SUPPLEMENTARY MATERIAL: SEMI-AUTOMATED OPEN WATER ICEBERG DETECTION FROM LANDSAT APPLIED TO DISKO BAY, WEST GREENLAND

Jessica SCHEICK, Ellyn M. ENDERLIN, Gordon HAMILTON

Power law, lognormal, and Weibull distributions have previously been used to describe Greenlandic and Antarctic iceberg size distributions (e.g. Savage and others, 2000; Tournadre and others, 2012; Enderlin and others, 2016; Kirkham and others, 2017; Sulak and others, 2017). The application of different statistical models to describe iceberg size distributions suggests that the physics of iceberg decay plays an important role in determining the size distribution of ice pieces (Savage, 2001), particularly when time and distance from the parent glacier and/or parent iceberg are considered. For example, a recent analysis by Kirkham and others (2017) suggests that at the time of calving icebergs follow a power law distribution which transitions to a lognormal distribution with distance from the calving location as different and increasingly fewer physical processes dominate the decay process.

The shape of any distribution function describing iceberg sizes (e.g. area, length, volume/mass) can be broadly described as highly skewed or heavy-tailed. As such, the data becomes easier to interpret when viewed in log-log space (Fig. S1). The selection of bin sizes to describe frequency data in log-log space inevitably plays a role in our interpretation of the data. Specifically, a probability density function (PDF) with linearly spaced bins (Fig. S1a) clearly displays an inflection point in the data. The location of the inflection point, here at iceberg surface areas of ~10000 m m^2, depends completely on the choice of bin size and has no physically-based interpretation. Thus, using a PDF with linearly spaced bins to describe iceberg size distributions makes it difficult to fit size distributions to the entire dataset unless a maximum x value is defined (Alstott and others, 2014), resulting in the unnecessary exclusion of a portion of the dataset. A PDF with logarithmically-spaced bins (Fig. S1b) effectively includes the larger icebergs in the distribution and smooths the inflection point, but the shape and slope of the curve are still influenced by the number of bins used. Alternatively, a complimentary cumulative density function (CCDF, Fig. S1c) provides a means of objectively fitting a size distribution without the need for determining ideal bin sizes (Alstott and others, 2014). This approach is commonly taken in available computational libraries designed for testing power law and other similar heavy-tailed distributions and is the method used here.

The large number of methods employed in the literature for fitting iceberg size distributions suggests the non-trivial nature of fitting empirical distributions to natural phenomenon. Unfortunately, it is all too common that the preferred model used to fit size distributions is chosen based primarily on a qualitative inspection of the data rather than robust statistical methods (Clauset and others, 2009). In the case of supposed power law distributions, the fitted parameters are often computed using a least squares fit to the data in log-log space, alternative distributions are not rigorously evaluated, and the statistical validity of the model for describing the dataset is not tested (Clauset and others, 2009). However, the limitations imposed by statistical rigor have the potential to effectively eliminate large portions of a measured dataset, in turn making it difficult to characterize a natural system and suggesting that a compromise between pure and applied mathematics is necessary to describe the stochasticity of natural phenomena in a consistent framework.

As a starting point to determine the best fit models to describe our data, we used the poweRlaw package (Gillespie, 2015) for the open-source statistical software R (R Core Team, 2018). The package contains easy-to-implement methods for testing power law, lognormal, and exponential fits of the form:

$$powerlaw : f(x) = x^{-\alpha}$$

$$lognormal : f(x) = \frac{1}{x} exp[-\frac{(ln(x) - \mu)^2}{2\beta^2}]$$

$$exponential : f(x) = e^{-\lambda x}$$

where α, μ, β, and λ are their respective fit parameters. It includes methods for determination of the best minimum x (x_{min}) value based on the Kolmogorov-Smirnov (KS) fit statistic, measures of model fit and estimates of parameter uncertainty using bootstrapping, and model intercomparisons using log likelihood-ratio testing (Vuong's method) to compare alternative distributions (Clauset and others, 2009). The process of fitting and testing a statistical model using the package is outlined in detail in Clauset and others (2009) and in the package's documentation. An iceberg size distribution from 31 August 2013 with potential model fits and relevant statistical parameters is shown in Figure S1 and Table S1. In this example case and for one other case tested (not shown), the exponential curve showed a visually poor fit to the data and exhibited very high x_{min} values with associated poor goodness of fit values for the x_{min} estimation. When compared to other models, the non-exponential models had a statistically significant better fit. As a result, the exponential model was not considered further as a potential distribution for the iceberg size distribution data.

A key step that drives the rest of the analysis for fitting a model distribution to any dataset begins with the determination of x_{min} values for each model. x_{min} is determined using the KS statistic as detailed in

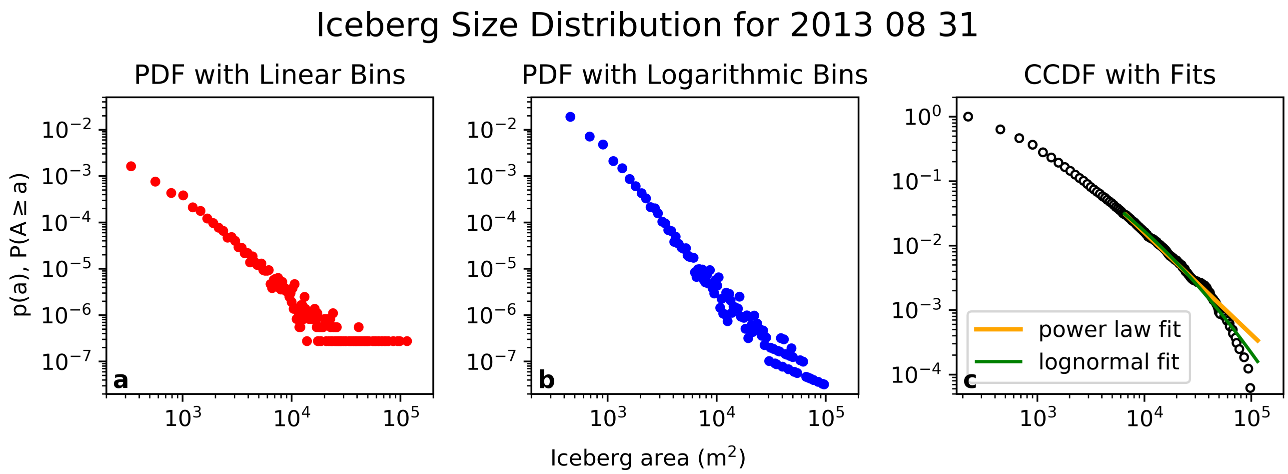

Fig. 1. Size distribution of icebergs delineated by the automated algorithm for the Landsat scenes collected 2013 08 31. a (b) shows the iceberg area probability density function (PDF) in log-log space with linear (log) bins. c) shows the complimentary cumulative distribution function (CCDF) for the dataset with modeled power law (yellow dashed) and lognormal fits (solid green). $n = 16145$, $n_{tail} = 492$.

Clauset and others (2009) and identifies the starting point beyond which the data can most accurately be described by a given distribution; over- or underestimation of this statistic quickly influences the value of fit parameters, with a too-high value being preferred to a too-low value (Clauset and others, 2009). Although the minimum iceberg size theoretically detectible in Landsat imagery would be one pixel (225 m^2), the x_{min} values recommended by the software for the example size distribution are an order of magnitude larger, though they are similar for both the power law and lognormal models. In order to compare two distributions, they must have equivalent x_{min} values. Thus, we compared the power law and lognormal models using both x_{min} values, and in both cases the p value was >0.1, suggesting we cannot reject the null hypothesis that one model is a better fit to the data with the sign of the returned ratio R indicating which model is better. For future interpretations wherein R is statistically significant, in our implementation of the package negative R values indicate the lognormal model is a better fit. A visual inspection of the power law and lognormal curves

Table 1. Iceberg size distribution fit parameters from poweRlaw for one Landsat scene (2013 08 31).

Powerlaw		**Lognormal**			**Comparison (power law x_{min})**		**Comparison (lognormal x_{min})**	
x_{min}	α	x_{min}	μ	β	R	p	R	p
6750	2.58	6525	5.59	1.67	-1.48	0.14	-1.64	0.101

fitted to the data provides qualitative confirmation that the distribution could readily be described by either model. Acknowledging that neither model necessarily provides a better fit to the data but in pursuit of a quantitative description of the shape of the iceberg size distribution curve, we ran a bootstrapping procedure with 1000 iterations using the power law model to determine the statistical significance of a power law fit and the uncertainty on the parameter estimate. The results of this bootstrapping suggest that a power law fit to the data is statistically significant ($p=0.181>0.1$). The fitted parameter (α), which is the slope of the power law fit, has a value of 2.58 ± 0.10. This value is notably larger than previously estimated values (Enderlin and others, 2016; Sulak and others, 2017) and the theoretically expected value of 1.5 (Åström and others, 2014), possibly suggesting that previous investigations have underestimated the fit parameter and/or the theoretically derived value does not apply to Disko Bay given the distance of the icebergs from the calving front. The portion of the dataset fitted by the statistical model contains enough values (n ~500) to suggest the obtained parameter estimates are reliable, though the number of observations in the data tail is small enough (<1000) that the algorithm's selection of x_{min} may be compromised in this case (Clauset and others, 2009).

The above analysis suffers from several important limitations. First, only three models are considered; these models were chosen based on their previous use in the literature, qualitative inspection of the data, and ease of comparison. However, alternative models not tested in the poweRlaw implementation might provide a superior fit to the data and/or be able to explain a larger portion of the dataset. Second, the x_{min} values calculated by the algorithm eliminate an overwhelmingly large proportion of the iceberg areas measured (often >50% of the data). This has the important consequences of reducing the likelihood of statistically significant outcomes that generally arise from a large dataset and failing to characterize the full range of data, thereby posing a challenge for assessing changes in characteristic iceberg size distributions. Third, where lognormal was the preferred model, the software does not enable computation of the statistical significance of the model fit. Thus, it is impossible to tell whether or not the lognormal fit is statistically valid, even if the parameter uncertainty is small. Together, these limitations suggest that perhaps the power law and lognormal models are too simplistic to represent the proportions of icebergs present across the full range of iceberg sizes. Alternative models such as the large number of rare events model may provide a suitable distribution, especially to capture the tail portions of the size distribution curve, which includes the comparatively rare but largest icebergs present in many regions. An alternative approach to fitting one model to the data would be to apply breakpoint regression or a related statistical technique that iteratively tests different models on portions of the data to determine a series of breakpoints within the dataset and fit the most appropriate model to each section of the data. Determining a more robust way to statistically model iceberg size distributions represents an important avenue for future work but is beyond the scope of this investigation.

In an attempt to address some of the limitations discussed, the iceberg size distributions were also compared using the powerlaw library for Python (Alstott and others, 2014), which is designed to implement the same statistical solutions as the R version but allows the comparison of additional distributions. The companion paper by Alstott and others (2014) also provides a more nuanced discussion for using the package to fit measured size distributions. A comparison of outputs from the poweRlaw and powerlaw packages for the 2013 08 31 iceberg size distributions confirms the dependence of the fitted parameters on the chosen x_{min} value but otherwise produces similar results. An inspection of the KS values for each possible x_{min} value shows that the absolute minimum chosen by the software is very similar to several other local minima and thus choosing a smaller x_{min} value that includes more of the data is not unreasonable (Alstott and others, 2014). The use of a smaller x_{min} value also does not change the conclusion that neither a power law nor a lognormal distribution provides a better fit to the data. Further, the use of the powerlaw library enables confirmation that neither a stretched exponential (i.e. Weibull distribution) nor an exponential provide a better fit to the data.

To characterize our data, we fit power law size distributions to all datasets using an x_{min} value of 1800. We acknowledge that this likely influences the fit parameter values but argue that it effectively limits data loss associated with high x_{min} values while minimizing the influence of large fluctuations in the smallest size fractions of icebergs. We choose this approach because, although in some cases a lognormal distribution might be more appropriate for describing the data, in general this relationship is tenuous and this approach provides consistency that enables comparison across our entire dataset as well as with previously computed values.

REFERENCES

Alstott J, Bullmore E and Plenz D (2014) powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE*, **9**(1) (doi: 10.1371/journal.pone.0085777)

Aström JA and 10 others (2014) Termini of calving glaciers as self-organized critical systems. *Nat. Geosci.*, **7**, 874–878 (doi: 10.1038/ngeo2290)

Clauset A, Shalizi CR and Newman MEJ (2009) Power-law dstributions in empirical data. 1–43 (doi: 10.1109/ICPC.2008.18)

Enderlin EM, Hamilton GS, Straneo F and Sutherland DA (2016) Iceberg meltwater fluxes dominate the freshwater budget in Greenland's iceberg-congested glacial fjords. *Geophys. Res. Lett.*, **43**(11), 287–294 (doi: 10.1002/(ISSN)1944-8007)

Gillespie C (2015) Fitting heavy tailed distributions: the poweRlaw package. *Jo. Stat. Softw.*, **64**(2), 257–266

Kirkham JD and 8 others (2017) Drift-dependent changes in iceberg size-frequency distributions. *Nature*, **7**(15991), 1–10

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Savage S (2001) *Aspects of Iceberg Deterioration and Drift*, chapter 12, 279–318. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-45670-4 (doi: 10.1007/3-540-45670-8)

Savage SB, Crocker GB, Sayed M and Carrieres T (2000) Size distributions of small ice pieces calved from icebergs. *Cold Reg. Sci. Technol.*, **31**(2), 163–172

Sulak DJ, Sutherland DA, Enderlin EM, Stearns LA and Hamilton GS (2017) Iceberg properties and distributions in three Greenlandic fjords using satellite imagery. *Ann. Glaciol.*, 1–15 (doi: 10.1017/aog.2017.5)

Tournadre J, Girard-Ardhuin F and Legrésy B (2012) Antarctic icebergs distributions, 2002–2010. *J. Geophys. Res.*, **117**(C5), C05004 (doi: 10.1029/2011JC007441)