We thank all three reviewers for thorough and helpful reviews. Although reviewers suggested major revisions, almost all suggestions focussed on the presentation. The reviewers' main commentsrequested a dedicated methods section, more description of the ConvNet method in the main text (as opposed to in the appendix), and more exposition of the ConvNet results. We have modified the manuscript accordingly and reorganized the manuscript as follows:

(1) We have expanded the description of data processing methods in Section 2, and added dedicated methods section (Section 3) that describes briefly the linear fit approach, and an extensive description of the implementation of the ConvNet. Much of the latter is information that has been moved from the appendix, which is now short. We have kept the data processing separate from the data analysis (ConvNet) methods, as the former describes processing of the data to provide accurate data, while the latter describes how the processed data are analyzed.
(2) We have moved much of the description of the ConvNet implementation from the appendix to the main text
(3) We have added an additional figure of ConvNet results. As suggested by review #1, this shows a 2D comparison of both the linear fit and ConvNet ice thickness predictions to the original 2D mapped data.

In addition to suggested changes, we have removed the section on relationships between level ice thickness and ridge keel depths, as this was a diversion from the main results comparing simple linear and 1-D theoretical models with the ConvNet for estimation of ice thickness from surface topography, and not really connected with the main results. We have shortened the introduction accordingly (as also suggested by reviewer #1) to keep it focussed on the main thrust of the paper.

All reviewer comments are addressed inline below.

REVIEWER #1

General comments
This manuscript introduces estimating deformed sea ice thickness with in-situ data using simple statistical methods and deep learning technique. Although I believe Convolutional Neural Network (CNN) can be an alternative way to retrieve sea ice thickness without snow depth and densities, the readability of this manuscript is low. I think major revisions are needed before publishing.

Specific comments
Introduction: the long introduction distracts the objective of this manuscript. The authors should concise previous literatures in the introduction. The authors should focus on more the objective of this study in the introduction.
The introduction has been shortened and as described above, we have removed much of the material relating to ridge morphology as that is not germane to the main results.

P2 L27: surface elevation normally means surface height with respect to Earth ellipsoid in the altimetric study. Surface elevation and freeboard are used the same meaning in this manuscript, which can confuse the reader. I would suggest change surface elevation to freeboard throughout the manuscript.
We have changed surface elevation or snow elevation to “snow freeboard”, and freeboard (when referring to ice freeboard) to “ice freeboard” throughout.

P2 L29: since the hydrostatic equilibrium equation depends on altimeter type (i.e., laser/radar) it would be good to mention which one.
As presented here, this equation is used as the universal 1-D buoyancy equation, and is valid independent of measurement. While different forms are used for different altimeters, this is merely a rearrangement of the terms based on what is measured (snow freeboard, ice freeboard, or potentially some horizon in between)

P7: This manuscript majorly covers the methodology for estimating sea ice thickness. I believe this manuscript should include method section in the main body for better understanding to readers. This manuscript needs a method encompasses the entire manuscript. Particularly, as CNN is rather highlighting in this manuscript CNN details should be in the body manuscript.
We have included a Methods section that includes both background for the linear fit analysis, and a more extensive description of the ConvNet method and implementation, which included moving much of the material from the

appendix. Some of the technical ConvNet details are kept in the Appendix to not distract from the main points. Note, that the separate "Data and Processing" section describes the methods for acquisition and processing of the data. We have expanded this description, but kept it separate because it applies to all the subsequent analysis.

Table 1: Please briefly explain in terms of sail angle.
Defined sail angle, and clarified with "and a range of slopes across the deformed surface are given"

P9 L7: There is no validation for the model in 3.1.1 and 3.1.2, which is not consistent throughout the manuscript. Do authors have a specific meaning without the validation?
These sections were intended to compare our observations with prior studies, and because the fits are poor they are not validated. However, since section 3.1.1 (comparison between level ice thickness and keel depth) was not very relevant to other results of the paper we have removed this section along with much of the background information on ridge morphology to keep the paper focussed.

Section 3.1.2 was intended to demonstrate potential relationships between surface roughness and ice thickness (as suggested by previous authors) to motivate the use of surface morphology to aid in ice thickness estimation. This section has been shortened and moved to Section 3.2.3 where the incorporation of surface roughness is included in the linear fit analysis.

P12 L15: Please briefly explain in terms of drill lines.
This section has been removed as it is distracting from the key results in the paper, since drilling data are not used in the analysis. A short statement on the relative accuracy of drilling data has been added to clarify the corrections to the AUV data in the Data and Processing section (P7 L3-10)

P13 L17: Why this particular range? (2.9-6.1)
This comes from Table 2. We have added a note to refer to Table 2. This section has been modified and moved to the discussion as it seeks to explain the differences between our fits and those of prior authors.

P13 L23: What is the basis for setting 5.9?
Following Fig. 10 (in new manuscript, 7 in old), this is the best fit line. As above, this section has been modified for clarity and moved to discussion. We have added a note to Fig 10.

P14 Figure 7: While freeboard (F) is mentioned in figure 7, surface elevation (F) is mentioned in the caption, which is not consistent.
This has been changed to snow freeboard.

P14 L8-17: this paragraph should be in the discussion.
This has been moved to the Discussion.

P16 Table 2: Please briefly explain in terms of Akaike Information Criterion (AIC).
We have added a note that the AIC attempts to minimize information loss, and that we use the lowest AIC to perform model selection.

P17 Table 3: Why the authors separate linear model (i.e., without constant vs. with constant).
As noted in Stefan Kern's review, the without-constant fit is an attempt to match physical conditions of hydrostatic equilibrium (and permit estimation of effective densities), whereas the with-constant fits are empirical and attempt to minimize fit error. The reorganization of the text separating out methods, results, and discussion should make this more clear. We use the fits with constant as a basis for comparison of how our ConvNet improves upon linear fits, and we use the fits without constant to estimate ice/snow densities in our data.

P17 L10-L18 - P18 L1-12: this part should be in the discussion.
We have kept part of this in the Results as it is reporting errors for some particular fit. As such, it is a result and fits best there, but we have moved the discussion of this to Discussion.

P18 L13: It would be better the authors include the spatial distribution of sea ice thickness derived by CNN with discussion.

This has been added to the manuscript as a new figure (Fig. 9) which shows that the the ConvNet prediction matches the spatial variability better than the linear fit (see below in response to Reviewer #2) for the figure) While we have produced this plot for each floe, we are electing to include one example as all floes show qualitatively similar results.

P18: the first paragraph of 3.3 should be in the methods.
The structure has been reworked as described above.

P19 L12: Normally this parameter setting is determined by trial and error.
There are several reasons for this choice. Our goal here is not to make the best possible network, but to make a good network that can be interpreted to physically justify why the network is working. Our method is predicated on the assumption that feature morphology is important to SIT prediction. Since our feature sizes are of similar scale, we felt this was a reasonable choice so that the ConvNet would learn features that are likely physically relevant to sea ice thickness variability. With our limited dataset, we cannot use too large a window as this would lead to too few unique samples; similarly, if we use too small of a window,relevant physical features would not be captured. We have also tested the network by halving the window size (which has the danger of not capturing relevant physical feature scales), and by decreasing resolution, with no significant effect on performance as described in the text. Because of the limited size of the dataset, we cannot test this dependence further.

P20 Figure 9: Some part of the caption of figure 9 should be in the main body. (from we also to the end).
Caption has been edited. We keep a reference to the linear fit as a description of what is in the figure.

P18-21: 3.3 predicting SIT with deep learning is quite mixed with methods, results, and discussion. please reorganize 3.3.
We have reorganized as described above.

P25 L15: what is meant by Figure #0?
The text reads "Feature #0", which is the first feature in the 8-bit vector, as shown in the referenced Figure.

P26 L31: as the validation of this method is spatially limited, this sentence should be corrected.
This has been corrected to estimating SIT with "comparable morphology"

Technical corrections
P2 L9: wieth -> with
P7 L6: With -> with
P7 L7: need references.
P9 L11: I don't see ratio of keel depth and snow-sail height in the Table 1. **This is Hs/Hk**
P17 Table: replace "no constant" with "without constant".
P19 L16: replace "CNN" with ConvNet.
P22 Figure 10: Figure 10 never mentioned before.
Thanks, these have been fixed.

Reviewer #2

Estimating Early-Winter Antarctic sea ice thickness from deformed ice morphology by Mei, J. M., et al.

General Comments: GC1: I note that a dedicated "Methods" section is missing completely. It is Data followed by Results. There are places where this seems ok for the flow of the manuscript but there are other places, e.g. Section 3.2 where this seems not to be optimal. The deeper I stepped into this section the more confused I got. At a certain point I got lost with density values and with regressions with or without intercepts or additional constants. This section would perhaps benefit from a clear up-front explanation of what you did / how you derived coefficients / which density values you choose (and why) / how you derive effective density values (and why)? Such an improvement in structure of the paper would possibly also reduce its length a bit here and there.
The paper has been restructured with the addition of a Methods section. The text and motivation for this analysis has in particular been improved by separation into Methods, Results, and Discussion sections, so that the basis for comparison with ConvNet and previous results is more clear. We have also improved the discussion of regressions and derivation of density values (discussed more in response to the next comment)

GC2: You treat the hydrostatic equation as a form of a linear fit. While one can see this as such a fit it would be very important to mention (even more) that the coefficients as you call them are based on density values and are computed based on physics. This is an important difference to the empirical linear fits used by Xie et al. or Ozsoy-Cicek et al. which are purely mathematical. To my opinion it would add to the understanding of your paper if you would clarify this even better at an appropriate position in your paper. I'd think that interpreting the CNN results into the direction that effective densities can be derived is very hypothetical - especially given the unknown (and non-existing) relationship between sea-ice and snow densities which are both involved. I note in this context that the issue of negative ice freeboards has neither been mentioned nor discussed. I guess it would not hurt to get back to it given the results published in Ozsoy-Cicek et al. (2013) and Yi et al. (2011).

We interpret the hydrostatic equation as a linear fit solely to provide density estimates, which is only relevant for a dataset that has snow depth measurements. We now stress this more in the text. The ConvNet results do not derive any effective densities; we only suggest that the network may be accounting for different effective densities.

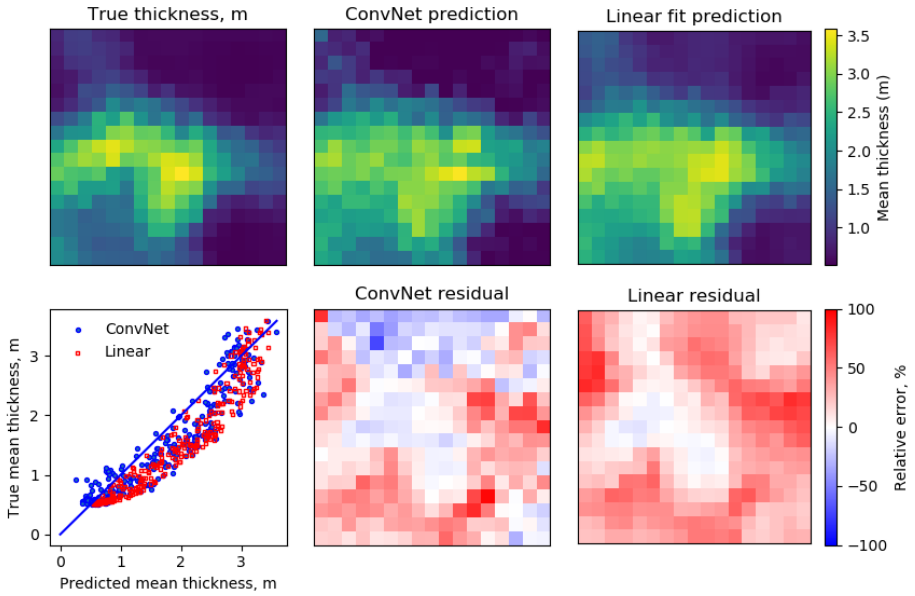
Similarly, the one-variable fits to surface elevation (F) can be interpreted as (given some snow/ice density) an average snow-to-ice ratio in the measured surface elevation. This again is not particularly prescriptive, as other datasets no doubt have different snow-ice ratios. However, when averaged over large enough areas, it is likely that the ice component in the measured surface elevation is low (and perhaps the snow = freeboard assumption is now reasonable in some, or many cases). This can be inferred from Xie/Ozsoy-Cicek's fits as their coefficients of 2-3 are equivalent to assuming F=D in our Eq. 1. We stress this more in the text that these no-constant fits are intended to check why our coefficients may be different to Xie/Ozsoy-Cicek's.

Also, we deliberately do not discuss negative freeboards in this case because in this dataset there are few negative freeboards sowhen averaged over 20m windows, our data have no negative freeboards (although at the 0.2m resolution there are some). and so they cannot be reasonably included in the article. Where they do exist, they are primarily on the flanks of ridges where their near-local effect will be negligible. We have added a sentence in the Data section to note that there were few negative freeboards.

Note, when applying either a linear fit or the ConvNet to surface topography data, we cannot know whether there are negative 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. We have added a note about potential effects of negative freeboards in the discussion (P 20 L3-7).

GC3: I am missing the presentation / discussion of more results of the ConvNet approach. What a reader might have loved to see is profiles of sea-ice thickness computed from the draft-snow depth-surface elevation measurements (your benchmark)and of the sea-ice thickness estimated with your approach. Ideally you are able to show at least one representative profile of each PIP used here. That way one will get a better handle on the actually estimated sea-ice thickness distribution compared to the measured one - in addition to the histograms shown.

This has been added to the manuscript. It is also shown here, using PIP8 as the test set. This plot requires considerable oversampling (here, it is oversampled at 4x, i.e. using a shift of 0.25 * window size of 20m), otherwise there are not enough points to make a useful visualization. The mean relative error of the ConvNet (trained on PIP 4, 7, 9) applied to this test set (PIP8) is 23%, vs. 31% for the linear fit (fitted to PIP 4, 7, 9) applied to this test set (PIP8). It shows that the ConvNet is better generalized to new datasets, and also shows the considerable biases that a linear fit can have when applied to a new dataset, presumably due to a varying snow/ice ratio. We have generated these for each floe, but choose to include one example in the manuscript as all are qualitatively similar.



GC4: ICESat-2 is up since September last year. After having read the paper I am wondering what the ultimate goal of your work is. Is it to create high-resolution validation data sets of the sea-ice thickness which are spatially distributed? Or is it to develop an algorithm which potentially could be applied to ICESat-2 data. For both cases, I believe the authors could stress the main motivation and future use of their work and product.
This is a good point, and we have added to the text in the introduction to better state the goal of the paper. The long-term goal here is to improve on ice thickness algorithms for ICESat-2. However, our results are not directly transferable as ICESat-2 only maps a straight line (well, the 3 beams are not quite enough to form a 2D surface). As such, the ConvNet approach is not appropriate. While there are also non-convolutional deep neural networks that could work on such data, although one would need to test whether ICESat-2 can sufficiently capture morphological features that are related to ice thickness.

An alternative approach is then to use other more extensive 3D-datsets and use the technique to identify what morphological metrics are best predictors for a variety of types (which is much of the reason for our investigation of possible physical basis for the ConvNet features). For example, with IceBridge data one could use this method to predict radar snow depth. Alternatively, additional datasets (for example more coincident AUV and surface topography, or coincident scanning LIDAR and EM-bird observations) could be used to relate these identified metrics directly to ice thickness. Then it may be possible to relate ConvNet metrics that are good predictors of thickness to analytical metrics, then the results could be used to optimize algorithms for ICESat-2

Our goal in the paper was to demonstrate (1) to show that sea ice surface morphology contains information that can be related to ice thickness so that linear fits can be improved upon, (2) deep learning has the potential to use this information to provide "optimal" predictions of ice thickness, and (3) these deep learning techniques respond to features that are likely physically meaningful and hence there is scope to use this physical information to provide better ice thickness predictions. We acknowledge that our particular deep learning architecture is not necessarily what would eventually be used in practice.

This was only briefly touched on in the Dicussion and Conclusion before, so we have expanded this discussion to better suggest a viable strategy.

Page 2, Line 14: I suggest to cite the paper by Behrendt et al., 2013, Sea ice draft in the Weddell Sea, Earth System Science Data, 5, measured by upward looking sonars to underline that also this ULS data is a valuable source - even though you write "sporadic"
Added, thanks.

Page 3, Line 1-9: I guess it would not hurt to perhaps again refer back to Kern and

Spreen (2015) who dedicated some work on the uncertainty analysis for ICESat sea-ice thickness retrieval and to Kern et al., 2016, Antarctic sea-ice thickness retrieval from ICESat: Inter-comparison of different approaches, Remote Sensing, 8(7), who inter-compared a number of sea-ice thickness retrieval approaches for the Antarctic and on which results the Li et al. (2018) paper cited is based upon. I am tempted to say that the fits used by Li et al. (2018) are based on the work of Ozsoy-Cicek et al. solely and not on the work of Xie et al. (2011).
Thank you for the correction on Li (2018). This and the additional references have been added to the introduction

Line 14: "do not yet understand the distribution" –> I am inclined to say that we do understand the physical mechanisms forcing the snow depth distribution around ridges very well. What we cannot yet do is, however, to measure this distribution accurately over an large enough area.
Changed to "we do not yet know the statistical distribution"

Line 10-19: In this paragraph, the work of Weissling and Ackley, 2011, Antarctic sea-ice altimetry: scale and resolution effects on derived ice thickness distribution, Ann. Glaciol. 52(57) might fit as well.
Thanks, added.

Lines 16/18: This is perhaps a good place to refer to the work of Hutchings et al., 2015, Comparing methods of measuring sea ice density in the East Antarctic, Ann. Glaciol., 56(69)
Thanks, added.

Lines 20-22: I agree that the unknown snow depth is one factor here. But isn't the fact that we don't know the keel morphology and distribution relative to what we see from above with a LIDAR contributing much more to a potential bias in estimated sea-ice thickness?
Yes, this is a good point, and we have clarified this in the text.

Line 23-27: As far as I know, Kern and Spreen (2015) focused quite a bit on ICESat and the uncertainties involved. I doubt, however, that this is the correct citation for the AMSR-E snow depth bias issue. I'd say the first to report this issue were Worby et al., 2008, Evaluation of AMSR-E snow depth product over East Antarctic sea ice using in situ measurements and aerial photography, J. Geophys. Res., 113. Their work was followed later by Ozsoy-Cicek et al., 2011, Intercomparison of Antarctic sea ice types from visual ship. RADARSAT-1 SAR, Envisat ASAR, QuikSCAT, and AMSR-E satellite observations in the Bellingshausen Sea, Ann. Glaciol., 52(57) or Kern et al., 2011, An intercomparison between AMSR-E snow depth and satellite C- and Ku-band radar backscatter data for Antarctic sea ice, Ann. Glaciol. 52(57). In the same Ann. Glaciol. volume you also find the paper by Markus et al., 2011, Freeboard, snow depth and sea-ice roughness in East Antarctica from in situ and multiple satellite data, Ann. Glaciol., 52(57). Another paper about the deficiencies of the AMSR-E snow depth product could be this one: Kern and Ozsoy-Cicek, 2016, Satellite Remote Sensing of snow depth on Antarctic sea ice: An inter-comparison of two empirical approaches, Remote Sensing, 8(6).
Thank you for the detailed notes; text and citation has been amended.

Line 33: "fewer such datasets exist" –> This applies to the Antarctic and I would mention this accordingly. In the Arctic there are way more draft measurements available and these have actually been used to develop draft-based sea-ice thickness estimation tools.
Fixed.

Figure 1: I love this figure. It could be even a tiny bit more realistic if the ice floe or sheet would not be continuous in the ridge / keel area.
This is fixed.

Page 4: Lines 3-12: I am wondering whether in the context of this discussion the work of Goebell, 2011, Comparison of coincident snow-freeboard and sea ice thickness profiles derived from helicopter-borne laser altimetry and electromagnetic induction sounding, J. Geophys. Res., 116 should be mentioned as well?
Thank you for this reference, though it is added a bit later for discussing the coefficients for T vs F in later sections.

Page 6: I suggest to mention / give answers to the following questions here: - Water depth in which the AUV was operating - How many AUV scans per "cake" were stitched together? - If multiples scans: Were all scans carried out into the same direction? Or parallel to each other in opposite directions? X-ing? - Did I understand correctly that per "cake" 4 surface elevation scans were carried out, each from one side of a 100 m x 100 m grid? Or did you actually fly over the area? - I assume the snow depth measurements were the last measurements carried out –> although it is logical it is worth to mention this. - It is not entirely clear how the about 2000 measurements per "cake" are distributed across the "cake" area and how this was technically realized. I assume that the measurements were carried out along parallel transects across the cake with a fixed transect-to-transect distance and that only the sampling along each single transect varies between 5 m and 0.1 m. - How is the reference sea-surface height computed and how accurate is it?
Added to text.
The AUV survey was done following Williams et al (2014), at a depth of 15-20m 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 snow depth was indeed done last. Your description of the snow sampling is mostly correct, the sampling along a transect is not purely 1D, as we could not necessarily walk over a ridge, and instead would walk around it, sampling the snow distribution at high resolution.

Line 29: "The ice thickness can ..." –> so there were no drillings?
There was one drilling line done per floe with 50 points at 2m resolution along the edge of the lidar scan. This was used for calibration purposes using the level ice. This is described in the Data section. An example of the comparison is shown below, but is not included in the manuscript for brevity. The freeboard measurements (middle panel) have poor agreement due to sampling bias (often, when drilling near a ridge, the most accessible point to drill is the lowest elevation). The error bars here refer to the max/min MagnaProbe/Lidar/AUV measurement in a +-1m range. The drilled draft measurements do not necessarily match the AUV measurements in ridges because there may be biases in drilling through thick ice, slight differences in location when thickness variability is extreme, and may tend to miss deeper loose blocks. The AUV will tend to detect the full draft, and samples at the sonar footprint resolution (as opposed to at a 2 inch drill hole point)

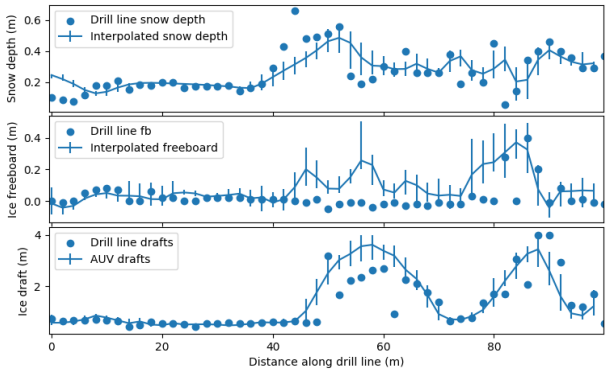


Figure 5: Readability of this figure would improve with an increase of its size.
Fixed

Page 11:
Line 18: "very similar value of 1.3" –> This very similar value needs two standard

deviations (2 time 0.1) to include that 1.5 values from other studies. Perhaps "similar" would do it?
Fixed

Figure 6: - See comment to Figure 5. - In the caption I would call the black line dashed rather than dotted.
Enlarged and fixed.

Page 12:
Line 11: "The lidar and AUV data were corrected by ..." –> I don't understand what needs to be corrected here. Is there a way you specify better what you did? Why is this a "correction"?
The correction is a simple offset to the entire lidar or AUV survey that is applied to account for the uncertainty in the AUV trim (which may cause slight depth offsets for different surveys) and the depth sensor, and lidar referencing to the sea surface (linked to your earlier comment). This has been clarified in the Data section.

Page 13: Line 12: "snow = surface elevation assumption" –> I recommend to mention that this is a strong assumption, that it applies to thin, perhaps medium thick first-year ice only (possibly only to the kind grown under quiescent conditions, i.e. not originating from the pancake-ice cycle), that it requires a certain snow load to be present, and that such an assumption can only be made if one is not interested in a really exact sea-ice thickness estimate.

This has been clarified so that it is clear this is a lower bound, and that in our case it does occur for the thinner, level ice.

However, we note that this is quite often close to being true, at least for regional means such that the large-scale error is small. Based on our own observations in the field over many cruises, it is broadly true for much first-year ice, and even can be reasonable for very thick ice where positive freeboard and negative freeboards cancel. While this is often reasonable on a regional scale, we agree it is not a good assumption at smaller scales, and presumably what creates much of the scatter in linear fits. Incidentally, this also suggests why using surface morphology may help improve predictions – even where a linear fit is accurate in the mean, it cannot capture this variability, while surface morphology may be suggestive of variations in the snow/ice freeboard ratio.

Line 17: "All our coefficients" –> Would you mind to refer to the place where you already mentioned these coefficients?
Added.

Line 21: "2.2-3.1 in Ozsoy-Cicek et al (2013)" –> I tried to figure out how you ended up with this range and potentially misunderstood something. If I check that paper, then - in Figure 5, which is possibly the one you got these numbers from, - I find regression lines with a considerable intercept between ∼10 cm and ∼30 cm, depending on the region, paired with this range in factor of F of 2.2 - 3.1. But these values are valid for positive ice freeboards only. When taking all ice freeboards into account, then the black lines (and numbers) in that Figure 5 apply. In addition to that: I could make sense to focus only on the Ross Sea results from that paper?
We used the positive freeboard regressions because all our 20m-averaged freeboards have no negative freeboards (although there are individual negative freeboard values at 0.2m resolution). If we take the 'all freeboards' coefficients, the range is 2.4-3.5, which does not affect our analysis. However, if we just take the Ross Sea coefficients then the range is 2.4-3.1, which again does not change the analysis. So we will just use 2.4-3.5 as our analysis would apply to both cases.

Lines 30-32: "This means that assuming ..." –> So what you state here basically is, that the linear regression approaches developed by Xie et al. and Ozsoy-Cicek et al. are of limited value? If so you could mention this and also refer to Kern et al. (2016) in Remote Sensing, where it is layed out that the linear regression approaches fail to provide a meaningful circum-Antarctic sea-ice thickness distribution.
Added. However, the approach of Xie et al and Ozsoy-Cicek et al may be reasonable at larger scales, and this is described in the text. We now state that such relationships should be used with caution.

Figure 7 - Some data points are annotated "snowy" –> I did not find an explanation of what this is in the text or in the caption. Where is the distinction between "snowy" and "ridged"? - caption: The ice density value given in line 4 of the caption differs from the one given in the text on page 13, line 25.
Density value changed. 'Snowy' surfaces are manually classified as those that have snow features (likely originating at the ridge, but the ridge is not in the window). The classification is purely meant for the feature analysis in the Discussion. This is described in the text, but now we also added "manually classified" to the caption.

Page 14: Line 3: "T = 2.45F + 0.21" –> is modified from Ozsoy-Cicek et al., now using unit meters instead of centimetres, correct? "for a winter Ross Sea" –> according to Ozsoy-Cicek et al. (2013) this is data from just one cruise in Sep./Oct. = much later in the season than PIPERS. In that sense your statement in Line 6 "same region/season"
should perhaps be changed? Also the spatial overlap (see Ozsoy-Cicek et al., 2013, Figure 1) is quite small.
Yes, we converted the equation to meters. Merging with your below comment for lines 15-17, we have removed "same season/region", and now point out that the proportion of deformed ice is varying and perhaps causes linear fits to not generalize well.

Line 4: So your intercept is -0.73 meters or -73 cm? That is quite large.
Yes; we discuss the reasons for this in the text (P19 L10-15).

Line 7: "nonzero freeboard" –> "nonzero ice freeboard"
OK, changed

Lines 15-17: Yes, I agree with your interpretation. However, it might make sense to also mention that the Ross Sea data used in Ozsoy-Cicek et al. (2013) was from a different part of the Ross Sea and from a different season and to my opinion indeed exhibits a totally different characteristics than the PIPERS data set collected 3-4 months earlier.
Yes, see 3 comments before this for summarized changes.

Page 15: Line 6: "additive constant" –> I don't understand what you mean by this. Did you add an intercept?
Yes, changed this to "we fit a linear regression both with and without a constant term" - we don't want to use 'intercept' as this only has meaning for a one-variable fit.

Lines 7 & 8: Isn't it surprizing that the coefficient for F fitted over all four PIPs of 10.4 is so close at the upper range of 10.6 for individual PIP fitting? Also: The range for the coefficient for D of the individual PIPs does not include the value found over all four PIPs. Is this logical?
The answer to both of these questions is that the multilinear fit fits both variables simultaneously, and so the fit for all floes combined is not a weighted average of each individual fit (your intuition would be correct for a one-variable fit – and indeed it is, the F-only, no intercept fits have coefficients of 6.5, 6.4, 4.8, 4.1 and the overall fit coefficient is 5.8)

Lines 9-12: - Your measured snow densities are considerably lower than those given by Sturm et al. (1998). Could it be that the latter were obtained in late winter / spring? - While I understand the concept behind the effective sea-ice density (voids filled with water included in the density estimate) I have problems to understand the concept of an effective snow density. What is this? In this context, I find your effective snow density value to be quite high. - I guess it would be good to learn how you ended up with the density values reported in Line 10. When I tried to insert your range for the factor for F (7.9 to 10.6) into an equation where D is zero, then I end up with densities between 897.9 and 931.0 kg/mˆ3. But of course, without further information from your side I cannot reproduce your numbers. - I find it quite surprizing that the standard error for the effective sea-ice density is so low compared to that of the effective snow density.
Sturm includes 2 Ross Sea cruises, one in May-July 1995 and one in Aug-Sept 1995, with snow densities of 350 and 390 kg/m3 respectively, although the May-July cruise would have somewhat older snow than in our case (due to

somewhat earlier dates and a later freeze-up for PIPERS)The snow does not have an "effective density" and this has been corrected in the text.
The standard errors are computed using the standard error of the linear regression and propagating them. This should not really be interpreted as an uncertainty value for the sea ice density, as it just means the multilinear fit has a (relatively) low error for the F coefficient (sea ice density variability can still contribute to displacement of any given point from the fit line) This has been clarified in the text.

You use a water density value which differs from those given at the beginning of this paragraph. Why? Where does this value originate from?
The water density of 1028 comes from CTD casts from PIPERS, whereas 1027 comes from Worby (2011). The difference is minor, but have added that this came from onboard measurements.

Lines 15-18: Please check these sentences. There is some repetition first and then something is missing.
Fixed

Lines 20-22: "For example, ..." –> Just to understand this: What you write here in the text is the comparison between using coefficients of ONE of the PIPs to estimate sea-ice thickness in another PIP while in Table 3 you show the comparison between using a joint coeffient of THREE PIPs to estimate sea-ice thickness in the remaining PIP. I just got confused a bit about why you write different things in the text than you actually show in Table 3 (and refer to in the subsequent sentence).
You are right. We had listed the results of using each ONE of the PIPs to show that the average error was not dominated by one particularly bad one, but you are right, it is better just to show the average of the THREE PIP fit applied to the fourth PIP.

Table 2: - "no int." means what? - Are you sure the AIC is a monotonic function even on the negative value range? I am just wondering whether the "smallest" AIC criterion does not need to be applied to absolute values? Could it be that these negative values have no proper meaning in case that the correlations are so low? - The subscript "adj" stands for "adjunct"? "adjusted"? If the latter adjusted to what? - What is the unit of the constant? It seems to be in meters?
No int. means the forced fit through the origin with no intercept. This has been added to the caption and we now refer to "F, D" and "F, D, Constant" fits for clarity.

The AIC is not dependent on the absolute measurement, and rather the difference between AIC values can be interpreted as a relative likelihood, e.g. if two models have AIC values A and B, with A<B, then the second model is exp(A-B) times as likely to minimize the information loss. This has been added to the text.

Adj stands for adjusted, and it is adjusted to account for varying sample sizes. This is mentioned in the caption.

The constant is in meters and this has been added to the column.

Page 17: Table 3, caption, last line: "zero freeboard = zero thickness condition"? Do you refer to ice freeboard here? Do you perhaps mean "zero snow depth"?
This has been deleted from the caption, as we now include the fit with constant term.

Line 1: This equation is something you could use in a "Methods" section (should you include one) to tackle my general comment GC1. I note however, that seemingly with this equation one can explain only parts of the entries in Table 2; the c3 times sigma part is not represented in Table 2.

The fit with sigma is only mentioned to show that it does not improve the fit (likely because sigma is itself highly correlated with mean surface elevation). It is not an important result by itself, and also would not fit into Table 2. The equation will be moved to the new Methods section.

Lines 5-9: It is still not clear to me how you discriminate between "snowy" and "level". Please add.

This is manually done based on whether the majority of the image was level or contained a visible snow feature in the lidar window. 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.

Page 18: Lines 3-6: I can in principle follow your argumentation that zero surface elevation (= zero snow depth) means zero sea-ice thickness. I would sign this if we consider larger scales. But on the scales investigated with the PIPs this is not necessarily true because under cold conditions and hence impermeable sea ice there will be many places with a negative ice freeboard. Even if we assume for simplicity that most of these will have a snow cover and hence potentially have a non-zero surface elevation, it is still likely that especially in the vicinity of ridges and/or where the ice is under lateral stress - at the scales of your measurements - you will have surface elevations close to zero or even negative ones paired with a non-zero sea-ice thickness. - This paragraph is again a good place to comment and/or underline the difference between the physically based coefficients used by Zwally et al. and similar papers and the empirically based coefficients used by Xie et al and similar papers. One could argue that the physically-based coefficients are more dependent on the validity of the hydrostatic assumption while the empirically based ones are not ... but I am not sure this holds.
On the scale of the actual linear regressions (20m), there are no negative (mean) surface elevations. However, we have decided to scrap the requirement of zero S.E. = zero thickness, as our ConvNet performs better than a linear fit with constant anyway. As stated above, the use of no constant is now only used to compare to the theoretical fit and estimate densities.

Page 19: Line 13/14: "20% of the data ..." –> this refers to the randomly selected data? If so, please stress so in the text.
This has been clarified.

Line 13 vs. Line 16 and remainder of the text: Please check your usage of "floe". From the text until here I got the impression that the PIPs are subsets of one floe. Here I get the impression that PIPs comprise several floes out of which a few are selected. Please clarify your terminology here.
Each PIP is sampled from a different floe, but is nevertheless a subset of that floe (i.e. only a portion of the floe is sampled. This has been clarified in the text.

Line 20: You state PIP8 here but in Figure 9 it seems you refer to PIP9. Please check.
Fig. 9 should read PIP8, thanks for noticing this.

Line 24: "epoch 881" –> does it make sense to refer to the Appendix here? Otherwise this information is perhaps a bit out of context.
Ok, this has been moved. Note that most of the appendix is now in the Methods section, in response to other reviewers and to not repeat information from the Methods.

Line 31: "are all negative" –> except for level ice.
Fixed

Line 33-35 and beyond: I doubt that this comparison should be presented as is. Aren't these data sets quite different? I wrote about the sub-set of data for the Ross Sea used in Ozsoy-Cicek et al. (2013) already. In Li et al. (2018), the data basis is ICESat footprint-scale estimates of the freeboard - hence we talk about one value for one footprint of which we do not know how well it covers how many different surfaces. The data set used in Ozsoy-Cicek et al. (2013) is at least based on multiple measurements conducted on one or more transects across a single floe.
This has been amended to compare against the profile mean RMS error of 11-15 cm from Ozsoy-Cicek et al (2013) Table 7 against our validation error. The training error, which is equivalent to a fit error, is not as good of a comparison because a ConvNet can/will overfit with an artificially low training error; model selection is done by choosing the best validation error, which is kept separate (but is similar to) the training set in order to try and reduce overfitting.

Figure 9: The description / caption of the figure needs to be improved. - Please annotate the images with a), b), c) ... - What is the value behind showing a continuous fit in addition to the bars? It extrapolates the bars towards non-existing data values. - What are the bin-sizes used? Are these the same in all three histograms shown? Do they always have the same borders (i.e. minimum and maximum value included into the count of a respectiv bar)? - The peak counts are obscured by the legend. This needs to be changed. - I suggest to add in each row which PIPs are used for what. - I suggest to stress in the caption that the last row shows a different range of thickness values. - The caption in line 3 says PIP7-9 but in line 2 it is PIP4, 7 and 8. What is correct? - The caption in line 4 says PIP9. True? - What is the unit at the y-axis in the histograms? - The mixed colors in the histograms originating from overlapping bars of different data sets are not easy to interpret. Perhaps you could either add these in the annotation (which is possible if you increase the size of Figure 9) or find a different way to show the counts of the different data sets. One way to do this would be that you use substantially narrower bars which you do not let overlap each other and center these at a specific thickness; then in the caption you might need to state that you display three bars centered at a specific thickness, separated horizontally for better visibility. That way the real differences in the distributions would become more clear.
This figure has been redone with just the outline of the histogram, binned at 0.4m, with additional labels added. The caption has been fixed, thanks for noting this.

The bin sizes were chosen to have equal numbers of bins as opposed to a constant bin size; this has been changed to constant bins of 0.4m.

Page 20: Lines 2-4: Perhaps you could put these numbers in context with the number of data points used to get these uncertainty estimates? I guess, in case of Ozsoy-Cicek et al. (2013) we are talking about 23 floes with an actually unknown number of measurements per transect. About how many measurements are we talking in your case?
We have 4 floes, but each floe has many measurements as the lidar/AUV data can be binned at varying resolutions. So the REM (relative error of the mean) is comparing floe mean with floe mean; although as you point out the floes may have different numbers of measurements. We have added a note that our test error is essentially taking 3 floes and applying the fit to a fourth floe, vs. fitting to 23 floes. It is probably easier to get a good fit with fewer floes, but we also expect poorer generalization with fewer floes, so we can reasonably infer that our fit is better generalized than a linear fit.

Page 21: Line 14: I suggest to remove the "see" –> at least I cannot see these results.
We have removed it.

Line 29: "isostatic assumption may no longer be valid" –> may be so. What do we know about spatial scales over which the isostatic assumption is valid? No too much I'd say - particularly for ridged ice. Perhaps this sentence could be deleted.
Deleted. In fact, we realized that the ConvNet does not require assuming isostacy.

Figure 10: - this figure belongs to section 4 and should be located within section 4 not before it. - Why do we have 3x3 imagettes for layer 1 but 16 for layer 3? - The size of 4 m and 8.8 m given in the caption, do these refer to the pixel size in these imagettes or to the imagette size itself? It seems as if the pixels in layer 1 are indeed smaller than in layer 3. - Instead of "as the lidar" you might want to write "as the surface elevation" - Is it in this context correct to assume that layer 3 has the unit meters while layer 1 is unitless? - What do the bright and dark pixels in layer 1 mean?
Moved to appropriate section. The pixels corresponding to the meters of the layers have been added to the caption. Due to the stride, each subsequent layer essentially halves in resolution. Darker colors indicate higher weights, though the actual weight values are not important. The reference to "surface elevation" has been fixed. All the layers are unitless as the weights are just a numerical weight value. The direction of the colorbar doesn't actually matter as the difference would just be a negative sign, easily accounted for in any of the subsequent hidden layers. This has been clarified in the text.

Page 22: Line 17 through Page 23: Line 7 and Figure 11: - Please provide a), b), ... in Figure 11; it aids referring to the images. - I suggest to mention Figure 11 before Figure 12. - You refer to Figure 11 in Line 5 but should perhaps also do it in Line 3 (strong correlation for feature #0) and again in Line 6. - I have difficulties to understand the continued mentioning of "effective densities". I doubt that with the CNN you can (and should) derive any conclusions about the effective density - especially because the densities for sea ice and snow do not necessarily co-vary. This brings be back to GC2.

Figure 11 and 12 have been switched. The ConvNet cannot give any conclusions about why it has learned its prediction; we simply try to give physically plausible explanations without asserting that these are true. Our goal is to show that this method can work in general for other datasets, and why we may expect this to be the case. We will attempt to better stress how speculative our discussion is.

Page 23: Line 10: "snowy surfaces" –> which still need to be defined in comparison to "level surfaces".

Yes, this has been clarified in the text.

Page 24: Line 4: "ridged and level surfaces are clearly distinguishable" –> I don't agree when I look at Figure 13 - unless I have perhaps misunderstood what the used tool is able to show. But my interpretation of this figure is that level, ridged and snowy symbols overlap well.

Ridge, snowy and level overlap somewhat, but Ridge and Level are more distinct with less of an overlap. This suggests their features are differently analyzed by the ConvNet.

Page 25: Lines 9/10: That prediction of snow depth from lidar input is possible as also been shown by Ozsoy-Cicek et al. (2013) and Kern and Ozsoy-Cicek (2016).

OK, added

Lines 11-13: I guess these two sentences could be deleted.

Removed

Page 26:

Lines 1-8: I am not a fan of these attempts to try to relate CNN features to (effective) snow density variations which may or may not be realistic and physically meaningfully linked to input parameters. To my opinion, this really requires a careful analysis and description of how the CNN "learns" from the input data and whether there is (within the CNN) a link to physics - which I doubt is the case.

We agree that the speculation that these features may be linked to snow density variations is highly speculative. Because the weights of these features (5, 7) are so small, we decided it is not that important and we have removed this from the text.

Lines 18: "thickness of a new dataset" –> you seem to have applied your approach to a different PIPERS data set. It might be really beneficial to show this example in the paper and not to just mention it. Particularly because you come back to this in your conclusions (Line 33, "unseen floe").

This was badly worded. By new/unseen floe, we mean a dataset on which the net is not trained (i.e. the test floe). We have changed these to "test dataset".

Page 27: Line 5: "it can account for a varying ice/snow density" –> I'd say that this is a hypothesis. It may be that the ConvNet is able to account for the different densities and perhaps even provide additional information about these - but the evaluation of whether this is the case and/or whether this is at all meaningful physically based on deep learning is not known and might not be over-stressed here.

We have softened this as a "possible" strength.

Lines 12-15: "Our error ..." I suggest to not overstress these inter-comparisons because these are based on completely different data sets and scales. After all, a real quality measure of your method will be its application to ICESat-2 data which should be the overall goal here - as is finally mentioned in the last paragraph.

We will keep the reference to the survey-wide mean RMSE from Ozsoy-Cicek 2013 because as explained in the results and discussion, we feel this is a reasonable comparison, but we temper the statement by saying these are different datasets. We will delete the comparison to Kern 2015 and satellite based estimates.

References: You need to go through the references list and complete it with respect to page numbers and journal volume and issue numbers. Also doi's are generally missing. Some journal abbreviations are not in place.
OK - done
Page 28: Line 31: I am not sure but I guess Figures in the Appendix need to be named differently to the main text. See the instructions for authors.
Thanks for bringing this to our attention.

Page 29: Line 2: Would you mind commenting on the layer sizes being first 4 m, then 8.4 m and subsequently 8.8 m? Do these "strange" values have to do with the pixel size of 0.2 m?
This is discussed in an earlier caption, but it is simply because of the stride of 2 halving the resolution each layer (so 0.2m, then 0.4m, then 0.8m), with window sizes of 20, 21 and 11 pixels. This has been clarified.

Figure 15: - What is a "training loss"? - It appears that after epoch ∼550 there is a small jump in validation and training error from a certain level before that epoch to a certain, lower level afterwards. Any explanation to this? - What explains the sudden increases in the training error from a low background of ∼15-16% MRE to the level of the validation error? It seems as if the result of the ConvNet even after that many epochs is still not stable?
The training loss is just the loss function (mean squared error) of the training set. This has been clarified in the caption.
The jump may be due the method being stochastic and this accounts for the error jumping around. Also, because we are optimizing mean squared error, this is correlated to but not exactly equivalent to optimizing the MRE (which is closer to optimizing mean absolute error). As the method is also stochastic, we could possibly get a smoother curve with more epochs and a smaller time step, but this increases training time. Again, the point here is to show that this method is effective for lidar datasets in general and not to propose that our architecture is the best possible one.

Typos:
Please replace "e.g" by "e.g." (a few incidences) Please check usage of "climatology" and replace all incidences in the paper by a more appropriate term.
Page 2: Line 9: witeh –> with, interrannual –> interannual
Page 9: Line 26/27: "beyond beyond" –> "beyond"
Page 15: Line 11: "which" –> "who"
Page 21: Line 28: "slighly" –> "slightly"
Page 28: Line 23: "assigining" –> "assigning"
Figure 14, caption, line 3: "optimzer" –> "optimizer"
Thanks for this.

REVEIWER #3

Dear TC editor and authors of the manuscript TC-2019-140,
The topic of the manuscript is interesting and the content useful for sea ice research. A neural network has been applied to sea ice thickness (SIT) estimation from lidar surface elevation. The introduction section is quite comprehensive. It is mentioned that it may be possible that the introduced method may be used to improve SIT estimation by lower resolution / larger footprint laser instruments (ICESat-2). The results have mostly been presented nicely and comprehensively.

The first referee already submitted quite comprehensive comments on the manuscript, and I'll just try to complement his comments. I agree with him that a major review is still required before publication

Here are some comments trying to improve the manuscript:
General comments:
1) I agree with the reviewer 1 that more results of the DCNN approach could be included.
We have added an additional figure into the DCNN Results section showing the spatial distribution (Fig. 9)

2) Also a more detailed technical description of the applied methods (DCNN) would be preferable, as already suggested by referee 1. This could be a Section of its own (not an Appendix). Also include the information of numbers of DCNN neurons used at each layer and how these numbers were selected.
We have added a Methods section with most of the material that was in the appendix.

3) Regarding e.g. icesat-2 data, it would be nice to have some experiments or at least approximation related to the effect of resolution to SIT estimation using the proposed method.
Unfortunately, ICESat-2 data is linear and not suitable as an input for our ConvNet, although see the response to reviewer #2 above regarding how these results provide a demonstration of deep learning techniques and a possible path to an improved ICESat-2 algorithm. We have discussed a halving of the resolution and its effect on the accuracy in the Results. We cannot reduce the resolution too much as each lidar scan is only 100m x 100m, which limits how large our window can be. We are now exploring this with ICEBridge data, but this will be a subsequent paper.

More detailed comments:
S. Introduction P5, L17-18 "... detailed snow depth measurement": Also include already here by which method the snow depth measurements were made (not in detail).
Added "manually-probed"

S. Data P6, L12 and L22: instruments are named, also include references to their technical specs, and also shortly write on the principle of the snow measuring device.
Added.

S. Data P 6-7: Division of the data sets used into training and test data sets (possibly also validation data set) could be clearly described in the data section already. Were the data sets the same for all the performed experiments? This seems to be described later in the deep learning section for the DCNN.
We have added a Methods section which describes this training/validation procedure and the test data set

S 3.1.1 P9, L18-19: Rather say "...Thickness of the level ice (L) forming a sail and its sail height (S)..."
Fixed, thanks.

S 3.1.1 P10, L8 "...for estimating sea ice thickness,..." -> "...for estimating sea ice thickness T,...". Possibly You could use SIT for sea ice thickness throughout the manuscript?
We have replaced sea ice thickness with SIT.

3.1.2 title could be "...mean sea ice thickness..." or "...mean ice thickness..." or even "mean SIT".
Changed to SIT

P10, Fig. 5. Make the figure larger, difficult to read in the printed version. Its width could e.g. be approximately the column width.
OK – done.

P11 S 3.1.2 L22: Describe the use of semivariogram in more detail. Did You make any experiments by varying the window size also?
We used the semivariogram to identify the optimal window size. We did try a half-sized window, as described in the text, but with somewhat worse results, likely because the windows fail to capture surface features.

P12 Fig. 6: Same thing as for Fig. 5, make larger.

OK – done

P14 Fig. 7: Same thing as for Fig. 5, make larger.
OK – done

P18 Fig. 8: Make the figure larger or make the box frames wider for better visibility. Include a legend describing the classes instead of writing it in the caption.
OK, done.

S 3.3, P 19 L5: The best-performing linear regression result has been given here for comparison. Have You any idea, could better results have been achieved by using a nonlinear approach with the same inputs, e.g. a multilayer perceptron neural network with the same inputs (plus an additional constant/intercept input)? Or are the dependencies really linear?
MLPs actually are less effective but more complicated than ConvNets due to their fully-connected style. This means the total number of parameters quickly becomes very high. We also use a nonlinear activation function in our ConvNet, as the dependencies are nonlinear, as you note. Moreover, we want a convolutional approach precisely because we believe the spatial information in the lidar 'image' is important and necessary for accurate SIT estimation.

P19 L12-13: "20m x20m windows", also give the window size in pixels here. Did You study the effect of the resolution to the result by using down-sampled data? Any idea, how would this possibly affect the estimation result? Possibly You could then get average SIT over a larger area? This could give an idea of the applicability of the method to coarser resolution data.
Yes, we have tried to halve the resolution. This is mentioned in the text (was formerly in the appendix) and results given (a modest degradation in performance). As mentioned above, we cannot keep halving the resolution as then our dataset becomes too small to do any meaningful convolution. We are attempting to do this in future studies using Operation IceBridge data.

Figs. 9, 11,12,1314 and 15: make bigger for better readability in the printed version.
Ok – done.

App. A: did You also vary the number of neurons at each level and how did this affect to the results? How were these parameters selected? Does there exist any "rules of thumb" for selecting the parameters (e.g. numbers of neurons) for DCNN's as there exist for Multilayer Perceptrons (as a function of the number of inputs and outputs)
We varied the number of filters at each layer and if there were too few, then the results were worse. There are no rules of thumb, other than to double the number of filters in each layer if the stride is 2 (as the dimensionality of the data is halved), which we did. Again, we stress that our architecture could be fine-tuned to improve accuracy even more, and we simply aim to show that this method can be applied to improve SIT estimates.

A: Also include execution times for the training and SIT estimation in the used hardware.
Added.

And yet one interesting aspect: As a researcher of microwave and optical EO imagery (over sea ice) I am also interested in possibilities of utilizing the existing imaging devices for SIT estimation. Typical high-resolution (HR) sensors covering a wide spatial area, such as HR SAR or optical/IR sensors, measure only the 2-D sea ice surface, not the elevation directly. However, it is possible to locate ice ridges and even estimate their sail width in HR EO imagery. There is some literature (e.g. Timco & Burden, 1997) relating the ridge parameters to each. However, I have not seen any good reference relating sail width (Ws) to sail height (Hs). This kind of relationship would be very useful for better estimating ice thickness from 2-D HR EO data. Could the authors comment on this topic i.e. how (well) the morphology could be derived/estimated from the available 2-D EO data/imagery and whether this relation could be utilized in SIT estimation? Possibly a deep neural network could be used after deriving some

ridge parameters from 2-D HR sea ice data form SAR/optical/IR, or even just training a DCNN with the data directly. This would naturally require a good data set with a large number of (nearly) simultaneous SIT measurements (possibly made by another validated remote sensing method, such as laser scanning).]

We agree convnets might have utility to determine potential relationships with other metrics and sea ice thickness. For example, it is reasonable to think that spatial variability in SAR signatures might be correlated with deformed ice percentage, ice types, etc, which are likely correlated with ice thickness. Alternatively, HR imagery may show features that are indicative of snow dunes or ridges. However, at present we have little basis to expect that any such relationships might be sufficiently reliable, and without carrying out such analysis, we feel this is too speculative to comment on. To take the example suggested by the reviewer, our experience with analysis of ridge morphology in the Antarctic (from our data and Icebridge) we have not seen any suggestion of a relationship between the sail height and width. Ridges identified in imagery may vary from well-behaved triangular ridges which may exhibit some relationship, to rubble fields, which likely do not. Note also that in our case, it appears that the CNN appears to heavily use the freeboard; without any freeboard information, we do not expect a CNN to be very accurate in predicting thickness. That said, we would agree that a CNN would likely be effective in identifying ice types (in analogy to how a trained analyst does this).

Since any suggestions here would be very speculative, we prefer to not discuss here, although we have added additional discussion in the conclusions relevant to ICESat-2, as requested by the other reviewers.

Estimating Early-Winter Antarctic Sea Ice Thickness From Deformed Ice Morphology

M. Jeffrey Mei[1,2], Ted Maksym[1], Blake Weissling[3], and Hanumant Singh[4]

[1]Department of Applied Ocean Science and Engineering, Woods Hole Oceanographic Institution, MA 02540, United States
[2]Department of Mechanical Engineering, Massachusetts Institute of Technology, MA 02139, United States
[3]Department of Geological Sciences, University of Texas El Paso, TX 79968, United States
[4]Department of Electrical and Computer Engineering, Northeastern University, MA 02115, United States

Correspondence to: Jeffrey Mei (mjmei@mit.edu)

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 ~~surface elevation~~ 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 ~~sea ice thickness~~ 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 ~~sea ice thickness~~ 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 ~~sea ice thickness~~ 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 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~~, which is much lower than the mean relative error for a linear fit (errors up to 47%)~~. 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 ~~sea ice thickness~~ SIT from ICESat-2 over current methods, especially at smaller ~~scale~~scales.

1 Introduction

Satellites have documented changes in sea ice extent (SIE) for decades (Parkinson and Cavalieri, 2012); however, sea ice thickness (SIT) is much harder to measure remotely. Declines in Arctic ~~sea ice thickness~~ SIT over the past several decades have been detected in under-ice upward-looking sonar 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 Cunningham, 2015), but any possible trends in Antarctic SIT are difficult to detect because of the presumably

relatively small changes, and difficulties in estimating ~~sea ice thickness~~ SIT in the Antarctic (Kurtz and Markus, 2012; Zwally et al., 2008). Because fully-coupled models 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 ~~sea ice thickness~~ 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 ~~witch~~ with any confidence the relatively small trends in thickness expected (Massonnet et al., 2013), or even interrannual 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; Ozsoy-Cicek 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)~~(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).

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 ~~surface elevation~~ 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 ~~surface elevation~~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 ~~surface elevation~~ snow freeboard F (i.e. snow depth + ice freeboard, ~~sometimes called snow freeboard;~~ see Fig. 1) and snow depth D measurements using the relation

$$T = \frac{\rho_w}{\rho_w - \rho_i} F - \frac{\rho_w - \rho_s}{\rho_w - \rho_i} D \tag{1}$$

for some densities of ice, water and snow ρ_i, ρ_w, ρ_s (Fig. 1). ~~In this article, freeboard refers exclusively to ice freeboard.~~ 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 (e.g. Kwok and Maksym, 2014), some assumption of snow depth has to be made, or an empirical fit to field observations is needed ~~(e.g Ozsoy-Cicek et al., 2013)~~(e.g. Ozsoy-Cicek et al., 2013). When averaging over multiple kilometers, and in particular during spring, it is common to assume that there is no ice component in the ~~surface elevation~~snow freeboard, i.e. $F = D$ in Eq. 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 (Ozsoy-Cicek et al., 2013), but these implicitly assume a constant proportion of snow within the ~~surface elevation~~ 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 ~~surface elevation~~snow freeboard) inferred from the empirical fits of Ozsoy-Cicek et al. (2013)~~and Xie et al. (2011)~~, which is equivalent to a more complex, regime-dependent set of snow assumptions.

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), 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). ~~Around such~~ The sea ice surface naturally has a varying proportion of deformed ice, which affects the sampling required to faithfully represent the distribution (Weissling and Ackley, 2011). Around deformed areas, both the ice freeboard and snow depth may be high, and we do not yet ~~understand the~~ know the statistical distribution of snow around such deformation features. ~~This means that~~ In this respect, local estimates of SIT are likely biased low as the average ice freeboard cannot be assumed to be zero. 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; 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 ~~surface elevation~~ snow freeboard measurements.

Many pressure ridges can be observed from above using airborne or terrestrial lidar scans ~~(e.g Dierking, 1995)~~(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 ~~surface elevation~~ snow freeboard measured by a lidar scan. Furthermore, the corresponding keel morphology given some surface (lidar) scan, and its effect on the SIT distribution, is not known. Among other factors, radar-based estimates of snow depth are known to be highly sensitive to surface roughness, weather and grain size (Stroeve et al., 2006; Markus and Cavalieri, 1998). ~~Kern and Spreen (2015)~~ Ozsoy-Cicek 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 underesti-

mated by a factor of two or more~~, and found an estimate of Antarctic sea ice thickness error from ICESat of around 50%, which is not easily reduced due to AMSR-E snow depths not reporting any uncertainty~~. Kern and Spreen (2015) also showed that the error estimate in the ~~sea ice thickness~~ 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%.

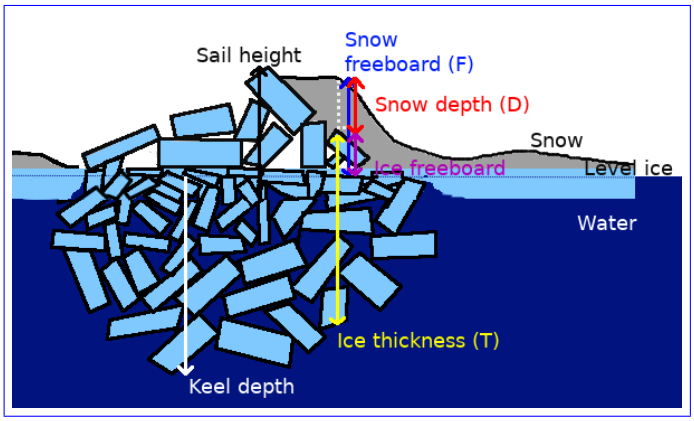


Figure 1. A schematic diagram of a typical first-year ridge. The ridge may not be symmetric, and peaks of the sail and keel may not coincide. The effective density of the ice is affected by the air gaps above water and the water gaps below water. T, D and F may be linked by assuming hydrostatic balance (Eq. 1).

Sea ice draft and ridge morphology may also be observed from below using sonar on autonomous underwater vehicles (AUVs) ~~(e.g Williams et al., 2015). As most of the sea ice is below water, using the mean draft as a direct estimate of the SIT gives lower errors than surface elevation-based methods. Moreover, the underside of the deformed ice surface does not have snow, making the morphological features less obscured.~~ (e.g. Williams et al., 2015). 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 ~~surface elevation~~ snow freeboard much better than with linear, low-resolution drilling profiles.

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,

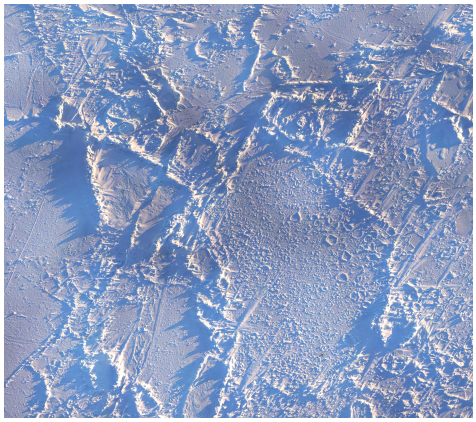


Figure 2. Drone imagery (180 m x 180 m) of heavily deformed ice in the Ross Sea, Antarctica. There are multiple ridges which cannot be easily separated. The ridge widths and slopes are varying and must be arbitrarily defined, leading to a variety of possible values. Image provided by Guy Williams.

2003), though much thicker (maximum keel depths $>$ 15 m) ridges have also been observed with AUVs (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. 2), 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 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.~~ Identifying how the morphology of deformed ice can inform estimates of ~~sea ice thickness~~ SIT is important for ~~reduce~~ reducing errors on SIT estimates~~, which~~ . 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 ~~sea ice thickness~~ SIT estimates (Kern and Spreen, 2015). For example, if assuming zero ice freeboard ($F = D$ in Eq. 1) with some known snow density, a 10% uncertainty in the sea ice density can lead to a 50% uncertainty in the ~~sea ice thickness~~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)~~(e.g. Worby et al., 2008

, 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 ~~sea ice thickness~~ 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.

In this paper, we aim to use a high-resolution dataset of deformed sea ice to develop better algorithms to estimate ~~sea ice thickness~~ SIT from surface topography. Unlike previous studies which have relied on low-resolution, 2D drilling transects, we use high-resolution, 3D characterization of the snow surface from terrestrial lidar, coincident with 3D ice draft from an autonomous underwater vehicle and detailed manually-probed snow depth measurements. In particular, having 3D coverage allows for the analysis of complex morphological features. ~~We first analyze our dataset to examine the morphology of the surveyed first-year ridges and potential relationships with ice thickness. Second~~First, we examine simple statistical relationships between ~~surface elevation~~ snow freeboard, snow depth and ice thickness~~and simple measures of local morphology~~, and compare with prior studies. We also estimate ~~effective~~ densities of ice~~and snow~~ /snow by comparing the fits with Eq. 1 and compare with field data. ~~Lastly~~Next, we use a deep learning convolutional neural network to improve estimates of local ice thickness by using complex, non-linear functions of 3D surface morphology. ~~We then discuss~~ Finally, we discuss the linear and ConvNet models and attempt to interpret how learned features in the neural network may be related to physically-meaningful morphological features, and consider possible extensions to this work on larger datasets.

Our goal here is to test whether complex surface morphological information can be used to improve sea ice thickness estimation. In this paper, we demonstrate this using high-resolution spatial surface topography, which is most applicable to airborne remote sensing data such as that obtained by NASA's Operation IceBridge (Kurtz, 2013). While a somewhat different approach would be required for linear data such as that obtained from the ICESat-2, this paper is a first test of proof-of-concept that using such information may be beneficial.

2 Data and Processing

The PIPERS (Polynas, Ice Production, and seasonal Evolution in the Ross Sea) expedition took place from early April to early June 2017 (Fig. 3). In total, 6 AUV ice draft surveys were taken of the undersides of deformed sea ice. Of these, 4 coincided with snow depth measurements and a lidar survey of the ~~surface elevation~~snow freeboard, 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. 4. There are two other AUV scans which lack lidar/snow measurements ~~but are included for draft-related~~ so are not included in our analysis. The AUV ~~scans were done using the~~ surveys were done with a Seabed-class AUV from the Woods Hole Oceanographic Institution ~~equipped~~ following Williams et al. (2015), with a swath multibeam sonar (Imagenex 837 DeltaT) at a depth of 15-20m 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 ~~was~~ were processed to correct for vehicle pose, then individual swaths were stitched together, with manual corrections to pitch and roll offsets of the sensors to minimize 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. The vertical error ~~for~~ in draft is estimated at 10 cm over deformed areas and <~~1~~ 3 cm for level areas (Williams et al., 2015). The scans were ultimately binned at 0.2 m horizontal resolution. The ~~surface elevation~~ snow freeboard scans were done with a Riegl ~~VZ-400 lidar , using four~~ VZ-1000 terrestial 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 300kHz, with an effective maximum range of 450m. 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 ~~product~~ 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. It was fitted 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. 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 ~~changing snow surfaces (near deformed ice)~~ deformed ice and lower resolution (~5m) over flat, level topography. 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. The ice thickness can then be calculated by taking (draft) + (~~surface elevation~~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 neglible because of the much thicker ice.

~~In addition to these 6 AUV scans from PIPERS, there were 14 additional AUV scans from other experiments (3 from SIPEX-II, 5 from IceBell and 6 from SeaState) combined for analysis (Williams et al., 2015; Thomson et al., 2018). SIPEX-II took place in East Antarctica in September 2012, IceBell in November 2010 in the Bellinghausen~~

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. 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. The order of the lidar correction is ~1 cm and the order of the AUV correction is

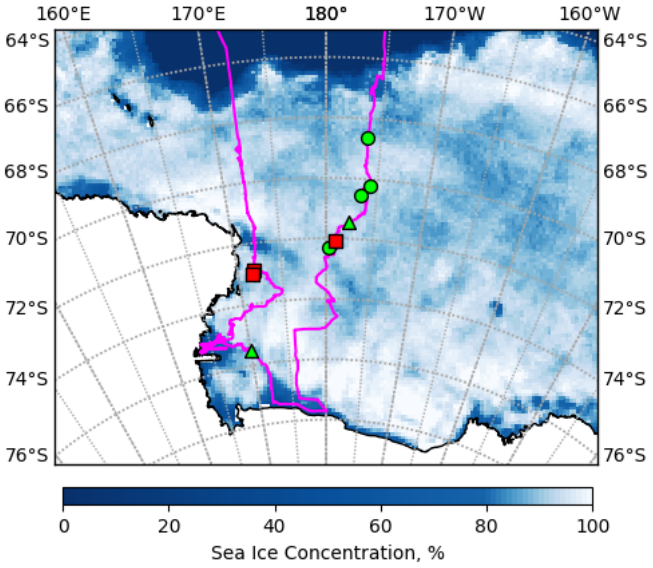


Figure 3. PIPERS track (magenta) with locations of ice stations labeled. ~~The~~ Stations with AUV scans ~~used in this paper~~ are shown in green (3, 4, 6, 7, 8 and 9) and the other stations (1, 2 and 5) are shown with red squares. Stations 4, 7, 8 and 9 (green circles) also have a ~~surface elevation~~ snow freeboard scan and snow depth measurements; these are shown in Fig. 4. 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).

~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.

Summary statistics for the floes sampled during PIPERS are in Table 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., 2 . The sail/~~Weddell seas, and SeaState in October 2015 in~~ 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 ~~Chukchi~~sail/~~Beaufort seas. Note that SeaState surveyed Arctic ice, but we include it for comparison as the AUV scans are primarily of thin, first-year ice. Tin and Jeffries (2003)found that Antarctic ridges are morphologically comparable to those from~~ 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~~which largely form from first-year ice.~~

3 ~~Results~~

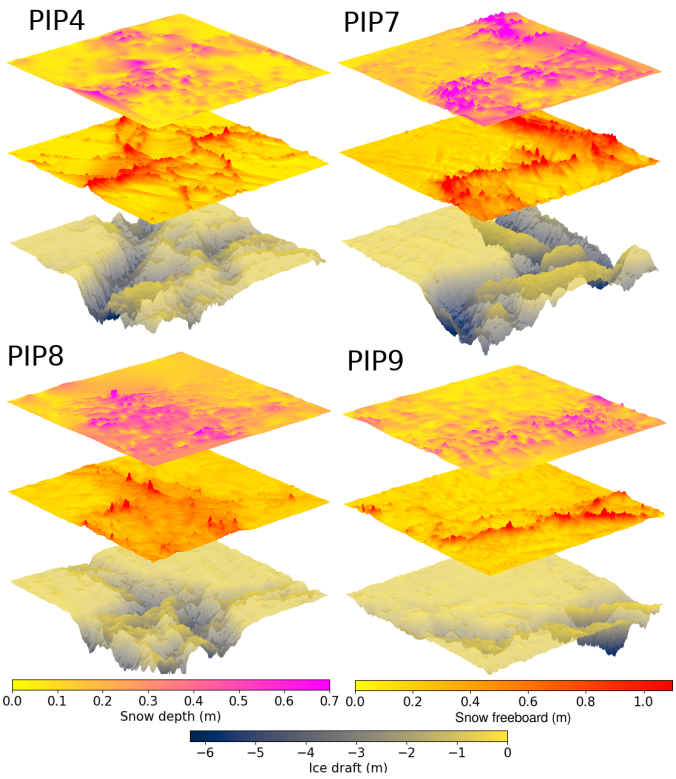


Figure 4. ~~Layer~~ Sea ice/snow layer cakes from PIPERS. The top layer is the snow depth (D), the middle layer is the lidar scan of the ~~surface elevation~~ snow freeboard (F), and the bottom layer is the AUV scan of the ice draft. The ice thickness is therefore given by ice draft + ~~surface elevation~~ snow freeboard - snow depth.

~~We attempt to statistically model sea ice thickness using surface-measurable metrics (e.g. mean and standard deviation of the surface elevation), in order to see the limitations of this method.~~ . 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.

3 Methods

3.1 ~~Estimation of Sea Ice Thickness With Surface-Based Metrics~~Linear regression approach

~~Following previous literature, we expect some relationship between sea ice morphology and thickness. Previous studies have used low-resolution, 2D surveys using drill lines and have found various correlations between certain metrics and the level/deformed ice thickness.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). For our PIPERS~~

Table 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 ~~surface elevation~~snow freeboard, σ), mean ~~surface elevation~~ 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

~~data, we calculate standard metrics such as sail/keel height and angle, summarized in Table 1. The sail/keel angles are not as well-defined for non-linear ridges, so a range of angles is given. 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. We compare the relationships for predicting thickness from prior studies that were fitted on low-resolution, 2D datasets with our higher-resolution, 3D dataset to see if the same relationships still hold.~~

3.1.1 ~~Estimating level ice thickness from keel depth~~

~~The simplest way to account for the morphology of a deformed surface is to simply measure its maximum height, as has often been done when quantifying ridge statistics from airborne lidar surveys (e.g Dierking, 1995; Petty et al., 2016). A common way of reporting this is by taking the ratio of the sail height and keel depth. Using the 99th percentile values for the sail/keel from Table 1, the ratio of keel depth and snow-sail height for our PIPERS dataset is 3.9, in line with a ratio of 3.6~~ 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 ~~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). Tin and Jeffries (2001a) also found a ratio of 29.6 for ice keel area/ice sail area, and a ratio of 10.4 for ice keel area/snow sail area. For our corresponding 3D dataset, our ice keel volume/ice sail volume ratios range from 11.6-19.0 and our ice keel volume/snow sail volume ratios from 4.6-6.4. Our ratios are somewhat lower than those of~~

~~Tin and Jeffries (2001a), perhaps because drill line measurements of snow~~one season/~~ice freeboard tend to be biased low due to selection bias.~~

~~It may also be possible to infer the level ice thickness given a measured sail height. 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 Arctic ridges in the Beaufort Sea (Tucker III et al., 1984; Melling and Riedel, 1996). When fitted to our PIPERS AUV dataset , we get $a = 6.7 \pm 0.7$. Note that here we use the mean draft of the level ice from the AUV dataset~~region, which is ~~very close to the mean thickness (and indeed, for early winter, over level ice, the $F = D$ assumption should be approximately true). 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 our full AUV dataset (20 scans, see Section 2) spanning a much wider range of keel depths (Fig. **??**). 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 by choosing the lowest AIC (Akaike Information Criterion, see 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 both of these are much lower than the coefficients found for Arctic ridges (15-20), 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, although 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.~~ important so that the ridges have consistent morphology.

~~This relationship could potentially be used the other way, by measuring the extreme value (sail height) and inferring the level ice thickness . Sail heights and deformed ice proportion are recorded when taking underway observations (ASPeCt)to estimate ice thickness, and so our empirical relationship may be applied to these data to give a basic estimate of sea ice volume in the Ross Sea. 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. This relationship was~~ 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, 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 ~~by initially working out the snow sail mean to draft mean ratio as 5.1, and then correcting for snow obscuring the deformed surface area to obtain a corrected ratio of 4.4. Using our own values for mean ice draft/freeboard, we coincidentally get the same corrected ratio of 4.4, and so we could expect the relationship above to hold for our dataset. However, our ridges are non-triangular, so we are not surprised to find that this relationship does not really hold with our PIPERS dataset, with a MRE of 56% (using the 99th percentile values for the sail height, which would be closer than the 100th percentile value to the effective height of an equivalent-volume triangular sail). Further analysis using surface roughness as a variable to estimate ice thicknessis detailed in Sections **??** and 4.1.3~~to highlight uncertainty when used with different ice conditions, and to our ConvNet predictions of ice thickness.

~~Level ice (draft) 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 the square-root fit. The mean relative errors (MRE) in predicting keel depths compares similarly, with MREs of 41%, 24% and 22% for the linear, square-root and monomial fits.~~

3.1.1 ~~Estimating mean thickness with surface roughness~~

~~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 (i.e. the~~ In order to motivate more complex methods in subsequent sections, we also use surface roughness (standard deviation, σ) ~~of the snow surface and its thickness, and also a linear relationship between the σ of the surface and σ of the draft, for 3 Ross Sea datasets from summer and autumn. Taking their ratio, we can estimate the linear relationship between the σ of the draft and the ice thickness, which we can compare to our AUV data (Fig. **??**). Tin and Jeffries found that the survey-wide mean thickness was 5.5 times the survey-wide snow-surface roughness, and their snow-surface roughness was 1/3.7 of the ice bottom roughness. So, approximately, the survey-wide mean thickness would be 5.5/3.7 times the ice bottom roughness, giving a factor of 1.5. As not all our AUV datasets have corresponding lidar/snow data, we use the mean draft as an estimate of the mean thickness. We find the same ratio of (survey-wide mean thickness)/(survey-wide bottom roughness) = 1.5±0.1 for our AUV dataset (Fig. **??**) . Here, our survey-wide scale is very close to the floe scale of Tin and Jeffries (2001b), but our surveys are not necessarily representative of the whole floe due to deliberate selection of deformed ice. To avoid confusion, we refer to our large-scale statistics as survey-wide instead of floe-wide; each survey (e.g. PIPERS) is comprised of multiple scans (e.g. PIP4, PIP7), from which smaller windows are taken.~~ to predict thickness, to demonstrate that surface morphological characteristics have some information that can be used to predict thickness.

~~Due to the seasonally- and regionally-varying snow cover, the relationship between snow-surface roughness and bottom roughness is unlikely to be consistent across different climatologies, so we only analyze the thickness~~

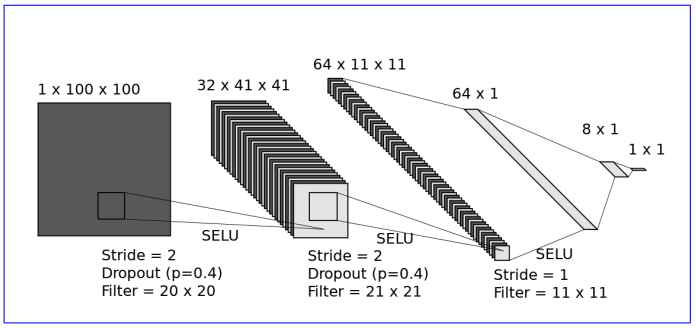


Figure 5. ConvNet architecture, using 3 convolutional layers and 2 fully-connected layers, for predicting the mean thickness (1 x 1 output) of a 20 m x 20 m (100 x 100 input) lidar scan window at 0.2 m resolution (LeNail, 2019). The (64 x 1) layer is made by reshaping the (64 x 1 x 1) output of the final convolutional layer, and so is visually combined into one layer. 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}.

3.2 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-surface roughness) ratio for the four full ice stations from Fig. 4 in the PIPERS dataset. This gives a ratio of 8.2 ± 0.5~~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. 5. The input consists of 20 x 20 m (100 x 100 pixel) windows, with 3 convolutional layers, with a ~~mean relative error of the mean thickness of 12%, which is higher than the ratio of 5.5 from Tin and Jeffries (2001b) . In this paper, we use the mean relative (percentage) error (MRE) of the predictions instead of the mean absolute error to prevent weighting prediction errors in thicker ice differently to thinner ice, which is important as the majority of sea ice surfaces are thinner, undeformed ice. The ratio of (survey-wide surface roughness) /(survey-wide bottom roughness) is $1/(6.5 \pm 0.5)$, which is also different to the corresponding ratio of 1~~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. We also used dropout (p=0.4) and augmentation (random 90^o rotations, horizontal~~/3.7 from Tin and Jeffries (2001b). Interestingly, combining them to get a ratio of (survey-wide mean thickness)~~vertical flipping) to reduce overfitting (Srivastava et al., 2014). An overview of ConvNet basics and full implementation details are given in the Appendix.

The training/~~(survey-wide bottom roughness) gives a very similar value of 1.3 ± 0.1, which agrees well with their value of 1.5. This may be because our scans captures larger features than a drill line can, and these may have different roughness values. Moreover, drill lines may suffer sampling biases as previously discussed.~~

~~We repeat the analysis using local snow surface roughness (σ of a 20m x 20m window at resolution 0.2 m) and local mean draft (Fig. **??**b) instead of scan-wide statistics. The window size was chosen~~ 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 points) for our analysis. ~~The MREs using the snow surface roughness to predict mean local thickness range from 23-37%when fitting for each survey separately. For comparison, we also try fitting the local mean thickness to the local draft roughness, which has higher MREs of 31-48% . In general, rougher surfaces correspond to thicker ice, although the nature of this relationship may be nonlinear at higher resolutions.~~

~~Similarly, we may expect rougher areas to trap more snow (Massom et al., 2001; Kwok and Maksym, 2014). Although Kwok and Maksym (2014) averaged the snow depth and surface roughness over a much larger scale (4 km scale at resolutions 1-10 m)~~ 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. Although the training error is directly analagous to the fit error for linear models for some dataset, ~~we also find snow accumulates preferentially in areas of deformation. We find that snow depth at a 20 m scale can be approximated by a linear function of the surface roughness (slope: 0.80, intercept: 0.12 m). This is a similar relationship to what they found (slope: 0.83-1.25, intercept: 0.07-0.18 m), despite their dataset being from a different region (Weddell/Bellinghausen Seas) and season (spring) than ours (Ross Sea/winter). 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 techniques to analyze deformed surfaces. RMS roughness does not account for spatial features, as any permutation of points within a grid would have the same standard deviation.~~

~~(a) Survey-wide RMS roughness (σ) vs. floe mean draft 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 dotted 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-II is from early spring. Fits to the individual datasets are color-coded. The MRE in the predicted mean thicknesses are 11%, 17%, 12%, and 18% for IceBell, PIPERS, SIPEX-II and SeaState respectively, and 33% for all data. (b) Local σ of the snow surface vs. local mean thickness, for PIPERS data only. The MREs for predicting mean local thickness range from 23-37%, with the fit for all PIPERS data (black dotted line) having a MRE of 33%. This is slightly lower than the MRE of 49% for predicting mean draft thickness using local draft roughness from PIPERS (not shown).~~

These results relating ridge morphology to other metrics largely agree (with some exceptions) with prior literature, despite the difference in resolutions and ridge thicknesses. This suggests that surface morphology can be related, at least to a limited extent, to sea ice thickness, and potentially improve on simple linear predictions of ice thickness. This is discussed in further in Section 4.1.3.

3.3 Estimating thickness using hydrostatic balance

To accurately calculate sea ice thickness without making assumptions of snow distribution, we need to use combined measurements of ice draft (AUV), surface elevation (lidar) and snow depth (probe). Here, we primarily use PIPERS data to focus on early-winter floes, and also because this is it is much easier to overfit with a ConvNet as the largest such datasetfrom the same season/region, which is important so that the ridges have consistent morphology. The lidar and AUV data were corrected by aligning the mean measurements of the level areas of the drill line. 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, and the presence of seawater-filled gaps that may be confused with the ice-ocean interface.

Assuming hydrostatic balance, Eq. 1 should hold for all datasets. However, drill lines have coarse resolution, which may not capture the local variability in elevation between drill points, and so drill points may not even be in hydrostatic balance. Drill lines are also only 2D, and so surface variations in the 3rd axis may not be accounted for in drilled points. Moreover, densities may vary spatially, in particular around deformed surfaces that may contain air/water gaps. Due to the difficulty in drilling sail peaks, freeboards from drill lines may also be undermeasured. Our 3D data should therefore more accurately sample ridged regions, and we should expect hydrostatic balance to hold.

Ozsoy-Cicek et al. (2013) compiled various Antarctic datasets to investigate the relationship between sea ice thickness, snow depth and surface elevation. Assuming hydrostatic equilibrium is reached over the window size (20 m), we use Eq. 1 to fit a regression for sea ice thickness (T) as a linear function of surface elevation (F), sometimes along with snow depth (D). This approach has been applied by Zwally et al. (2008); Worby et al. (2011); Yi et al. (2011); Xie et al. (2013) over a variety of scalestraining 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. From here onwards, analysis of the ConvNet refers to the one using PIP8 as a test set, though using a different one would yield qualitatively similar analysis.

4 Results

4.1 Linear model results

4.1.1 Fitting to surface elevation snow freeboard only

Although we have snow depth measurements in addition to surface elevation snow freeboard measurements, in general there are far fewer snow data and so we first try to fit with just surface elevationsnow freeboard, by making some snow depth assumptions.

For level topography, where the snow = surface elevation assumption is supposedly valid (set $F = D$ in Equation 1), Eq. 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 negligble ($D = 0$), which may be the case at the sail peak, Eq. 1 becomes $T = 9.4F$. These values become lower and upper bounds for fitting k in $T = kF$.

All our coefficients, which range from 2.9-6.1, fall between these two extremes of snow-only freeboard and ice-only freeboard, with the lowest value of 2.9 being for level topography, as expected. Our sampled areas are likely not representative of typical area-averaged deformation rates of sea ice due to these survey areas being selected for their heavy deformation, and so the fitted coefficients for individual floes and the "All" category are considerably higher than 2.7. In contrast, the coefficients for F of 2.8-3.0 in Xie et al. (2011), and 2.2-3.1 in Ozsoy-Cicek et al. (2013) , suggest that at floe-wide and larger scales, there is enough level ice that the snow = surface elevation assumption is valid, at least for this region/season. It is also possible that these drill lines have undersampled ridged ice. Our coefficient of 5.79 is much higher, which suggests that there is some non-zero component of ice freeboard in the surface elevation measurements. For example, if we assume typical snow/ice densities of 300 kgm^{-3}/920 kgm^{-3}, we can estimate that snow, on average, comprises 54% of the measured surface elevation, which means Eq. 1 simplifies to $T = 5.8F$, as in Fig. 10. In further support of this, our dataset has mean snow depths for the four surveys ranging from 16-26 cm, and mean surface elevations ranging from 24-37 cm, implying non-zero mean ice freeboards from 6.5-11 cm. If the proportion of ice to snow were constant (and their effective densities, too), then the best-fit line would have no scatter. This is not the case in Fig. 10, and indeed the standard deviation of ice freeboardacross 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.

The sea ice thickness (T) as a function of measured surface elevation (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 2 is shown in black. Assuming mean snow and ice densities of 300 and 925 kgm^{-3}, this implies a mean proportion of 62% snow and 38% ice in the surface elevation. 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 kgm^{-3} (or effective ice density $<$915 kgm^{-3}.)

Note that our sampled region is not representative of the floe, as we have deliberately chosen a heavily-deformed area, and so the amount of ice freeboard is higher in our survey, in contrast to large-scale averaged surface elevations like Xie et al. (2013) and Ozsoy-Cicek et al. (2013). This suggests that at large scales for some seasons/regions, it may be reasonable to assume that the mean freeboard is zero, but this is not the case at smaller scales.

Ozsoy-Cicek et al. (2013) found the fitted linear regression between T and F as $T = 2.45F + 0.21$ for a winter Ross Sea dataset. With our PIPERS dataset, our equivalent fit is $T = 7.67F - 0.73$, with 23% MRE. Using the relationship from Ozsoy-Cicek et al. (2013) on our dataset, the MRE is 36%, and the error in estimating the overall survey mean thickness is

41%, despite being from the same climatology. This means that relationships from other datasets from the same region/season do not generalize well, especially if the proportion of deformed ice (and hence nonzero freeboard) is significant.

It is possible to interpret our negative intercept as a bias due to fitting a linear model across two roughness regimes. From above, the two regime extremes (no-ice vs. no-snow contribution to surface elevation) give coefficients of 2.7 and 9.4 for F. 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). 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 no-intercept fit with F for the level category, which is essentially a null fit (as over 90% of the thickness values are clustered around 0.5 m, with surface elevations varying over a narrow 5 cm range). In contrast, the coefficients for F from Ozsoy-Cicek et al. (2013); Xie et al. (2011) are all ~3, because these studies average over multiple floes and have a sufficiently small proportion of deformed surface area to assume a negligible mean freeboard. This could also explain why their intercepts are positive.

4.1.2 Fitting to surface elevation and snow depth

Using typical values of 910 kgm^{-3} for ice density, 1027 kgm^{-3} for water density and 323 kgm^{-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 kgm^{-3}, 1023.9 kgm^{-3} and 300 kgm^{-3}, giving a freeboard coefficient of 9.4 and a snow coefficient of 6.7. We compare these to our results of the (multi)linear regressions in Table 2. We include fits with an additive constant, even though this is unphysical, to see how well a linear model can predict SIT. Our coefficients when fitting over all 4 floes are 10.4 for F This approach has been applied by Ozsoy-Cicek et al. (2013) and 6.8 for D, which are comparable to those inferred from Zwally et al. (2008), although there is considerable variation between the floes (7.9-10.6 Xie et al. (2013) in order to obtain empirical relationships between SIT and snow freeboard. All our fitted coefficients are shown in Table 2. 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 F; 3.9-6.3 for D). As discussed below, this suggests a lack of generalization in the fits. Assuming the density of seawater is fixed at 1028 kgm^{-3}, this gives bounds for the effective densities and standard errors of sea ice and snow as 929.4±3.5 kgm^{-3} and 356.3±57.2 kgm^{-3}. The snow density is in line with Sturm et al. (1998), which found mean densities of 360-390 kgm^{-3} during winter in the Ross Sea, as well as the measured snow densities from PIPERS (245-300 kgm^{-3}). The ice (effective) density errors here are only for our PIPERS dataset and may not apply to other samples from the Ross Sea in winter as the effective density is affected by the proportion of ridged ice. 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) and if the estimate of ρ_s decreases, ρ_i increases. For example, if ρ_i = 935 kgm^{-3} (unusually, but not impossibly high for the effective density of ridged ice, which includes some proportion of seawater - see Timco and Frederking (1996), and also note that this is the effective density, including some proportion of seawater), the best estimate for ρ_s becomes 312 kgm^{-3}, which is closer to

~~Although each individual floe has MREs of 10-20% when fitting with an intercept, the large variations in coefficients and constants suggests that the linear model does not generalize well between floes. For example, using the relationship from PIP7, PIP8 or PIP9 on PIP4 gives 30%, 33% or 35% MRE respectively, compared to 10% error for the PIP4 coefficients; using the PIP4, PIP7, or PIP8 fits on PIP9 gives 67%,~~ example, a difference in AIC of 70 ~~% or 43% mean error, compared with 21% for using PIP9 coefficients. Table 3 summarizes the fit and test errors for using each of the floes as the test set. The fit MREs range from 17-20% (fitting with~~ between the two best models in the 'All' category implies that the likelihood that the model with $\{F,$ constant~~) and 25-36% (without), and the test MREs range from 23-34% (with constant) and 12-59% (without). Using typical values for snow~~$\}$ 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~~/~~ice density from literature mentioned above (giving F and~~ snow ratio such that introducing D ~~coefficients of 8.8 and 6.0 (Worby et al., 2001) or 9.4 and 6.7 (Zwally et al., 2008))gives MREs of 26% in both cases, and errors in estimating overall mean thickness of $\sim 15\%$. The high variability in the test errors suggests that statistical relationships may not generalize to future datasets, even those from the same climatology. This is an important limitation of applying empirical fits from small datasets.~~

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. Ozsoy-Cicek et al. (2013) found that $T = 2.45F + 0.21$ for a early-spring Ross Sea dataset). Using the fitted relationship from Ozsoy-Cicek 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 5.1.

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. 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.

4.1.2 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 Ozsoy-Cicek et al. (2013), and the without-constant fit is intended to be compared against Eq. 1 to estimate sea ice/snow densities. The coefficients are reported in Table 2 and the fit/test MREs are reported in Table 3. 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. Using typical values of 910 kgm^{-3} for ice density, 1027 kgm^{-3} for water density and 323 kgm^{-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

Table 2. 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. 7) to incorporate roughness into the fits (Section 4.1.3). 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

kgm^{-3}, 1023.9 kgm^{-3} and 300 kgm^{-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 2).

Assuming a density of seawater during PIPERS of 1028 kgm^{-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±3.5 kgm^{-3} and 356.3±57.2 kgm^{-3}. The snow density is in line with Sturm et al. (1998), who found mean densities of 350 and 380 kgm^{-3} during autumn/early-winter and winter/spring, respectively, in the Ross Sea, as well as the measured snow densities from PIPERS (245-300 kgm^{-3}). The measured PIPERS snow densities may be biased low because they were measured at level areas, and possibly do not repesent 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

Table 3. A compilation of the MRE of different fitting methods. Coefficients for the linear fits are shown in Table 2 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 ~~(Section 4.2)~~ achieves the best generalization in the fit, even though there are individual anomalous cases. For example, the ~~linear, no-constant~~ F-only fit using ~~PIP4~~ PIP7 as a test set has a low test error ~~of 12% despite having a high~~ than fit error~~of 36%. This is~~ , which simply ~~coincidental in~~ means that the ~~scatter~~ average snow/ice ratio for ~~the fit~~ PIP7 is ~~so high that the best-fit coefficients end up close~~ similar to the ~~corresponding coefficients~~ averaged snow/ice ratio for the ~~PIP4 floe~~other floes. The ~~F-only (no constant) column is included as this~~ F only fit is ~~directly~~ most comparable to our ConvNet ~~method,~~ as ~~they~~ neither use ~~surface elevation as~~ the ~~only input (no~~ snow depth ~~) and maintains the zero freeboard = zero thickness condition~~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%**

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) and if the estimate of ρ_s decreases, ρ_i increases. For example, if ρ_i = 935 kgm^{-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 kgm^{-3}, which is closer to the measured 300 kgm^{-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 snow freeboard Ozsoy-Cicek et al. (e.g. 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 5.1, and motivates the introduction of surface roughness (σ) as an additional variable in our linear fit.

4.1.3 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 ~~(Kwok and Maksym, 2014), and we have previously shown that the~~ and the mean thickness(Kwok and Maksym, 2014; Tin and Jeffries, 2001b). Our data also show a reasonable relationship between SIT and

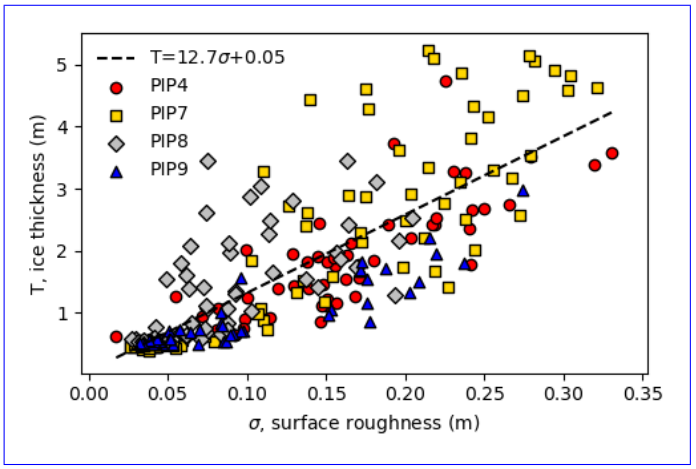


Figure 6. Predicting mean ice thickness with just the surface roughness (σ) as the input, with MRE 33%. The best-fit line is also shown, with R^2=0.65.

surface σ~~has some prediction power for the mean thickness~~ , though it is weaker than fits to the freeboard (Fig. ~~??b~~6). Adding the roughness as a third variable to the fit ~~, so that $T = c_1F + c_2D + c_3\sigma(+c_0)$, gives slightly lower fit MREs of 16-20% (with constant) and 24-34% (without) and slightly higher test MREs of 14-28% (with constant) and 28-58%(without)~~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~~. Furthermore, there~~ , 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 ~~surface elevation~~snow freeboard, even if it makes dimensional sense.

Instead, we can try using the roughness as a regime selector. To do this, ~~firstly~~ the lidar windows were classified manually into snowy surface, level surface, ridged surface and deformed surface categories (Fig. 7). 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 2.

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% ~~(with constant) and 17-19% (without),~~ assuming 20-50% of the surface is deformed. This is slightly better than for fitting the 'all' category in Table 2 (20% MRE~~with constant and 31% without~~), suggesting that distinguishing topographic regimes improves

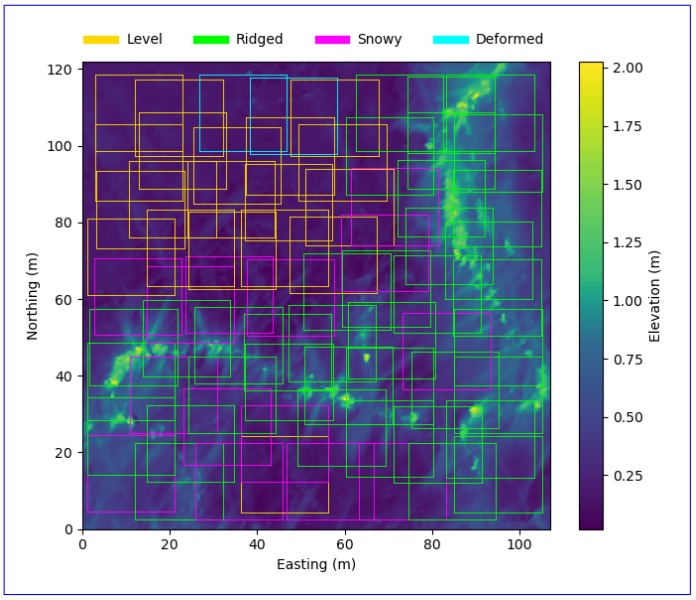


Figure 7. An example lidar scan from a station (~~PIP4~~PIP7) with the manually classified segments. ~~Yellow = level surface, green = ridged surface (possibly with snow), magenta = snowy surface and blue = deformed surface (excluded from analysis).~~ Snow features are clearly visible emanating from the L-shaped deformation. Deformed (blue) surfaces were excluded from the analysis.

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 (~~19-25% when fitting with constant, 34-36~~24% when fitting~~without, and in each case averaging the results over all possible test floe choices~~). This does not improve much on the generalization ~~if not using a two-regime model~~from the two-variable linear fit, where the test ~~MREs, again averaged over all test cases, are~~ MRE was 28%~~(with constant) and 34% (without) . Although the distinguishing of regimes may improve the model fits, it does not improve the test errors, again because this is likely too simplistic.~~ .

~~In reality, there is no reason why ice thickness should be non-zero given a zero surface elevation and snow depth. For all cases, the AIC and mean error is lower when fitting with an intercept, but a negative intercept, as in all our fits, sets a lower bound on values for F and D that return non-negative values of T. This reduces the applicability of these statistical models for thin ice. Taking the zero-intercept models only, estimating the survey-wide thickness has relatively low error (10-20%), but estimating local thickness has much higher errors (30-50%), which means that variations in ice thickness, such as those around deformation areas, cannot be precisely estimated. This motivates more advanced techniques to decrease the MRE, especially those that also maintain the physicality of zero surface elevation = zero thickness. Given that high surface elevation values may be either snowfeatures or ridges (leading to very different ice thicknesses), we need to distinguish these surface features in a~~

non-arbitrary way. In particular, we want to account for the complex deformation morphology, which we expect to be better predictors of thickness than the simplistic metrics used previously.

4.2 Predicting SIT with deep learning

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 sea ice thickness, as they are able to learn complex structural metrics, in addition to simplistic roughness metrics like σ. Our input is a windowed lidar scan (surface elevation) 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 best-performing linear regression has a mean error of 23%, though this includes an unphysical intercept, and also does not generalize to other datasets from the Ross Sea. We seek our model to improve on this error rate, while also being generalized enough to apply to different floes (from the same climatology), and also maintaining the physicality of zero surface elevation = zero thickness. Due to our limited dataset, comprised entirely of Ross Sea data in early winter, we do not expect our results to necessarily apply to other regions or seasons, which may have different snow distributions, ridging frequencies, or other causes of morphological differences. We intend simply to demonstrate the potential for these techniques to improve estimates of SIT. We do, however, expect our methods to generalize to new floes from the same region/season as our data.

Full details for our ConvNet are given in the Appendix. The training set consisted of randomly-selected 20 m x 20 m windows from three PIPERS ice stations. We chose 20 m windows, as in Section 3.1, by inspecting the semivariogram. 20% of the data were excluded from training so that it could be used as a validation set. The remaining floe 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 CNN was trained four times, with a different floe used as the test floe each time. These results are shown in Table 3. Although the training error is directly analagous to the fit error for linear models for some dataset, it is much easier to overfit with a ConvNet as the trainingerror can be made arbitrarily low. As a result, we compare our validationerror to the linear fit errors, and also use our test errors as a test of the generalization of our model. From here onwards, analysis of the ConvNet refers to the one using PIP8 as a test set, though using a different one would yield qualitatively similar analysis.

4.2 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.

The input windows were randomly flipped and rotated in integer multiples of 90^o to help improve model generalization, and dropout layers ($p = 0.4$) were added after the first and second convolutional layers to reduce overfitting (Srivastava et al., 2014)

~~.~~The best validation error ~~of 15.5%occurred at epoch 881~~was 15%, corresponding to a training error of ~~11.3~~11% (Fig. 8a and b). The mean test error (on the excluded floe) was ~~20.0~~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. 8c).

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 (~~23-24%~~ 24% for the 3-variable fit in Section 4.1.3) than our ConvNET (20%), the range of errors is much greater, with test MREs of ~~23-34~~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~~and also a constant term. These constants, as shown in Table 2, are all negative~~, which is generally not available. These fits also typically include a negative constant (Table 2), 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 ~~surface elevation~~ snow freeboard only, which is using the same input data as the ConvNet, have considerably higher ~~MREs (13-74%~~test MREs (23-39%, see Table 3). For sake of comparison to models that use RMS error such as Ozsoy-Cicek et al. (2013), the ~~RMS errors for our validation and test datasets were 6-11~~ validation RMS error for our survey-averaged mean thickness values is 2 cm, which is ~~considerably~~ lower than the RMS ~~errors of > 50 cm for single drill point measurements in drill lines given in Ozsoy-Cicek et al. (2013), and also lower than the RMS error of 26 cmfrom using one varying-density layer at a 70 m scale from Li et al. (2018).~~ error of 11-15 cm from Ozsoy-Cicek 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 Ozsoy-Cicek 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 ~~surface elevation~~snow freeboard.

~~It is not entirely fair to compare the single-point RMS errors to our survey-averaged RMS errors due to the different length scales; a better comparison is in estimating mean scan-wide thicknesses (~100 m length scale). This error should not be confused with the mean error in estimating the local window means, which is what is being optimized by the ConvNet. The mean error in estimating survey-wide thickness (averaging through all 4 datasets) is 5% for the training set, 6% for the validation set and 11% for the test set. The 5-11% MREs for the scan-wide mean thickness predictions correspond to RMS errors of 1-3 cm, which is much lower than the best-performing linear regression for a Ross Sea floe-wide dataset from Ozsoy-Cicek et al. (2013) with an RMS error of 16 cm.~~

As shown in Fig. 8c, 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. 8, red points/lines) despite having ~~only a slightly higher test MRE~~ similar fit MREs (Table 3). ~~In any case , the~~ The ConvNet also successfully reproduces the spatial variability of the SIT distribution better than the linear fit (Fig. 9). 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

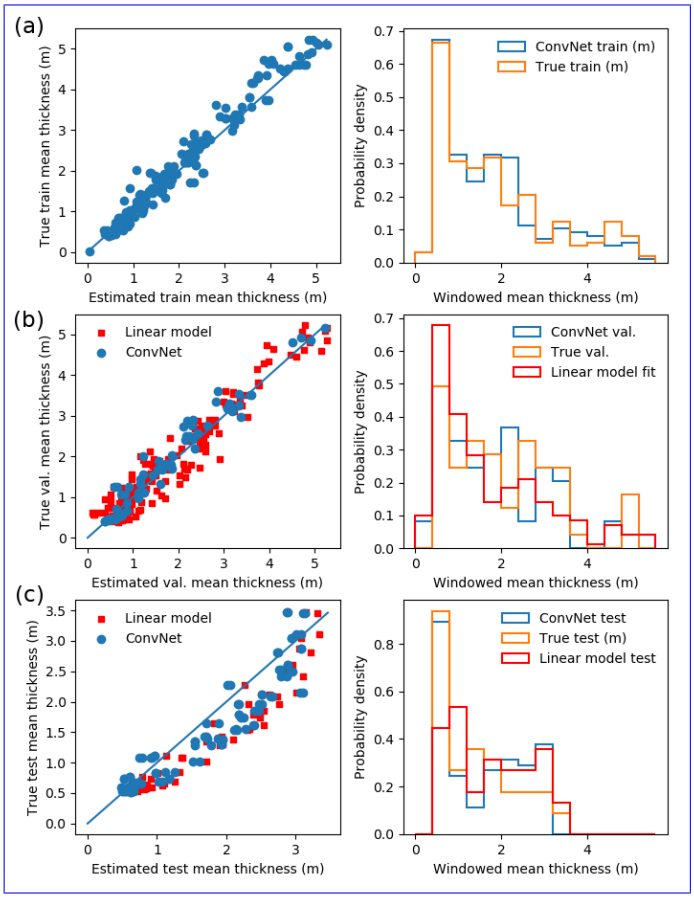


Figure 8. ConvNet results~~. The top panels show~~ , with (a) the learned ConvNet model applied to the training data (80% of randomly sampled 20m x 20m windows from PIP4, PIP7, ~~PIP8~~PIP9), with MRE 12%~~; the middle panels show~~ , (b) the learned ConvNet model applied to the validation data (remaining 20% of the randomly sampled 20m x 20m windows from ~~PIP7-9~~PIP4, PIP7, PIP9) with MRE 16% ~~; the bottom panels show~~ 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 ~~PIP9~~PIP8, as a check against learning self-similarity, with MRE ~~19~~20% ~~. The panels on the right show the resulting thickness distribution, both as a histogram~~ (ConvNet) and ~~as a continuous function. We also show the best~~ 32% (linear model~~fitted for the PIP4, PIP7~~)). In each case, ~~PIP8 in~~ the ~~middle panels (red) which has~~ left panel shows a ~~comparable MRE of 20%~~scatter plot with the predicted and true thicknesses, and the ~~results of this model applied to PIP9 (test floe) in~~ right panel shows the ~~bottom panels (red) (MRE: 34%)~~resulting thickness distribution. Our results suggest slight overfitting, as the test ~~scatter~~ error is higher than the training ~~scatter~~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).

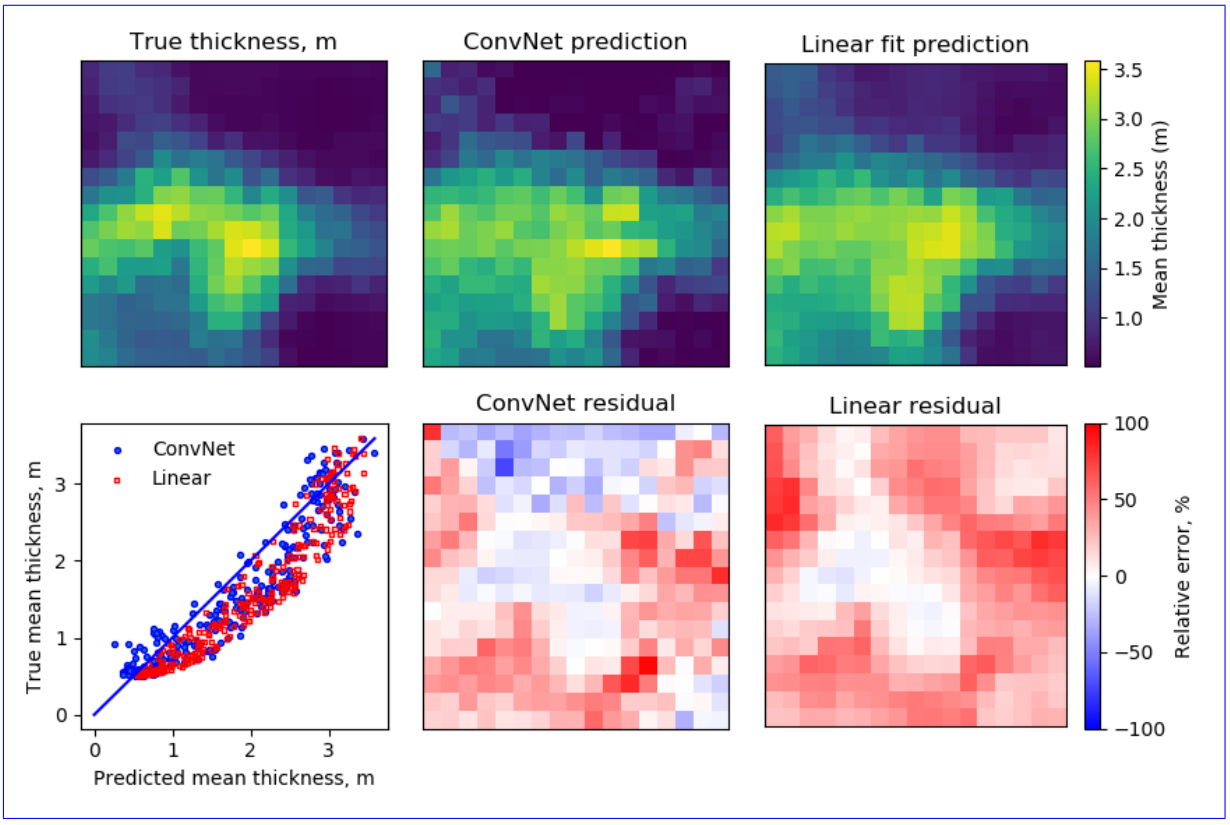


Figure 9. Ice thickness profile of the test set (PIP8), using the linear fit ($T = c_1 F + c_0$) and ConvNet model, both done with PIP4, 7 and 9 as inputs. The input windows are 20 m x 20 m, with a stride of 5 m in each direction, so there is a considerable oversampling. The mean residual for the linear 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 2) predictions are consistently biased high, which is also apparent in the linear model residual.

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 signficantly 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 ~~new floes~~ test sets from the same ~~climatology. An additional advantage of the ConvNet~~ region/season. In contrast, linear models, which do not generalize well to new datasets, have a considerable bias (Fig. 9), despite having an ostensibly good fit. Analysis of why the ConvNet may be performing better than linear fits is given in Section 5.2.

5 Discussion

5.1 Possible causes for poor linear fit

Our linear regression results for fitting $T = c_1 F + c_0$ have markedly different coefficients from drill line data from the same region/season (Ozsoy-Cicek et al., 2013). Here we discuss possible reasons for their differences. The first difference is that our value for $c_1 = 7.67$ (Table 2) is much higher. This is almost certainly because our 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), which for our data occurs for level thin ice where there is some snow load, Eq. 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 negligble ($D = 0$), which may be the case at the sail peak, Eq. 1 becomes $T = 9.4F$. These values become lower and upper bounds for fitting c_1 in $T = c_1 F$ (without the constant c_0). The best fit value for c_1 is ~~that it does not require specifying~~ 5.8 when fitting to the full dataset (Fig. 10), 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 kgm^{-3}/920 kgm^{-3}, this implies that snow, on average, comprises 54% of the measured snow freeboard. Using these values, Eq. 1 simplifies to $T = 5.8F$, as in Fig. 10. 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. 10, 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 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. 4.1.3, Fig. 6), this does not improve the test errors because this is likely too simplistic (even within a ridge, the ratio of snow/ice ~~densities, but instead implicitly accounts for the (potentially spatially-varying) densities with its filters (discussed below) . The ConvNet also gives an output thickness of 4×10^{-2} m (essentially zero)when the input is a zero array, which is physically appropriate~~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 4.2).

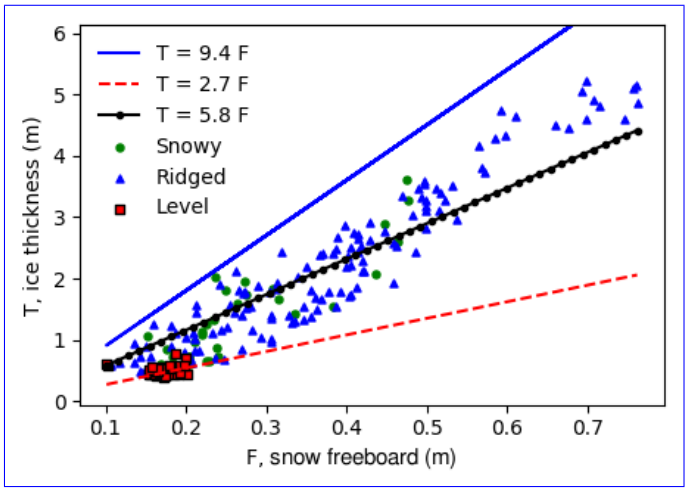


Figure 10. The SIT (T) as a function of measured snow freeboard (F). As expected, all points lie between the two extreme regimes (no ice freeboard and no snow freeboard). The level surfaces mostly have no ice freeboard, as expected, though there is some scatter that suggests a varying component of ice freeboard. The best fit line for all windows from Table 2 is shown in black. Assuming mean snow and ice densities of 300 and 920 kgm^{-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 kgm^{-3} (or effective ice density $<$915 kgm^{-3}.)

Unlike our approach, the fits in Ozsoy-Cicek 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 Bellinghausen 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 Ozsoy-Cicek 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 Ozsoy-Cicek 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.

~~When applying this model~~

5.2 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, ~~see~~ Maksym et al., in prep) ~~with different climatology (different seasonand different region)~~from a different season/region, the MRE is 69%, and the ~~error of predicting the survey mean~~ 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 signficantly 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 ~~range of climatologies~~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 ~~slighly~~ slightly higher test errors. ~~It is not surprising that a smaller window has a higher error, as the isostatic assumption may no longer be valid.~~ 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 for 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 ~~surface elevation~~ snow freeboard measurements naturally gives the mean snow depth via Eq. 1.

6 ~~Discussion~~

5.1 ~~ConvNet metric analysis~~

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. ~~Here, we~~ We can try to analzye 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 (analagous 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. 11. 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 ∼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 ∼ 250 m (Yi et al., 2015), we expect better feature identification, as the entire width of a ridge can be resolved within a filter.

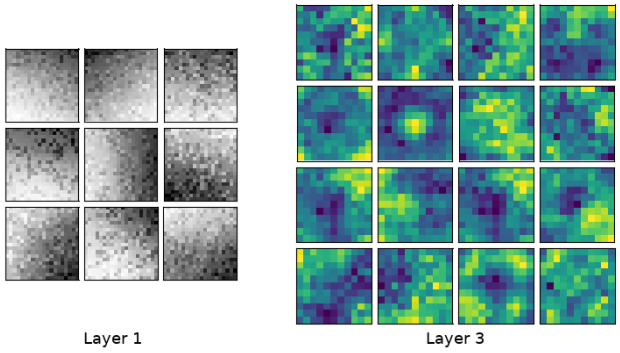


Figure 11. Typical weights learned in the first and last convolutional layers. Weights learned from the third layer are shown using the same colormap as the ~~lidar~~ snow freeboard in Fig. 7 to facilitate comparison. Darker colors indicate lower weights, but the actual values are not important. The 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. 5)

.

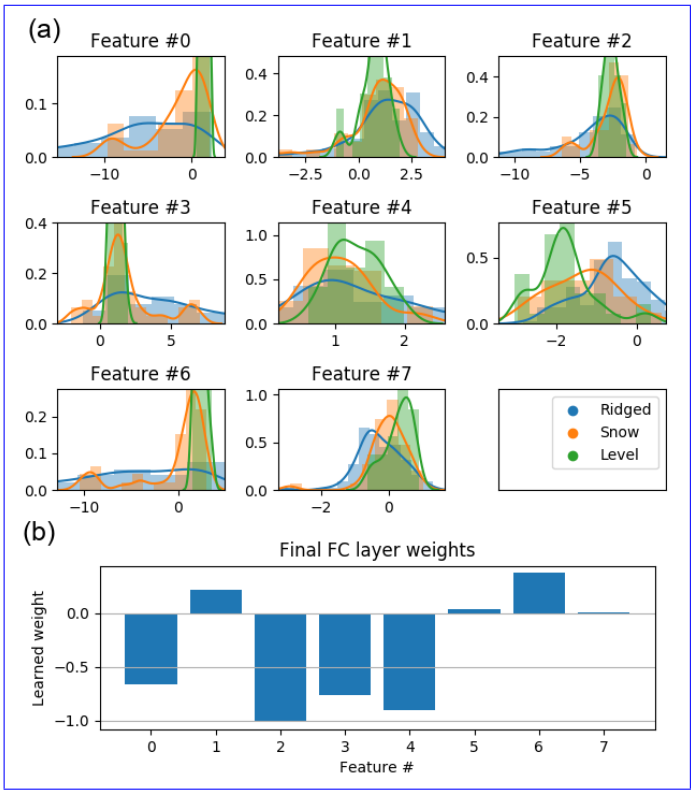


Figure 12. (a) Distribution of the final (8 x 1) layer activations for the level, ridged and snow categories from Fig. 7, and (b) the learned weights for the final fully-connected hidden layer. To generate the final thickness estimate, the activations in (a) are multiplied with the weights in (b), then summed.

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. 12a, grouped by category (level, ridged, snowy). These should correspond to (unspecified) metrics, which are linearly combined with the weights shown in Fig. 12b. 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 ~~elevation~~snow freeboard (Fig. 13d); for snowy surfaces, these three features had a slightly weaker correlation ($0.88 < |R| < 0.96$) to the mean ~~elevation~~snow freeboard; and for level surfaces, features #1 and #5 had a slight correlation ($|R| = 0.67$ and 0.80 respectively) to the mean ~~elevation~~ snow freeboard (Fig. 13a). However, features that correlated to the ridged surface mean ~~elevation~~ snow freeboard did not correlate to the level surface mean ~~elevation~~snow freeboard, and vice versa (Fig. 13b and c). This suggests that the mean ~~elevation~~ snow freeboard for level surfaces is treated differently (e.g. given a different effective density) than other categories.

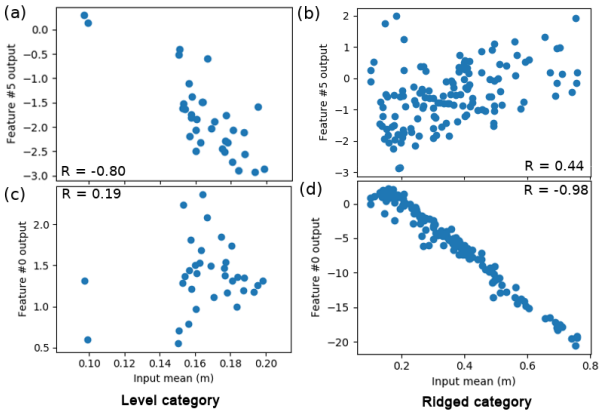


Figure 13. Scatter plot showing correlations between features and real-life metrics. Here, features #0 and #5 correlate strongly to the mean elevations of the level and ridged surfaces respectively, but not the other way around. This suggests that the level and ridged surfaces are treated differently, 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.

For ridged surfaces, in addition to the mean ~~elevation~~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 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 ~~surface elevation~~ 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 ~~sea ice thickness~~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 ~~these categories~~different surface types. However, as suggested by the considerable overlap in the distributions in Fig. 8, these categories may also not be 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. 14). However, the ridged category is quite dispersed, and may even consist of different classes of deformation which should not be grouped all together. Never-

~~(a) Distribution of the final (8 x 1) layer activations for the level, ridged and snow categories from Fig. 7, 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.~~

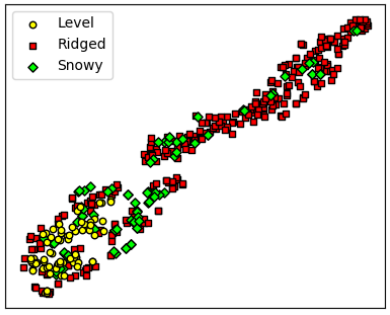


Figure 14. The t-SNE diagram for the encoded input, using the first fully-connected layer (feature vector of size 64) (Maaten and Hinton, 2008). The level and ridged categories are most clearly clustered, although the snowy category may also be a cluster. There is some overlap between the snowy/ridged 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. 7 and 12 are pertinent, but perhaps not the most relevant, for estimating SIT given different surface conditions.

theless, 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 ~~validation error is only slightly lower than the fit error for fitting $T \propto \sigma$ (Fig. **??**b, with MRE 33%), and the~~ test error is worse than the linear model, and has twice the test MRE of our ConvNet with ~~the raw surface elevation~~ snow freeboard (test MRE: 20%). ~~Moreover, a simple statistical model of thickness as a linear function of surface elevation, snow depth and RMS roughness also only does marginally better (MRE 30%). This suggests both that the surface elevation means are important, and also that these means are differently treated for different features, as was speculated in the previous paragraph.~~

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 ~~surface elevation~~snow freeboard (e.g. Ozsoy-Cicek et al., 20 , with the exception of ridged areas, it is not surprising that the demeaned input is not ~~a good~~ as good a predictor of the snow depth. However, when metrics obtained from the demeaned ~~elevation~~ snow freeboard (such as roughness) are combined with the mean ~~elevation~~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 ~~may be predicted given some lidar input, which~~ predictions. This is promising for applications to larger datasets such as OIB or ICESat-2.

~~Interestingly, predicting mean draft thicknesses using demeaned AUV windows gave low errors of 9%/11%/14%, suggesting that ice-surface morphology is a far better predictor of thickness than snow-surface morphology. This is not surprising, given that snow obscures the deformed ice surface. To compensate, the mean elevation becomes much more important.~~

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. 12. 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 ~~surface elevation~~ 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.

~~Feature~~ Features #~~7, which has positively- and negatively-skewed distributions for level/ridged categories respectively, centered on the snowy category distribution, may be accounting for variations in snow density. For example, ridges may have less wind-packed snow due to the shielding effects of the ridge (and hence less dense snow), whereas level surfaces may have wind-packed, denser snow. In contrast, feature~~ 5 and #~~5 has the level/ridged distributions skewed the other way around. Because the weight is positive, but the values are mostly negative, this most strongly reduces the thickness estimate for level surfaces. This may be equivalent to reducing the effective density of the surface due to the presence of snow, which would be reduced the most for level surfaces (that have mostly snow) , whereas the ridged category would have a minor correction (and so the feature values are mostly near 0)~~7 both show some distinguishing of the different surface types, although the weights are so small for these features (Fig. 12b) 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.

6 Summary and 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 Ozsoy-Cicek 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 ~~relative error increasing by factors of 2-4 when estimating local or floe-wide thicknesses~~MRE increasing substantially, likely due to differing snow/ice proportions in the ~~surface elevation~~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 ~~scan-wide~~ survey-wide thicknesses of a ~~new~~ 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. ~~We showed that using demeaned AUV topography has the same low error in predicting mean thickness as using the surface elevation. However, lidar~~ Lidar surveys are much easier to conduct than AUV surveys, and so a ~~more~~ viable method for obtaining more data for future studies is to use a high-resolution lidar scan, combined with coarser measurements of mean ~~sea ice thickness~~ 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.

~~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 sea ice thickness. Our ConvNet has comparable accuracy to a linear fit (16% MRE vs. 20% MRE for fitting PIP4-8), but it has much better generalization to an unseen floe (20% MRE vs. 28% MRE for applying the best linear fit). This linear fit uses both an unphysical constant term, as well as snow depth data that is not needed for the ConvNet. If comparing to a linear fit with no constant and without snow depth data, then the linear fit has a far higher fit error (43% MRE) and far worse generalization (47% MRES) than the ConvNet. The low test error for the ConvNet suggests that surface morphology, as identified by the ConvNet, may be consistent between different floes from the same climatology, and that this morphology may inform estimates of SIT.~~

~~Another~~ Another possible strength of our proposed ConvNet is that it ~~can~~ 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. ~~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), which is considerably lower and at higher resolution than current satellite-based estimates (~50%, or 80 cm, see Kern and Spreen (2015)), and the resulting mean scan-wide SIT also has lower errors (RMS error: 2-3 cm) than empirical methods (16 cm, see Ozsoy-Cicek et al. (2013)).~~

Although our ConvNet would be greatly improved with more training data, it is promising that local ~~sea ice thickness~~ SIT can be accurately predicted given only ~~surface elevation~~ snow freeboard measurements. More extensive lidar/AUV/snow measurements from different regions/seasons would improve the ConvNet generalization~~, but because high-resolution ice thickness and snow depth are not needed, other, simpler-to-obtain data sets (e.g. coincident scanning lidar and EM-induction ice thickness measurements) can also be used with this technique.~~.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 ~8 m resolution are also resolved by OIB lidar data (resolution 1-3 m). ~~Using~~

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 ~~. With more data, the low errors of deep learning-based methods may yield high-resolution, low-error SIT estimates that may be able to verify modest interannual variability~~or Operation IceBridge.

Competing interests. The authors declare they have no competing interests.

Author contributions. JM and TM conceived of the research idea~~and~~, and JM, TM and BW collected field data. The manuscript was written by JM and edited by TM. HS wrote the code for processing the AUV data and JM wrote the code for analyzing the data.

Data availability. The ~~entire PIPERS dataset is expected to be uploaded to a publicly-accessible portal in late 2019. However, all the~~ PIPERS layer cake data used here ~~will be uploaded upon publication of this paper.~~ is available at: Jeffrey Mei, M. and T. Maksym (2019) "Sea Ice Layer Cakes, PIPERS 2017" U.S. Antarctic Program (USAP) Data Center. doi:10.15784/601207.10.15784/601192

Acknowledgements. This work was supported by U.S. National Science Foundation grants ANT-1341606, ANT-1142075~~and the U.S. Office of Naval Research N00014-13-1-0434~~, ANT-1341717 and NASA grant NNX15AC69G. The authors would like to thank ~~Blake Weissling for providing the lidar data, and~~ Jeff Anderson for ~~technical support in the AUV missions~~overseeing the AUV surveys. Guy Williams and Alek Razdan were instrumental for collecting the AUV data. The crew onboard the RV N. B. Palmer were also essential to the success of the PIPERS expedition.

Appendix A: ConvNet Details

For a comprehensive introduction to deep learning, the reader is directed to Shalev-Shwartz and Ben-David (2014). Here we will give the details of our ConvNet and explain the importance of chosen parameters.

Convolutional Neural Networks, commonly known as ConvNets, are a class of deep neural networks that convolve filters (matrices that contain weighting coefficients, or weights) through the input array. The input array is typically an image, and the learned filters typically correspond to basic edge detections in initial layers, and more complex features in later layers (e.g. Krizhevsky et al., 2012). Here, we use the lidar elevation scan as an input, due to its similarity to a grayscale image.

Like other deep learning methods, ConvNets 'learn' by updating their weights. This is done through comparing the output of the prediction with the true output, using the derivative of a loss function (here, mean squared error) propagated through the layers in reverse (backpropagation). The weight update rule, in its most basic form, is $w_{i+1} = w_i + \eta \frac{\partial E}{\partial w_i}$, for some weight w, loss function E and learning rate η. The value of η is important to ensure convergence: too high, and the filters may not converge (and may even diverge); too low, and the filters may take too long to converge. In order to introduce nonlinearities in the network, a nonlinear activation function is used at each layer. Typically, this is done with a Rectified Linear Unit (ReLU), which zeros out all negative activations. We chose a scaled exponential linear unit (SELU), which has been found to improve convergence (Klambauer et al., 2017), as ReLUs sometimes lead to dead weights when dealing with many negative values. As convolutions by default shift by 1 pixel at a time, this leads to considerable overlap and large output sizes at each layer. To combat this, the filters can shift by a different number; this is called the stride.

ConvNets are normally used in image classification problems due to their ability to discern features. The output would be a probability vector ~~assigining~~ assigning likelihood of different classes, with the highest one being the prediction. ConvNets can also be applied to regression problems (e.g. Levi and Hassner, 2015) by simply changing the output to be one number. Here, we make the output 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. Similarly, the lidar inputs were scaled by 2.0 to keep them between 0-1. The values are unscaled during model evaluation. ConvNet inputs, when dealing with image classification, are often standardized to have a mean of 0 and a variance of 1, but this was not done here as we want to use the mean and variance (roughness) of the elevation to predict the mean ice thickness.

~~The proposed architecture is shown in Fig. 5. We use multiple convolutional layers to try to capture morphological features, along with fully connected layers at the end to combine the learned features.~~ We tried networks with 2, 3 and 4 convolutional

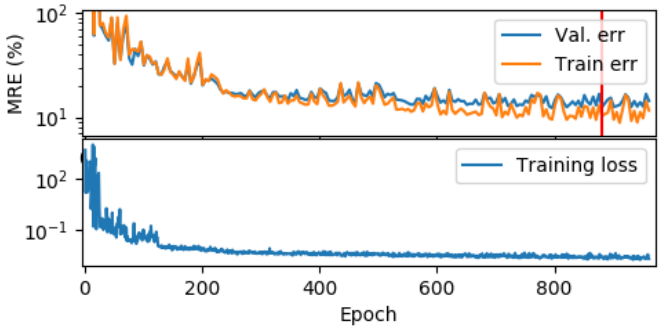


Figure A.1. Training errors, validation errors and training losses shown on a logarithmic scale. Although the training loss continues to slowly drop after the epoch with the lowest validation error (red line, at epoch 881), validation error stays relatively flat, suggesting that the ConvNet is overfitting after this epoch. The gradual decrease in MRE is less smooth than the training loss because the loss function is mean squared error, whereas the MRE is proportional to the mean absolute error.

layers and 1 or 2 fully connected layers with a variety of filter sizes and found the one shown in Fig. 5, 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, ~~following Section ??~~as discussed in Section 3.2. 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 8 hours.

~~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. The optimzer 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}.~~

The input windows were randomly flipped and rotated in integer multiples of 90^o 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, as shown in Fig. A.1.

References

Akaike, H.: A new look at the statistical model identification, in: Selected Papers of Hirotugu Akaike, pp. 215–222, Springer, doi:10.1007/978-1-4612-1694-0_16, 1974.

Baldi, P.: Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37–49, 2012.

Behrendt, A., Dierking, W., Fahrbach, E., and Witte, H.: Sea ice draft in the Weddell Sea, measured by upward looking sonars, Earth Syst. Sci. Data, 5, 209–226, doi:10.5194/essd-5-209-2013, 2013.

Brock, J. C., Wright, C. W., Clayton, T. D., and Nayegandhi, A.: LIDAR optical rugosity of coral reefs in Biscayne National Park, Florida, Coral Reefs, 23, 48–59, doi:10.1007/s00338-003-0365-7, 2004.

Dierking, W.: Laser profiling of the ice surface topography during the Winter Weddell Gyre Study 1992, Journal of Geophysical Research: Oceans, 100, 4807–4820, doi:10.1029/94jc01938, 1995.

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

Ekeberg, O.-C., Høyland, K., and Hansen, E.: 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, doi:10.1016/j.coldregions.2014.10.003, 2015.

Fons, S. W. and Kurtz, N. T.: Retrieval of snow freeboard of Antarctic sea ice using waveform fitting of CryoSat-2 returns, The Cryosphere, 13, 861–878, doi:10.5194/tc-13-861-2019, 2019.

Goebell, S.: 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, doi:10.1029/2009jc006055, 2011.

Haas, C.: 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, 1–16, doi:10.1016/s0165-232x(97)00019-0, 1998.

Haas, C., Liu, Q., and Martin, T.: 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, 3111–3123, doi:10.1080/014311699211642, 1999.

Haas, C., Lobach, J., Hendricks, S., Rabenstein, L., and Pfaffling, A.: Helicopter-borne measurements of sea ice thickness, using a small and lightweight, digital EM system, Journal of Applied Geophysics, 67, 234–241, doi:10.1016/j.jappgeo.2008.05.005, 2009.

Harms, S., Fahrbach, E., and Strass, V. H.: Sea ice transports in the Weddell Sea, Journal of Geophysical Research: Oceans, 106, 9057–9073, doi:10.1029/1999jc000027, 2001.

Haumann, F. A., Gruber, N., Münnich, M., Frenger, I., and Kern, S.: Sea-ice transport driving Southern Ocean salinity and its recent trends, Nature, 537, 89, doi:10.1038/nature19101, 2016.

Holland, M. M., Bitz, C. M., Hunke, E. C., Lipscomb, W. H., and Schramm, J. L.: Influence of the sea ice thickness distribution on polar climate in CCSM3, Journal of Climate, 19, 2398–2414, doi:10.1175/jcli3751.1, 2006.

Holland, P. R., Bruneau, N., Enright, C., Losch, M., Kurtz, N. T., and Kwok, R.: Modeled trends in Antarctic sea ice thickness, Journal of Climate, 27, 3784–3801, doi:10.1175/jcli-d-13-00301.1, 2014.

Hutchings, J. K., Heil, P., Lecomte, O., Stevens, R., Steer, A., and Lieser, J. L.: Comparing methods of measuring sea-ice density in the East Antarctic, Annals of Glaciology, 56, 77–82, doi:10.3189/2015aog69a814, 2015.

Kaleschke, L., Girard-Ardhuin, F., Spreen, G., Beitsch, A., and Kern, S.: 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.

Kern, S. and Spreen, G.: Uncertainties in Antarctic sea-ice thickness retrieval from ICESat, Annals of Glaciology, 56, 107–119, doi:10.3189/2015aog69a736, 2015.

Kern, S., Ozsoy-Çiçek, B., and Worby, A.: Antarctic sea-ice thickness retrieval from ICESat: Inter-comparison of different approaches, Remote Sensing, 8, 538, doi:10.3390/rs8070538, 2016.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, https://arxiv.org/abs/1412.6980, 2014.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S.: Self-normalizing neural networks, in: Advances in neural information processing systems, pp. 971–980, http://papers.nips.cc/paper/6698-self-normalizing-neural-networks, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105, 2012.

Kurtz, N.: 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.

Kurtz, N. and Markus, T.: Satellite observations of Antarctic sea ice thickness and volume, Journal of Geophysical Research: Oceans, 117, doi:10.1029/2012jc008141, 2012.

Kurtz, N. T., Farrell, S. L., Koenig, L. S., Studinger, M., Harbeck, J. P., et al.: A sea-ice lead detection algorithm for use with high-resolution airborne visible imagery, IEEE Transactions on Geoscience and Remote Sensing, 51, 38–56, doi:10.1109/tgrs.2012.2202666, 2012.

Kwok, R. and Cunningham, G.: Variability of Arctic sea ice thickness and volume from CryoSat-2, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 20140 157, doi:10.1098/rsta.2014.0157, 2015.

Kwok, R. and Maksym, T.: 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, 4141–4167, doi:10.1002/2014jc009943, 2014.

Kwok, R. and Rothrock, D.: Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008, Geophysical Research Letters, 36, doi:10.1029/2009gl039035, 2009.

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

Leppäranta, M. and Hakala, R.: The structure and strength of first-year ice ridges in the Baltic Sea, Cold Regions Science and Technology, 20, 295–311, doi:10.1016/0165-232x(92)90036-t, 1992.

Levi, G. and Hassner, T.: Age and gender classification using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 34–42, 2015.

Li, H., Xie, H., Kern, S., Wan, W., Ozsoy, B., Ackley, S., and Hong, Y.: 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, doi:10.1016/j.rse.2018.09.031, 2018.

Li, S. and Chan, A. B.: 3d human pose estimation from monocular images with deep convolutional neural network, in: Asian Conference on Computer Vision, pp. 332–347, Springer, 2014.

Maaten, L. v. d. and Hinton, G.: Visualizing data using t-SNE, Journal of machine learning research, 9, 2579–2605, 2008.

Maksym, T. and Markus, T.: Antarctic sea ice thickness and snow-to-ice conversion from atmospheric reanalysis and passive microwave snow depth, Journal of Geophysical Research: Oceans, 113, doi:10.1029/2006jc004085, 2008.

Markus, T. and Cavalieri, D. J.: Snow depth distribution over sea ice in the Southern Ocean from satellite passive microwave data, Antarctic sea ice: physical processes, interactions and variability, pp. 19–39, doi:10.1029/ar074p0019, 1998.

Markus, T., Massom, R., Worby, A., Lytle, V., Kurtz, N., and Maksym, T.: Freeboard, snow depth and sea-ice roughness in East Antarctica from in situ and multiple satellite data, Annals of Glaciology, 52, 242–248, doi:10.3189/172756411795931570, 2011.

Markus, T., Neumann, T., Martino, A., Abdalati, W., Brunt, K., Csatho, B., Farrell, S., Fricker, H., Gardner, A., Harding, D., et al.: The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): science requirements, concept, and implementation, Remote sensing of environment, 190, 260–273, doi:10.1016/j.rse.2016.12.029, 2017.

Massom, R. A., Eicken, H., Hass, C., Jeffries, M. O., Drinkwater, M. R., Sturm, M., Worby, A. P., Wu, X., Lytle, V. I., Ushio, S., et al.: Snow on Antarctic sea ice, Reviews of Geophysics, 39, 413–445, doi:10.1029/2000rg000085, 2001.

Massonnet, F., Mathiot, P., Fichefet, T., Goosse, H., Beatty, C. K., Vancoppenolle, M., and Lavergne, T.: A model reconstruction of the Antarctic sea ice thickness and volume changes over 1980–2008 using data assimilation, Ocean Modelling, 64, 67–75, doi:10.1016/j.ocemod.2013.01.003, 2013.

Melling, H. and Riedel, D. A.: 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, 11 975–11 991, doi:10.1029/96jc00284, 1996.

Ozsoy-Cicek, B., Kern, S., Ackley, S. F., Xie, H., and Tekeli, A. E.: 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, 1092–1111, doi:10.1016/j.dsr2.2010.10.031, 2011.

Ozsoy-Cicek, B., Ackley, S., Xie, H., Yi, D., and Zwally, J.: Sea ice thickness retrieval algorithms based on in situ surface elevation and thickness values for application to altimetry, Journal of Geophysical Research: Oceans, 118, 3807–3822, doi:10.1002/jgrc.20252, 2013.

Parkinson, C. and Cavalieri, D.: Antarctic sea ice variability and trends, 1979–2010, The Cryosphere, 6, 871–880, doi:10.5194/tc-6-871-2012, 2012.

Petty, A. A., Tsamados, M. C., Kurtz, N. T., Farrell, S. L., Harbeck, J. P., Feltham, D. L., and Richter-Menge, J. A.: Characterizing Arctic sea ice topography using high-resolution IceBridge data, The Cryosphere, 10, 1161, doi:10.5194/tc-10-1161-2016, 2016.

Rothrock, D., Percival, D., and Wensnahan, M.: 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, doi:10.1029/2007jc004252, 2008.

Shalev-Shwartz, S. and Ben-David, S.: Understanding machine learning: From theory to algorithms, Cambridge university press, doi:10.1017/cbo9781107298019, 2014.

Shu, Q., Song, Z., and Qiao, F.: Assessment of sea ice simulations in the CMIP5 models, The Cryosphere, 9, 399–409, doi:10.5194/tc-9-399-2015, 2015.

Sibson, R.: A brief description of natural neighbour interpolation, Interpreting multivariate data, 1981.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research, 15, 1929–1958, 2014.

Stammerjohn, S., Martinson, D., Smith, R., Yuan, X., and Rind, D.: 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, doi:10.1029/2007jc004269, 2008.

Stroeve, J. C., Markus, T., Maslanik, J. A., Cavalieri, D. J., Gasiewski, A. J., Heinrichs, J. F., Holmgren, J., Perovich, D. K., and Sturm, M.: Impact of surface roughness on AMSR-E sea ice products, IEEE Transactions on Geoscience and Remote Sensing, 44, 3103–3117, doi:10.1109/TGRS.2006.880619, 2006.

Strub-Klein, L. and Sudom, D.: A comprehensive analysis of the morphology of first-year sea ice ridges, Cold Regions Science and Technology, 82, 94–109, doi:10.1016/j.coldregions.2012.05.014, 2012.

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

Sturm, M., Morris, K., and Massom, R.: The winter snow cover of the West Antarctic pack ice: its spatial and temporal variability, Antarctic sea ice: physical processes, interactions and variability, pp. 1–18, doi:10.1029/ar074p0001, 1998.

Thomson, J., Ackley, S., Girard-Ardhuin, F., Ardhuin, F., Babanin, A., Boutin, G., Brozena, J., Cheng, S., Collins, C., Doble, M., et al.: Overview of the Arctic sea state and boundary layer physics program, Journal of Geophysical Research: Oceans, 123, 8674–8687, doi:10.1002/2018jc013766, 2018.

Timco, G. and Burden, R.: An analysis of the shapes of sea ice ridges, Cold regions science and technology, 25, 65–77, doi:10.1016/s0165-232x(96)00017-1, 1997.

Timco, G. and Frederking, R.: A review of sea ice density, Cold regions science and technology, 24, 1–6, doi:10.1016/0165-232x(95)00007-x, 1996.

Timco, G. and Weeks, W.: A review of the engineering properties of sea ice, Cold regions science and technology, 60, 107–129, doi:10.1016/j.coldregions.2009.10.003, 2010.

Timco, G. W. and Sayed, M.: Model tests of the ridge-building process in ice, 1986.

Tin, T. and Jeffries, M.: 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.

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

Tin, T. and Jeffries, M. O.: Morphology of deformed first-year sea ice features in the Southern Ocean, Cold regions science and technology, 36, 141–163, doi:10.1016/s0165-232x(03)00008-9, 2003.

Tucker III, W. and Govoni, J.: Morphological investigations of first-year sea ice pressure ridge sails, Cold Regions Science and Technology, 5, 1–12, doi:10.1016/0165-232x(81)90036-7, 1981.

Tucker III, W. B., Sodhi, D. S., and Govoni, J. W.: Structure of first-year pressure ridge sails in the Prudhoe Bay region, in: The Alaskan Beaufort Sea, pp. 115–135, Elsevier, doi:10.1016/b978-0-12-079030-2.50012-5, 1984.

Turner, J., Bracegirdle, T. J., Phillips, T., Marshall, G. J., and Hosking, J. S.: An initial assessment of Antarctic sea ice extent in the CMIP5 models, Journal of Climate, 26, 1473–1484, doi:10.1175/jcli-d-12-00068.1, 2013.

Urabe, N. and Inoue, M.: Mechanical properties of Antarctic sea ice, Journal of Offshore Mechanics and Arctic Engineering, 110, 403–408, doi:10.1115/1.3257079, 1988.

Weissling, B. and Ackley, S.: Antarctic sea-ice altimetry: scale and resolution effects on derived ice thickness distribution, Annals of Glaciology, 52, 225–232, doi:10.3189/172756411795931679, 2011.

Willatt, R. C., Giles, K. A., Laxon, S. W., Stone-Drake, L., and Worby, A. P.: Field investigations of Ku-band radar penetration into snow cover on Antarctic sea ice, IEEE Transactions on Geoscience and remote sensing, 48, 365–372, doi:10.1109/tgrs.2009.2028237, 2009.

Williams, G., Maksym, T., Wilkinson, J., Kunz, C., Murphy, C., Kimball, P., and Singh, H.: Thick and deformed Antarctic sea ice mapped with autonomous underwater vehicles, Nature Geoscience, 8, 61–67, doi:10.1038/ngeo2299, 2015.

Williams, G. D., Maksym, T., Kunz, C., Kimball, P., Singh, H., Wilkinson, J., Lachlan-Cope, T., Trujillo, E., Steer, A., Massom, R., et al.: Beyond point measurements: Sea ice floes characterized in 3-D, Eos, Transactions American Geophysical Union, 94, 69–70, doi:10.1002/2013eo070002, 2013.

Wingham, D., Francis, C., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S., Mallow, U., Mavrocordatos, C., et al.: CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields, Advances in Space Research, 37, 841–871, doi:10.1016/j.asr.2005.07.027, 2006.

Worby, A., Jeffries, M., Weeks, W., Morris, K., and Jana, R.: 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, 28 441–28 455, doi:10.1029/96jc02737, 1996.

Worby, A., Bush, G., and Allison, I.: Seasonal development of the sea-ice thickness distribution in East Antarctica: Measurements from upward-looking sonar, Annals of glaciology, 33, 177–180, doi:10.3189/172756401781818167, 2001.

Worby, A. P., Geiger, C. A., Paget, M. J., Van Woert, M. L., Ackley, S. F., and DeLiberty, T. L.: Thickness distribution of Antarctic sea ice, Journal of Geophysical Research: Oceans, 113, doi:10.1029/2007jc004254, 2008.

Worby, A. P., Steer, A., Lieser, J. L., Heil, P., Yi, D., Markus, T., Allison, I., Massom, R. A., Galin, N., and Zwally, J.: 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, 1125–1136, doi:10.1016/j.dsr2.2010.12.001, 2011.

Xie, H., Ackley, S., Yi, D., Zwally, H., Wagner, P., Weissling, B., Lewis, M., and Ye, K.: Sea-ice thickness distribution of the Bellingshausen Sea from surface measurements and ICESat altimetry, Deep Sea Research Part II: Topical Studies in Oceanography, 58, 1039–1051, doi:10.1016/j.dsr2.2010.10.038, 2011.

Xie, H., Tekeli, A. E., Ackley, S. F., Yi, D., and Zwally, H. J.: Sea ice thickness estimations from ICESat Altimetry over the Bellingshausen and Amundsen Seas, 2003–2009, Journal of Geophysical Research: Oceans, 118, 2438–2453, doi:10.1002/jgrc.20179, 2013.

Yi, D., Zwally, H. J., and Robbins, J. W.: 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, 43–51, doi:10.3189/172756411795931480, 2011.

Yi, D., Harbeck, J. P., Manizade, S. S., Kurtz, N. T., Studinger, M., and Hofton, M.: 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, 1403–1410, doi:10.1109/tgrs.2014.2339737, 2015.

Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, in: European conference on computer vision, pp. 818–833, Springer, doi:10.1007/978-3-319-10590-1_53, 2014.

Zwally, H. J., Yi, D., Kwok, R., and Zhao, Y.: ICESat measurements of sea ice freeboard and estimates of sea ice thickness in the Weddell Sea, Journal of Geophysical Research: Oceans, 113, doi:10.1029/2007jc004284, 2008.