, et al. A Sea-ice Lead Detection Algorithm for Use with High-resolution Airborne Visible Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):38–56, 2012. doi: 10.1109/tgrs.2012.2202666.

R. Kwok. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, 13(10):105005, 2018. doi: 10.1088/1748-9326/aae3ec.

R. Kwok and C. Haas. Effects of Radar Side-lobes on Snow Depth Retrievals from Operation IceBridge. *Journal of Glaciology*, 61(227):576–584, 2015. doi: 10.3189/2015JoG14J229.

R. Kwok and S. Kacimi. Three Years of Sea Ice Freeboard, Snow Depth, and Ice Thickness of the Weddell Sea from Operation IceBridge and CryoSat-2. *The Cryosphere*, 12(8):2789–2801, 2018. doi: 10.5194/tc-12-2789-2018.

R. Kwok and T. Maksym. Snow Depth of the Weddell and Bellingshausen Sea Ice Covers from IceBridge Surveys in 2010 and 2011: An Examination. *Journal of Geophysical Research: Oceans*, 119(7):4141–4167, 2014. doi: 10.1002/2014jc009943.

R. Kwok and D. Rothrock. Decline in Arctic Sea Ice Thickness from Submarine and ICESat Records: 1958–2008. *Geophysical Research Letters*, 36(15), 2009. doi: 10.1029/2009gl039035.

R. Kwok, B. Panzer, C. Leuschen, S. Pang, T. Markus, B. Holt, and S. Gogineni. Airborne Surveys of Snow Depth Over Arctic Sea Ice. *Journal of Geophysical Research: Oceans*, 116(C11), 2011. doi: 10.1029/2011jc007371.

R. Kwok, G. F. Cunningham, S. S. Manizade, and W. B. Krabill. Arctic sea ice freeboard from icebridge acquisitions in 2009: Estimates and comparisons with ICESat. *Journal of Geophysical Research: Oceans*, 117(C2), 2012. doi: 10.1029/2011JC007654.

R. Kwok, N. T. Kurtz, L. Brucker, A. Ivanoff, T. Newman, S. L. Farrell, J. King, S. Howell, M. A. Webster, J. Paden, C. Leuschen, J. A. MacGregor, J. Richter-Menge, J. Harbeck, and M. Tschudi. Intercomparison of snow depth retrievals over Arctic sea ice from radar data acquired by Operation IceBridge. *The Cryosphere*, 11(6):2571–2593, 2017. doi: 10.5194/tc-11-2571-2017. URL `https://tc.copernicus.org/articles/11/2571/2017/`.

M. Lange, P. Schlosser, S. Ackley, P. Wadhams, and G. Dieckmann. δ18O concentrations in sea ice of the Weddell Sea. *J. Glaciol*, 36:315–323, 1990.

T. S. Ledley. Snow on Sea Ice: Competing Effects in Shaping Climate. *Journal of Geophysical Research: Atmospheres*, 96(D9):17195–17208, 1991. doi: 10.1029/91jd01439.

A. LeNail. NN-svg: Publication-ready Neural Network Architecture Schematics. *The Journal of Open Source Software*, 4:747, 2019. doi: 10.21105/joss.00747.

M. Leppäranta. A review of analytical models of sea-ice growth. *Atmosphere-Ocean*, 31(1):123–138, 1993. doi: 10.1080/07055900.1993.9649465.

M. Leppäranta and R. Hakala. The Structure and Strength of First-year Ice Ridges in the Baltic Sea. *Cold Regions Science and Technology*, 20(3):295–311, 1992. doi: 10.1016/0165-232x(92)90036-t.

H. Li, H. Xie, S. Kern, W. Wan, B. Özsoy, S. Ackley, and Y. Hong. Spatio-temporal Variability of Antarctic Sea-ice Thickness and Volume Obtained from ICESat Data Using an Innovative Algorithm. *Remote Sensing of Environment*, 219:44–61, 2018. doi: 10.1016/j.rse.2018.09.031.

S. Li and A. B. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014. doi: 10.1007/978-3-319-16808-1_23.

D. Lindell and D. Long. Multiyear Arctic Ice Classification Using ASCAT and SSMIS. *Remote Sensing*, 8(4):294, Mar 2016. ISSN 2072-4292. doi: 10.3390/rs8040294.

R. D. Lindsley and D. G. Long. Enhanced-Resolution Reconstruction of ASCAT Backscatter Measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):2589–2601, 2016. doi: 10.1109/tgrs.2015.2503762.

J. Liu, Y. Zhang, X. Cheng, and Y. Hu. Retrieval of Snow Depth Over Arctic Sea Ice Using a Deep Neural Network. *Remote Sensing*, 11(23):2864, 2019. doi: 10.3390/rs11232864.

W. Liu, A. V. Fedorov, S.-P. Xie, and S. Hu. Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate. *Science Advances*, 6(26): eaaz4876, 2020. doi: 10.1126/sciadv.aaz4876.

I. Loshchilov and F. Hutter. Fixing Weight Decay Regularization in Adam, 2018. URL `https://openreview.net/forum?id=rk6qdGgCZ`.

J. Ludescher, N. Yuan, and A. Bunde. Detecting the statistical significance of the trends in the Antarctic sea ice extent: an indication for a turning point. *Climate dynamics*, 53(1-2):237–244, 2019. doi: 10.1007/s00382-018-4579-3.

L. v. d. Maaten and G. Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

T. Maksym and M. O. Jeffries. Phase and compositional evolution of the flooded layer during snow-ice formation on Antarctic sea ice. *Annals of Glaciology*, 33: 37–44, 2001. doi: 10.3189/172756401781818860.

T. Maksym and T. Markus. Antarctic Sea Ice Thickness and Snow-to-ice Conversion from Atmospheric Reanalysis and Passive Microwave Snow Depth. *Journal of Geophysical Research: Oceans*, 113(C2), 2008. doi: 10.1029/2006jc004085.

R. D. C. Mallett, I. R. Lawrence, J. C. Stroeve, J. C. Landy, and M. Tsamados. Brief Communication: Conventional Assumptions Involving the Speed of Radar Waves in Snow Introduce Systematic Underestimates to Sea Ice Thickness and Seasonal Growth Rate Estimates. *The Cryosphere*, 14(1):251–260, 2020. doi: 10.5194/tc-14-251-2020.

T. Markus and D. J. Cavalieri. Snow Depth Distribution Over Sea Ice in the Southern Ocean from Satellite Passive Microwave Data. *Antarctic Sea Ice: Physical Processes, Interactions and Variability*, pages 19–39, 1998. doi: 10.1029/ar074p0019.

T. Markus, R. Massom, A. Worby, V. Lytle, N. Kurtz, and T. Maksym. Freeboard, Snow Depth and Sea-ice Roughness in East Antarctica from in Situ and Multiple Satellite Data. *Annals of Glaciology*, 52(57):242–248, 2011. doi: 10.3189/172756411795931570.

T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding, M. Jasinski, R. Kwok, L. Magruder, D. Lubin, S. Luthcke, J. Morison, R. Nelson, A. Neuenschwander, S. Palm, S. Popescu, C. Shum, B. E. Schutz, B. Smith, Y. Yang, and J. Zwally. The Ice, Cloud, and Land Elevation Satellite-2 (ICESat-2): Science Requirements, Concept, and Implementation. *Remote Sensing of Environment*, 190:260–273, 2017. doi: 10.1016/j.rse.2016.12.029.

J. Marshall and K. Speer. Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nature Geoscience*, 5(3):171–180, 2012. doi: 10.1038/ngeo1391.

R. A. Massom and S. E. Stammerjohn. Antarctic sea ice change and variability–physical and ecological implications. *Polar Science*, 4(2):149–186, 2010. doi: 10.1016/j.polar.2010.05.001.

R. A. Massom, M. R. Drinkwater, and C. Haas. Winter Snow Cover on Sea Ice in the Weddell Sea. *Journal of Geophysical Research: Oceans*, 102(C1):1101–1117, 1997. doi: 10.1029/96JC02992.

R. A. Massom, H. Eicken, C. Hass, M. O. Jeffries, M. R. Drinkwater, M. Sturm, A. P. Worby, X. Wu, V. I. Lytle, S. Ushio, K. Morris, P. A. Reid, S. G. Warren, and I. Allison. Snow on Antarctic Sea Ice. *Reviews of Geophysics*, 39(3):413–445, 2001. doi: 10.1029/2000RG000085.

F. Massonnet, P. Mathiot, T. Fichefet, H. Goosse, C. K. Beatty, M. Vancoppenolle, and T. Lavergne. A Model Reconstruction of the Antarctic Sea Ice Thickness and Volume Changes Over 1980–2008 Using Data Assimilation. *Ocean Modelling*, 64:67–75, 2013. doi: 10.1016/j.ocemod.2013.01.003.

G. A. Maykut. Large-scale heat exchange and ice production in the central Arctic. *Journal of Geophysical Research: Oceans*, 87(C10):7971–7984, 1982. doi: 10.1029/jc087ic10p07971.

M. J. Mei and T. Maksym. A Textural Approach to Improving Snow Depth Estimates in the Weddell Sea. *Remote Sensing*, 12(9), 2020. ISSN 2072-4292. doi: 10.3390/rs12091494.

M. J. Mei, T. Maksym, B. Weissling, and H. Singh. Estimating Early-winter Antarctic Sea Ice Thickness from Deformed Ice Morphology. *The Cryosphere*, 13(11):2915–2934, 2019. doi: 10.5194/tc-13-2915-2019.

H. Melling and D. A. Riedel. Development of Seasonal Pack Ice in the Beaufort Sea during the Winter of 1991–1992: A View from Below. *Journal of Geophysical Research: Oceans*, 101(C5):11975–11991, 1996. doi: 10.1029/96jc00284.

J. Mortin, S. E. Howell, L. Wang, C. Derksen, G. Svensson, R. G. Graversen, and T. M. Schrøder. Extending the QuikSCAT record of seasonal melt-freeze transitions over Arctic sea ice using ASCAT. *Remote Sensing of Environment*, 141:214 – 230, 2014. ISSN 0034-4257. doi: 10.1016/j.rse.2013.11.004.

O. Mussells, J. Dawson, and S. Howell. Navigating pressured ice: Risks and hazards for winter resource-based shipping in the Canadian Arctic. *Ocean & Coastal Management*, 137:57 – 67, 2017. ISSN 0964-5691. doi: 10.1016/j.ocecoaman.2016.12.010.

T. Newman, S. L. Farrell, J. Richter-Menge, L. N. Connor, N. T. Kurtz, B. C. Elder, and D. McAdoo. Assessment of radar-derived snow depth over Arctic sea ice. *Journal of Geophysical Research: Oceans*, 119(12):8578–8602, 2014. doi: 10.1002/2014JC010284. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JC010284.

S. V. Nghiem, R. Kwok, S. H. Yueh, and M. R. Drinkwater. Polarimetric signatures of sea ice: 1. Theoretical model. *Journal of Geophysical Research: Oceans*, 100 (C7):13665–13679, 1995. doi: 10.1029/95JC00937.

V. Onana, N. T. Kurtz, S. L. Farrell, L. S. Koenig, M. Studinger, and J. P. Harbeck. A Sea-ice Lead Detection Algorithm for Use with High-resolution Airborne Visible Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):38–56, Jan. 2013. doi: 10.1109/TGRS.2012.2202666.

A. H. Orsi, G. C. Johnson, and J. L. Bullister. Circulation, mixing, and production of Antarctic Bottom Water. *Progress in Oceanography*, 43(1):55–109, 1999. doi: 10.1016/s0079-6611(99)00004-x.

B. Özsoy-Çiçek, S. Kern, S. F. Ackley, H. Xie, and A. E. Tekeli. Intercomparisons of Antarctic Sea Ice Types from Visual Ship, Radarsat-1 SAR, ENVISAT ASAR, QUIKSCAT, and AMSR-E Satellite Observations in the Bellingshausen Sea. *Deep Sea Research Part Ii: Topical Studies in Oceanography*, 58(9-10):1092–1111, 2011. doi: 10.1016/j.dsr2.2010.10.031.

B. Özsoy-Çiçek, S. Ackley, H. Xie, D. Yi, and J. Zwally. Sea Ice Thickness Retrieval Algorithms Based On In Situ Surface Elevation and Thickness Values for Application to Altimetry. *Journal of Geophysical Research: Oceans*, 118(8):3807–3822, 2013. doi: 10.1002/jgrc.20252.

J. Paden, J. Li, C. Leuschen, R.-M. F., and R. Hale. IceBridge Snow Radar L1b Geolocated Radar Echo Strength Profiles, Version 2. (updated 2019), 2014.

B. Panzer, D. Gomez-Garcia, C. Leuschen, J. Paden, F. Rodriguez-Morales, A. Patel, T. Markus, B. Holt, and P. Gogineni. An Ultra-wideband, Microwave Radar for Measuring Snow Thickness on Sea Ice and Mapping Near-surface Internal Layers in Polar Firn. *Journal of Glaciology*, 59(214):244–254, 2013. doi: 10.3189/2013JoG12J128.

C. Parkinson and D. Cavalieri. Antarctic Sea Ice Variability and Trends, 1979–2010. *The Cryosphere*, 6(4):871–880, 2012. doi: 10.5194/tc-6-871-2012.

C. L. Parkinson. A 40-y record reveals gradual Antarctic sea ice increases followed by decreases at rates far exceeding the rates seen in the Arctic. *Proceedings of the National Academy of Sciences*, 116(29):14414–14423, 2019. doi: 10.1073/pnas.1906556116.

A. A. Petty, M. C. Tsamados, N. T. Kurtz, S. L. Farrell, J. P. Harbeck, D. L. Feltham, and J. A. Richter-Menge. Characterizing Arctic Sea Ice Topography Using High-resolution IceBridge Data. *The Cryosphere*, 10(3):1161, 2016. doi: 10.5194/tc-10-1161-2016.

T. Phillips. Cyclone tracks for the region south of 60S for 1979 - 2018 derived from 6-hourly ERA-Interim reanalysis mean sea level pressure (MSLP) fields. *UK Polar Data Centre, Natural Environment Research Council, UK Research & Innovation*, 2020. doi: 10.5285/D3B5D87D-C882-4FED-9D47-14C73BE43BCA.

L. M. Polvani and K. L. Smith. Can natural variability explain observed antarctic sea ice trends? New modeling evidence from CMIP5. *Geophysical Research Letters*, 40 (12):3195–3199, 2013. doi: 10.1002/grl.50578.

M. Porat and Y. Y. Zeevi. Localized Texture Processing in Vision: Analysis and Synthesis in the Gaborian Space. *IEEE Transactions on Biomedical Engineering*, 36(1):115–129, 1989. doi: 10.1109/10.16457.

R. Ressel, A. Frost, and S. Lehner. A Neural Network-based Classification for Sea Ice Types on X-band SAR Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3672–3680, July 2015. ISSN 2151-1535. doi: 10.1109/JSTARS.2015.2436993.

D. Rothrock, D. Percival, and M. Wensnahan. The Decline in Arctic Sea-ice Thickness: Separating the Spatial, Annual, and Interannual Variability in a Quarter Century of Submarine Data. *Journal of Geophysical Research: Oceans*, 113(C5), 2008. doi: 10.1029/2007jc004252.

S. Schmidtko, K. J. Heywood, A. F. Thompson, and S. Aoki. Multidecadal warming of Antarctic waters. *Science*, 346(6214):1227–1231, 2014. doi: 10.1126/science.1256117.

J. Schramm, M. Holland, J. Curry, and E. Ebert. Modeling the thermodynamics of a sea ice thickness distribution: 1. Sensitivity to ice thickness resolution. *Journal of Geophysical Research: Oceans*, 102(C10):23079–23091, 1997. doi: 10.1029/97jc01297.

L. Shi, J. Karvonen, B. Cheng, T. Vihma, M. Lin, Y. Liu, Q. Wang, and Y. Jia. Sea Ice Thickness Retrieval from SAR Imagery Over Bohai Sea. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 4864–4867, July 2014. doi: 10.1109/IGARSS.2014.6947584.

Q. Shu, Z. Song, and F. Qiao. Assessment of Sea Ice Simulations in the CMIP5 Models. *The Cryosphere*, 9(1):399–409, 2015. doi: 10.5194/tc-9-399-2015.

R. Sibson. A Brief Description of Natural Neighbour Interpolation. *Interpreting Multivariate Data*, 1981.

E. P. Simoncelli and W. T. Freeman. The Steerable Pyramid: A Flexible Architecture for Multi-scale Derivative Computation. In *Proceedings., International Conference on Image Processing*, volume 3, pages 444–447. IEEE, 1995. doi: 10.1109/icip.1995.537667.

E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable Multiscale Transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992. doi: 10.1109/18.119725.

L.-K. Soh and C. Tsatsoulis. Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-occurrence Matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):780–795, 1999. doi: 10.1109/36.752194.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

S. Stammerjohn, D. Martinson, R. Smith, X. Yuan, and D. Rind. Trends in Antarctic Annual Sea Ice Retreat and Advance and Their Relation to El Niño–Southern Oscillation and Southern Annular Mode Variability. *Journal of Geophysical Research: Oceans*, 113(C3), 2008. doi: 10.1029/2007jc004269.

S. Stammerjohn, R. Massom, D. Rind, and D. Martinson. Regions of rapid sea ice change: An inter-hemispheric seasonal comparison. *Geophysical Research Letters*, 39(6), 2012. doi: 10.1029/2012gl050874.

S. Stammerjohn, T. Maksym, R. Massom, K. Lowry, K. Arrigo, X. Yuan, M. Raphael, E. Randall-Goodwin, R. Sherrell, and P. Yager. Seasonal sea ice changes in the Amundsen Sea, Antarctica, over the period of 1979–2014. *Elem Sci Anth*, 3, 2015. doi: 10.12952/journal.elementa.000055.

A. Steer, P. Heil, C. Watson, R. A. Massom, J. L. Lieser, and B. Özsoy-Çiçek. Estimating Small-scale Snow Depth and Ice Thickness from Total Freeboard for East Antarctic Sea Ice. *Deep Sea Research Part Ii: Topical Studies in Oceanography*, 131:41–52, 2016. doi: 10.1016/j.dsr2.2016.04.025.

J. C. Stroeve, T. Markus, J. A. Maslanik, D. J. Cavalieri, A. J. Gasiewski, J. F. Heinrichs, J. Holmgren, D. K. Perovich, and M. Sturm. Impact of Surface Roughness on AMSR-E Sea Ice Products. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3103–3117, 2006. doi: 10.1109/TGRS.2006.880619.

J. C. Stroeve, V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier. Trends in Arctic Sea Ice Extent from CMIP5, CMIP3 and Observations. *Geophysical Research Letters*, 39(16), 2012. doi: 10.1029/2012gl052676.

L. Strub-Klein and D. Sudom. A Comprehensive Analysis of the Morphology of First-year Sea Ice Ridges. *Cold Regions Science and Technology*, 82:94–109, 2012. doi: 10.1016/j.coldregions.2012.05.014.

M. Studinger. IceBridge ATM L1B Elevation and Return Strength, Version 2. *Boulder, Colorado USA: NASA DAAC at the National Snow and Ice Data Center*, 2018. doi: 10.5067/19SIM5TXKPGT.

M. F. Stuecker, C. M. Bitz, and K. C. Armour. Conditions leading to the unprecedented low Antarctic sea ice extent during the 2016 austral spring season. *Geophysical Research Letters*, 44(17):9008–9019, 2017. doi: 10.1002/2017gl074691.

M. Sturm and J. Holmgren. An Automatic Snow Depth Probe for Field Validation Campaigns. *Water Resources Research*, 54(11):9695–9701, 2018. doi: 10.1029/2018wr023559.

M. Sturm and R. A. Massom. Snow and sea ice. *Sea ice*, 2:153–204, 2009. doi: 10.1002/9781444317145.ch5.

M. Sturm, K. Morris, and R. Massom. The Winter Snow Cover of the West Antarctic Pack Ice: Its Spatial and Temporal Variability. *Antarctic Sea Ice: Physical Processes, Interactions and Variability*, pages 1–18, 1998. doi: 10.1029/ar074p0001.

S. Suzuki and K. Abe. Topological Structural Analysis of Digitized Binary Images by Border Following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985. ISSN 0734-189X. doi: 10.1016/0734-189X(85)90016-7.

N. Swart and J. Fyfe. The influence of recent antarctic ice sheet retreat on simulated sea ice area trends. *Geophysical Research Letters*, 40(16):4328–4332, 2013. doi: 10.1002/grl.50820.

B. Tan, L. Wang, P. Lu, Z. Li, and E. Feng. A Novel Strategy to Analyse the Form Drag on Pressure Ridges and the Air–ice Drag Coefficient in the Northwestern Weddell Sea. *Applied Mathematical Modelling*, 58:158–165, 2018. doi: 10.1016/j.apm.2017.09.046.

L. Tian, H. Xie, S. F. Ackley, J. Tang, A. M. Mestas-Nuñez, and X. Wang. Sea-ice freeboard and thickness in the Ross Sea from airborne (IceBridge 2013) and satellite (ICESat 2003–2008) observations. *Annals of Glaciology*, pages 1–16, 2019. doi: 10.1017/aog.2019.49.

G. Timco and R. Burden. An Analysis of the Shapes of Sea Ice Ridges. *Cold Regions Science and Technology*, 25(1):65–77, 1997. doi: 10.1016/s0165-232x(96)00017-1.

G. Timco and R. Frederking. A Review of Sea Ice Density. *Cold Regions Science and Technology*, 24(1):1–6, 1996. doi: 10.1016/0165-232x(95)00007-x.

G. Timco and W. Weeks. A Review of the Engineering Properties of Sea Ice. *Cold Regions Science and Technology*, 60(2):107–129, 2010. doi: 10.1016/j.coldregions.2009.10.003.

G. W. Timco and M. Sayed. Model Tests of the Ridge-building Process in Ice. 1986.

T. Tin and M. Jeffries. Quantitative Identification of Antarctic First Year Pressure Ridges and Preliminary Results on Ridge Morphology. In *Proceedings of the International Conference on Port and Ocean Engineering Under Arctic Conditions*, 2001a.

T. Tin and M. O. Jeffries. Sea-ice Thickness and Roughness in the Ross Sea, Antarctica. *Annals of Glaciology*, 33:187–193, 2001b. doi: 10.3189/172756401781818770.

T. Tin and M. O. Jeffries. Morphology of Deformed First-year Sea Ice Features in the Southern Ocean. *Cold Regions Science and Technology*, 36(1-3):141–163, 2003. doi: 10.1016/s0165-232x(03)00008-9.

M. Tiuri, A. Sihvola, E. Nyfors, and M. Hallikaiken. The Complex Dielectric Constant of Snow at Microwave Frequencies. *IEEE Journal of Oceanic Engineering*, 9(5): 377–382, Dec. 1984. ISSN 2373-7786. doi: 10.1109/JOE.1984.1145645.

E. Trujillo, K. Leonard, T. Maksym, and M. Lehning. Changes in Snow Distribution and Surface Topography Following a Snowstorm on Antarctic Sea Ice. *Journal of Geophysical Research: Earth Surface*, 121(11):2172–2191, 2016. doi: 10.1002/2016JF003893.

M. Tschudi, C. Fowler, J. Maslanik, J. Stewart, and W. Meier. Polar Pathfinder Daily 25 km EASE-Grid Sea Ice Motion Vectors, Version 3. *Boulder, Colorado USA: NASA DAAC at the National Snow and Ice Data Center*, 2016. doi: 10.5067/O57VAIT2AYYY.

W. Tucker III and J. Govoni. Morphological Investigations of First-year Sea Ice Pressure Ridge Sails. *Cold Regions Science and Technology*, 5(1):1–12, 1981. doi: 10.1016/0165-232x(81)90036-7.

W. B. Tucker III, D. S. Sodhi, and J. W. Govoni. Structure of First-year Pressure Ridge Sails in the Prudhoe Bay Region. In *The Alaskan Beaufort Sea*, pages 115–135. Elsevier, 1984. doi: 10.1016/b978-0-12-079030-2.50012-5.

J. Turner, T. J. Bracegirdle, T. Phillips, G. J. Marshall, and J. S. Hosking. An Initial Assessment of Antarctic Sea Ice Extent in the CMIP5 Models. *Journal of Climate*, 26(5):1473–1484, 2013. doi: 10.1175/jcli-d-12-00068.1.

J. Turner, J. S. Hosking, T. J. Bracegirdle, G. J. Marshall, and T. Phillips. Recent changes in Antarctic sea ice. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2045):20140163, 2015. doi: 10.1098/rsta.2014.0163.

J. Turner, T. Phillips, G. J. Marshall, J. S. Hosking, J. O. Pope, T. J. Bracegirdle, and P. Deb. Unprecedented springtime retreat of Antarctic sea ice in 2016. *Geophysical Research Letters*, 44(13):6868–6875, 2017. doi: 10.1002/2017gl073656.

J. Turner, M. V. Guarino, J. Arnatt, B. Jena, G. J. Marshall, T. Phillips, C. Bajish, K. Clem, Z. Wang, T. Andersson, et al. Recent decrease of summer sea ice in the Weddell Sea, Antarctica. *Geophysical Research Letters*, page e2020GL087127, 2020. doi: 10.1029/2020gl087127.

P. Uotila, P. R. Holland, T. Vihma, S. J. Marsland, and N. Kimura. Is Realistic Antarctic Sea-ice Extent in Climate Models the Result of Excessive Ice Drift? *Ocean Modelling*, 79:33–42, 2014. doi: 10.1016/j.ocemod.2014.04.004.

N. Urabe and M. Inoue. Mechanical Properties of Antarctic Sea Ice. *Journal of Offshore Mechanics and Arctic Engineering*, 110(4):403–408, 1988. doi: 10.1115/1.3257079.

M. Van Den Broeke. The semi-annual oscillation and Antarctic climate. Part 4: a note on sea ice cover in the Amundsen and Bellingshausen Seas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 20(4):455–462, 2000. doi: 10.1002/(sici)1097-0088(20000330)20:4<455::aid-joc482>3.0.co;2-m.

L. Vasershtein. Markov processes over denumerable products of spaces describing large system of automata. *MR0314115 (47# 2667)*, (3):64–72, 1969.

R. Vickers and G. Rose. Short pulse radar measurements of layered ice and snow (Development of radar system for remote measurement of snow and ice thickness). *NASA. Manned Spacecraft Center 4th Ann. Earth Resources Program Rev.*, 3, 1972.

J. Wang, Q. Yuan, H. Shen, T. Liu, T. Li, L. Yue, X. Shi, and L. Zhang. Estimating snow depth by combining satellite data and ground-based observations over Alaska: A deep learning approach. *Journal of Hydrology*, 585:124828, 2020a. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2020.124828.

L. Wang, K. A. Scott, L. Xu, and D. A. Clausi. Sea Ice Concentration Estimation during Melt from Dual-pol SAR Scenes Using Deep Convolutional Neural Networks: A Case Study. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4524–4533, Aug. 2016. ISSN 1558-0644. doi: 10.1109/TGRS.2016.2543660.

X. Wang, H. Xie, Y. Ke, S. F. Ackley, and L. Liu. A Method to Automatically Determine Sea Level for Referencing Snow Freeboards and Computing Sea Ice Thicknesses from Nasa IceBridge Airborne Lidar. *Remote Sensing of Environment*, 131:160–172, 2013. doi: 10.1016/j.rse.2012.12.022.

X. Wang, W. Jiang, H. Xie, S. Ackley, and H. Li. Decadal variations of sea ice thickness in the Amundsen-Bellingshausen and Weddell Seas retrieved from ICESat and IceBridge laser altimetry, 2003-2017. *Journal of Geophysical Research: Oceans*, 125(7):e2020JC016077, 2020b. doi: 10.1029/2020JC016077.

B. Weissling and S. Ackley. Antarctic Sea-ice Altimetry: Scale and Resolution Effects on Derived Ice Thickness Distribution. *Annals of Glaciology*, 52(57):225–232, 2011. doi: 10.3189/172756411795931679.

B. Weissling, M. Lewis, and S. Ackley. Sea-ice Thickness and Mass at Ice Station Belgica, Bellingshausen Sea, Antarctica. *Deep Sea Research Part Ii: Topical Studies in Oceanography*, 58(9):1112–1124, 2011. ISSN 0967-0645. doi: 10.1016/j.dsr2.2010.10.032. Antarctic Sea Ice Research during the International Polar Year 2007-2009.

R. C. Willatt, K. A. Giles, S. W. Laxon, L. Stone-Drake, and A. P. Worby. Field Investigations of Ku-band Radar Penetration into Snow Cover on Antarctic Sea Ice. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1):365–372, 2009. doi: 10.1109/tgrs.2009.2028237.

G. Williams, T. Maksym, J. Wilkinson, C. Kunz, C. Murphy, P. Kimball, and H. Singh. Thick and Deformed Antarctic Sea Ice Mapped with Autonomous Underwater Vehicles. *Nature Geoscience*, 8(1):61–67, 2015. doi: 10.1038/ngeo2299.

G. Williams, D. Turner, T. Maksym, and H. Singh. Near-coincident mapping of sea ice from above and below with UAS and AUV. In *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*, pages 1–6. IEEE, 2018. doi: 10.1109/auv.2018.8729799.

G. D. Williams, T. Maksym, C. Kunz, P. Kimball, H. Singh, J. Wilkinson, T. Lachlan-Cope, E. Trujillo, A. Steer, R. Massom, et al. Beyond Point Measurements: Sea Ice Floes Characterized in 3-D. *Eos, Transactions American Geophysical Union*, 94(7):69–70, 2013. doi: 10.1002/2013eo070002.

D. Wingham, C. Francis, S. Baker, C. Bouzinac, D. Brockley, R. Cullen, P. de Chateau-Thierry, S. Laxon, U. Mallow, C. Mavrocordatos, et al. CryoSat: A Mission to Determine the Fluctuations in Earth's Land and Marine Ice Fields. *Advances in Space Research*, 37(4):841–871, 2006. doi: 10.1016/j.asr.2005.07.027.

A. Worby, M. Jeffries, W. Weeks, K. Morris, and R. Jana. The Thickness Distribution of Sea Ice and Snow Cover during Late Winter in the Bellingshausen and Amundsen Seas, Antarctica. *Journal of Geophysical Research: Oceans*, 101(C12):28441–28455, 1996. doi: 10.1029/96jc02737.

A. Worby, G. Bush, and I. Allison. Seasonal Development of the Sea-ice Thickness Distribution in East Antarctica: Measurements from Upward-looking Sonar. *Annals of Glaciology*, 33:177–180, 2001. doi: 10.3189/172756401781818167.

A. P. Worby, C. A. Geiger, M. J. Paget, M. L. Van Woert, S. F. Ackley, and T. L. DeLiberty. Thickness Distribution of Antarctic Sea Ice. *Journal of Geophysical Research: Oceans*, 113(C5), 2008. doi: 10.1029/2007jc004254.

A. P. Worby, A. Steer, J. L. Lieser, P. Heil, D. Yi, T. Markus, I. Allison, R. A. Massom, N. Galin, and J. Zwally. Regional-scale Sea-ice and Snow Thickness Distributions from in Situ and Satellite Measurements Over East Antarctica during SIPEX 2007. *Deep Sea Research Part Ii: Topical Studies in Oceanography*, 58(9):1125–1136, 2011. doi: 10.1016/j.dsr2.2010.12.001.

H. Xie, S. Ackley, D. Yi, H. Zwally, P. Wagner, B. Weissling, M. Lewis, and K. Ye. Sea-ice Thickness Distribution of the Bellingshausen Sea from Surface Measurements and ICESat Altimetry. *Deep Sea Research Part Ii: Topical Studies in Oceanography*, 58(9):1039–1051, 2011. doi: 10.1016/j.dsr2.2010.10.038.

H. Xie, A. E. Tekeli, S. F. Ackley, D. Yi, and H. J. Zwally. Sea Ice Thickness Estimations from ICESat Altimetry Over the Bellingshausen and Amundsen Seas, 2003–2009. *Journal of Geophysical Research: Oceans*, 118(5):2438–2453, 2013. doi: 10.1002/jgrc.20179.

J. Yackel, T. Geldsetzer, M. Mahmud, V. Nandan, S. E. L. Howell, R. K. Scharien, and H. M. Lam. Snow Thickness Estimation on First-Year Sea Ice from Late Winter Spaceborne Scatterometer Backscatter Variance. *Remote Sensing*, 11(4), 2019. ISSN 2072-4292. doi: 10.3390/rs11040417.

D. Yi, H. J. Zwally, and J. W. Robbins. ICESat Observations of Seasonal and Interannual Variations of Sea-ice Freeboard and Estimated Thickness in the Weddell Sea, Antarctica (2003–2009). *Annals of Glaciology*, 52(57):43–51, 2011. doi: 10.3189/172756411795931480.

D. Yi, J. P. Harbeck, S. S. Manizade, N. T. Kurtz, M. Studinger, and M. Hofton. Arctic Sea Ice Freeboard Retrieval with Waveform Characteristics for NASA's Airborne Topographic Mapper (ATM) and Land, Vegetation, and Ice Sensor (LVIS). *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1403–1410, 2015. doi: 10.1109/tgrs.2014.2339737.

Q. Yu, C. Moloney, and F. M. Williams. SAR Sea-ice Texture Classification Using Discrete Wavelet Transform Based Methods. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 5, pages 3041–3043 vol.5, June 2002. doi: 10.1109/IGARSS.2002.1026863.

N. Zakhvatkina, V. Smirnov, and I. Bychkova. Satellite SAR Data-based Sea Ice Classification: An Overview. *Geosciences*, 9(4):152, Mar. 2019. ISSN 2076-3263. doi: 10.3390/geosciences9040152.

N. Y. Zakhvatkina, V. Y. Alexandrov, O. M. Johannessen, S. Sandven, and I. Y. Frolov. Classification of Sea Ice Types in Envisat Synthetic Aperture Radar Images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2587–2600, May 2013. ISSN 1558-0644. doi: 10.1109/TGRS.2012.2212445.

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1-53.

X. Zhang and J. E. Walsh. Toward a Seasonally Ice-covered Arctic Ocean: Scenarios from the IPCC AR4 Model Simulations. *Journal of Climate*, 19(9):1730–1747, 2006. doi: 10.1175/jcli3767.1.

X. Zhang, W. Dierking, J. Zhang, J. Meng, and H. Lang. Retrieval of the Thickness of Undeformed Sea Ice from Simulated C-band Compact Polarimetric SAR Images. *The Cryosphere*, 10:1529–1545, 2016. doi: 10.5194/tc-10-1529-2016.

L. Zhou, J. Stroeve, S. Xu, A. Petty, R. Tilling, M. Winstrup, P. Rostosky, I. R. Lawrence, G. E. Liston, A. Ridout, M. Tsamados, and V. Nandan. Inter-comparison of Snow Depth Over Sea Ice from Multiple Methods. *The Cryosphere Discussions*, 2020:1–35, 2020. doi: 10.5194/tc-2020-65.

H. J. Zwally, D. Yi, R. Kwok, and Y. Zhao. ICESat Measurements of Sea Ice Freeboard and Estimates of Sea Ice Thickness in the Weddell Sea. *Journal of Geophysical Research: Oceans*, 113(C2), 2008. doi: 10.1029/2007jc004284.