Quantifying uncertainty in Pareto estimates of global lake area

This is an Accepted Manuscript of an article published in Limnology & Oceanography: Methods
DOI: 10.1002/LOM3.10536

Jemma Stachelek
Earth System Observations
Los Alamos National Laboratory
Los Alamos, NM, USA, 87544
jsta@lanl.gov

Abstract

Size is a critical factor determining the rate and occurrence of specific lake processes such as carbon sequestration and greenhouse gas emissions and emerging evidence suggests that small lakes in particular have particularly large CO^2 flux rates. Because we do not have a complete census of all lakes, upscaling estimates of such processes to small lakes at broad spatial scales requires the use of lake size-abundance distributions rather than empirical measurements of area. Existing lake census efforts are incomplete such that as lakes become smaller, they are more likely to be omitted either because they are too small to be resolved from remote sensing products or because of limited ground surveying effort (i.e. "censoring" of small lakes relative to large lakes). The present study explores one potential shortcoming of prior approaches estimating global lake area using lake size-abundance distributions. Namely, that these prior approaches rely on frequentist curve fitting techniques combined with an ad-hoc cutoff determination strategy (visual inspection to determine a likely censoring point). This yields an over-exact lake area estimate that is typically reported with no uncertainty bounds. I show how these shortcomings can be addressed with a Bayesian model that produces larger estimates of lake area uncertainty relative to the typical approach. When used as part of a sensitivity analysis, such an approach has the potential to enable more robust intercomparisons among studies of aquatic processes upscaling.

1 Introduction

Size is a critical factor determining the contribution of lakes to global biogeochemical cycling. As such, existing evidence suggests that lakes are key components of overall carbon burial, sequestration, and greenhouse gas emissions with small lakes having particularly high CO^2 flux rates (DelSontro et al., 2018; Keller et al., 2021). When upscaled to broad spatial extents, the high flux rates of these small lakes translate to a disproportionately large influence on overall CO^2 emissions (Pi et al., 2022). The calculation of such fluxes requires two terms 1) an areal flux rate and 2) the distribution of total lake area. As with all quantitative estimates at broad spatial scales, uncertainty estimates are of great interest both because they allow for intercomparisons among similar studies and because they allow for comparisons between ecosystems.

The techniques typically used to determine uncertainty in areal flux rates differ greatly from those used to determine uncertainty in total lake area distributions. In the former case, uncertainties can typically be calculated in straightforward manner from literature compilations (DelSontro et al., 2018; Keller et al., 2021). In the latter case, uncertainty in total lake area distributions is challenged by the fact that no existing database is a complete census of all lakes (Messager et al., 2016) and the distribution of individual lake areas spans a range of approximately 7 orders of magnitude. The largest lakes are so big they could otherwise be classified as inland seas (> 104 km^2) while the smallest are barely larger than a regulation sized soccer (football) field (> 10^{-3} km^2).

A consequence of these challenges is that uncertainty varies within the distribution as a function of lake size. The area of the largest lakes is known with a high degree of certainty while the area of the smallest lakes is unknown below a certain unknown threshold. The area of a small lake can be unknown either because it is too small to be resolved from remote sensing products or because of limited ground surveying effort. The omission or "censoring" of small lakes occurs because we know that small lakes exist but below a certain threshold, we have limited knowledge of their exact areas (Hamilton et al., 1992). Such censoring errors may have an outsized impact on upscaled estimates of aquatic processes such as CO^2 flux because small lakes have been shown to have particularly large CO^2 flux rates (DelSontro et al., 2018; Pi et al., 2022).

Estimating total lake area from a sample of lakes requires a conceptual model of how lakes are formed (i.e. the data generating process). Typically, lake areas are treated as arising from a fractal generating process due to the fact that landform topography, which determines the placement of lakes, can itself be treated as a fractal generating process. Indeed, many other geomorphological phenomena that are dependent on landform topography such as coastline length are often well-described by fractal generating processes (Newman, 2005). A challenge in modelling such data generating processes for lake areas is that large lakes likely follow a different data generating process than that of the smallest lakes. Whereas small lakes are constrained by landform topography, large lakes are essentially unconstrained by local landform topography and are instead constrained by the placement and arrangement of continents (Goodchild, 1988). As large lakes become even larger, they have a greater probability of intersecting a continent edge and becoming an estuary or embayment rather than a lake. Consequently, lake databases are said to be truncated on large lakes because we know that large lakes are essentially fixed in space and cannot occur in any given location (Hamilton et al., 1992).

Given that no existing database is a complete census of all lakes, yet we have near exact estimates of the area of large lakes, estimating total lake area requires a method of dealing both with the fact that lake databases are 1) truncated at large lakes and 2) censored at small lakes. Prior studies estimating global lake area have mostly not dealt with the first issue (but see Seekell and Pace (2011)). Note that in some instances the estimation process can be modified to limit its effect on the results (see Methods section). In contrast to the first issue, prior studies have typically addressed the second issue by specifying an ad-hoc cutoff value below which empirically determined lake areas are discarded and subsequently back-calculated. Hereafter, I refer to this strategy as the "cutoff" method.

In the present study, I explore two different methods for calculating uncertainty bounds around global lake area. First, I calculate these bounds using a frequentist approach for estimating total lake area (i.e. the cutoff method) via a simulation study. Second, I compare these frequentist uncertainty bounds with those calculated using a Bayesian framework. I carry out these demonstrations using a simulated dataset so that the ability of each method to recover the "true" parameter values can be evaluated. The use of a simulated dataset has the further benefit of avoiding potential confounding factors such as heterogeneity of survey effort or unknown data precision. The following analysis assumes that total global lake area has minimal temporal variability or at least minimal spatial trend. As a result, reported total lake areas can be thought of as “anticipated” or equilibrium long-run lake area rather than true totals representing any specific point in time.

2 Methods

2.1 Data description

I compared the properties of a simulated dataset of lake areas (described below) against that of the HydroLAKES dataset (Messager et al., 2016). HydroLAKES was created as a compilation of existing broad scale lake datasets including the SRTM Water Body Data (Slater et al., 2006) and contains information on lakes between 0.1 and 500 km^2. The total area of the HydroLAKES dataset is 2.67 mil. km^2.

2.2 Model overview

Lake areas are typically treated as arising from a scale-invariant fractal generating process (Winslow et al., 2015; Downing et al., 2006; McDonald et al., 2012; Goodchild, 1988; Hamilton et al., 1992). This means that the number of lakes in one size class is proportional to the number of lakes in the preceding size class irrespective of their magnitudes. The numerical form describing such a process is a power-law function. One of the statistical tools often used to model data that follow a power-law function is the Pareto distribution which has a probability density function (pdf) of:

$$pdf(A) = \alpha x_{min}^{\alpha} A^{-(\alpha+1)} \tag{1}$$

where A is lake area, a controls the "shape" of the distribution and x_{min} controls the "scale" of the distribution (Shalizi, 2017). Lake area studies using the Pareto distribution do not typically use the pdf directly. Instead, they use the inverse (complementary) cumulative distribution function (ccdf) (i.e. quantile function):

$$ccdf(A) = \frac{x_{min}}{(1-A)^{\frac{1}{\alpha}}} \tag{2}$$

The reason for using the ccdf is two-fold. First, it stabilizes model estimates in the lower tail of the distribution (Newman, 2005). This can be seen from the simulated data in Figure 1 where the Pareto pdf contains a lot of noise in the tail but the ccdf appears smoothed. The smoothing of the tail is also desirable because it functions as a way of dealing with the truncated nature of lake databases (i.e. the area of large lakes is known exactly). The second reason for using the ccdf is that it provides a computational shortcut for estimating the Pareto shape parameter a because it is numerically equivalent to the slope of the ccdf in log-log space (Downing et al., 2006).

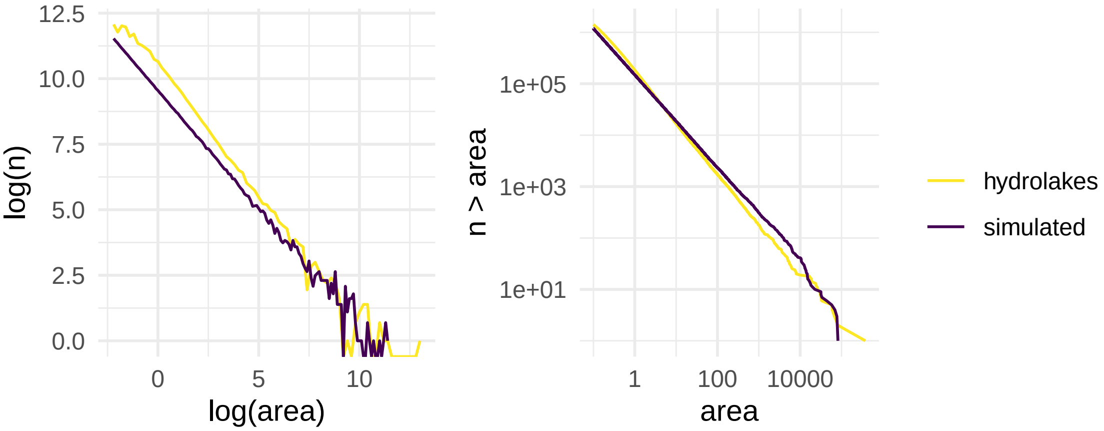

Figure 1: Realization of a Pareto (A) probability density function and (B) complementary cumulative distribution function compared against equivalent calculations on the HydroLAKES dataset.

For evaluation purposes, I generated a simulated dataset of 153,000 lake areas following the Pareto distribution using inverse transform sampling (Newman, 2005). Lakes in my simulated dataset have a minimum and maximum area of approximately 0.1 and 81,000 km^2 respectively. This maximum was chosen to be approximately as large as Lake Superior but less than the Caspian Sea following (Lehner and Döll, 2004). The number of lakes in the simulated dataset was adjusted so that the total "true" area of the dataset (3.28 mil km^2) would approximately match (but still exceed) the total reported in the HydroLAKES dataset (2.93 mil km^2). I simulated a censored lake dataset by excluding approximately 60% of lakes in the total dataset. I approximated the "true" lake area total by constructing the empirical distribution function (edf) of the data which approximates the underlying Pareto cdf (Newman, 2005). Then I used this estimate of the cdf slope to generate cdf estimates for the censored lakes. I combined these cdf estimates with the edf values from the known lakes before calculating the sum of the inverted distribution 2.

I estimated the Pareto shape parameter a in a frequentist framework by calculating the the slope of the edf in log-log space using linear regression in R (Team et al., 2018). I evaluated uncertainty in both a and total lake area in a Bayesian framework using Stan (Stan Development Team, 2017). Instead of computing on the edf (as in the frequentist case), I computed directly on the pdf with a Stan model (See Stachelek, 2022). I used uninformative gamma priors for both the x_{min} and a parameters following Scollnik (2007). I ran the model with four chains of 25,000 iterations and used the Stan defaults for thinning and burn-in which specify no thinning and discarding the first half (12,500) of iterations. I examined model fits to ensure that all models had acceptable convergence of MCMC chains.

Finally, I evaluated empirical uncertainty in total lake area using the size-dependent detection/non-detection estimates reported by Cheruvelil et al. (2021) which found that about 80% of lakes < 1 km^2 and 95% of lakes between 1 and 10 km^2 were accurately represented in the US National Hydrograph Network. I implemented a random sampling (n=500) procedure using these detection fractions whereby empirical uncertainty is reported as the quantiles of the resulting distribution.

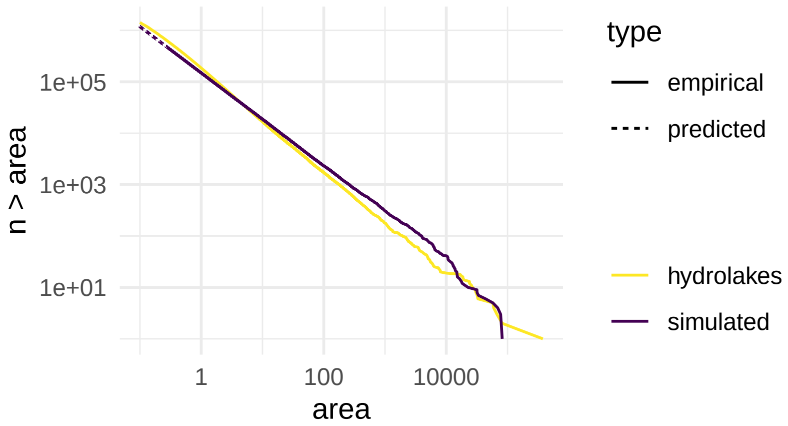

Figure 2: Simulated lake area (purple line) where the portion calculated from the edf (>= 0.1 km^2) is denoted by a solid line and the portion calculated from the transformed cdf (< 0.1 km^2) is denoted by a dashed line. For comparison, the edf of the HydroLAKES dataset (solid yellow line) is also shown.

Table 1: Comparison between the true total lake area in the simulated dataset against frequentist and Bayesian estimates. The true total is a point estimate while uncertainty is displayed via differences between the 95th and 5th quantile.

	Q50 (mil km^2)	Q5 (mil km^2)	Q95 (mil km^2)	Q95-Q5
True (point estimate)	3.278	-	-	-
True (80% non-detect)	3.205	3.205	3.205	673
Frequentist	3.274	3.274	3.274	14
Bayesian	3.199	3.158	3.273	115172

3 Assessment

The total area of the simulated dataset (3.28 mil. km^2, Table 1) was larger than the empirical total of the HydroLAKES (2.67 mil. km^2) dataset. Visual inspection of the frequentist results, which are computed on the edf, appeared to produce a reasonable density estimate for small, censored lakes (Figure 2). Although frequentist estimates of total lake area were close to the "true" value, uncertainty bounds were unrealistically small compared to an empirical estimate of uncertainty (Table 1).

In contrast to the fixed a and fixed total lake area estimates produced by the frequentist approach, I found substantial variability in both a (95% CI: 0.65, 1.02) and total area using a Bayesian approach (Figure 3, 4). In particular, the Bayesian 95% credible intervals for a encapsulated the true value (Figure 3). Despite larger uncertainty estimates using a Bayesian approach (i.e. more closely matching the empirical estimate of uncertainty), the true value of total lake area was underestimated. The reason for this underestimation is likely owed to incomplete “back calculation” in the cutoff method (i.e. back calculated lakes did not extend all the way to the smallest lakes in the simulated dataset).

4 Discussion

I have shown that the even with realistic estimates of the density of small censored lakes, the typical frequentist cutoff method does not reasonably capture uncertainty in total lake area (Table 1). However, the results of the Bayesian approach by contrast, indicate substantial uncertainty in both total lake area and the underlying Pareto shape parameter a used to derive these estimates (Figure 3, 4). The combination of accurate lake area accounting and realistic uncertainty estimates offered by this approach may provide a more principled way to assess the relative influence of different lake area classes on global lake fluxes (CO^2, CH^4, etc.) beyond ad-hoc study intercomparisons or even meta-analysis on collections of prior studies.

The Bayesian model estimate of total lake area was an underestimate of the true total lake area of the simulated datasets (Figure 4). Some of this underestimation is because the Pareto shape parameter a was estimated from the truncated dataset (to simulate the fact that observed datasets such as HydroLAKES have some unknown degree of truncation) and was lower (~0.82) than the true value of 0.9. This lower value of a resulted in lower than expected total lake

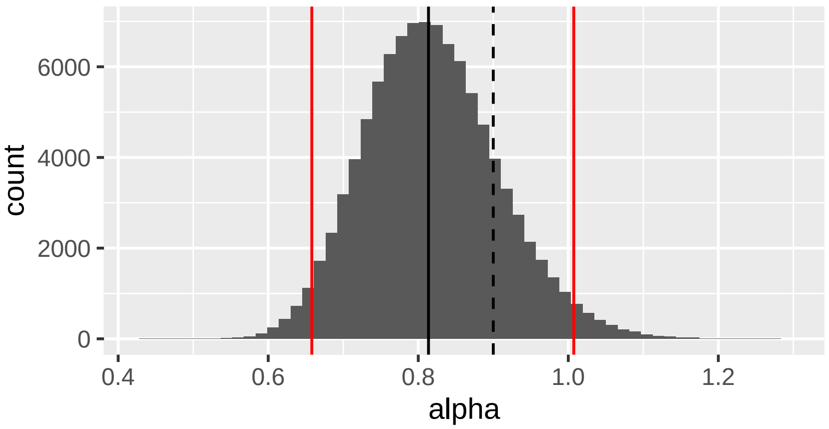

Figure 3: Median (black line) and central 95 percent interval estimates of alpha (red lines). Here the 'true' alpha is 0.9.

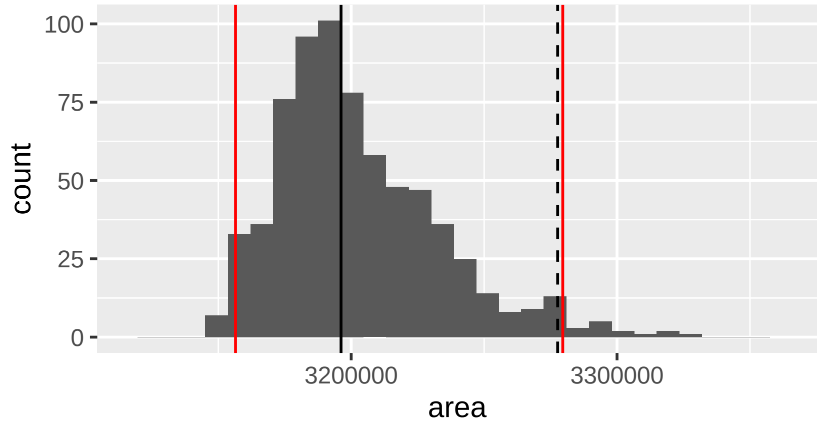

Figure 4: Median (black line) and central 95 percent interval estimates of total lake area (red lines). Here the true total lake area is marked with a dashed vertical line.

area. Such underestimation would increase with a greater proportion of censoring. Another likely contributor to lake area underestimation is differences in the properties of the simulated data compared to an empirical dataset like HydroLAKES. Note that the simulated dataset diverges notably from HydroLAKES for the largest waterbodies because these are few in number such that randomness in the simulation is more obvious (Figure 1). Taken together, these factors (i.e. less and less truncation and better data on the largest waterbodies) may explain the steady increase in estimates of global lake area through time from approximately 3 to 5 million km^2 as lake area databases have improved their accuracy (Downing et al., 2006; Lehner and Döll, 2004; Verpoorter et al., 2014). Future work on estimating global lake area may consider implementing a sensitivity analysis looking at the response of total area estimates to variation in the degree of censoring.

Another area where further sensitivity analyses may be warranted is in the generation of total lake area estimates from Pareto realizations. Typically this is done a single time such that any particular reported estimate comes only from a single realization. To generate the simulated dataset for the present study, I generated multiple realizations with the goal of arriving at one that approximately matched the total reported in HydroLAKES. Although each realization matched the Pareto properties of HydroLAKES, the sum of many of the realizations differed markedly from each other and from HydroLAKES. This demonstrates the need for generating multiple Pareto estimates of total lake area for aquatic process upscaling as part of a sensitivity analysis rather than taking only a single realization.

A caveat of the present study is that it is difficult to confirm whether or not any particular dataset follows a fractal generating or Pareto process without strong prior knowledge. As a result, there is some degree to which we cannot know whether the distribution of lake areas truly aligns with the Pareto distribution (Seekell and Pace, 2011). Indeed data following many heavy-tailed distributions such as the lognormal or negative exponential can appear to be equivalent to power-law distributions such as the Pareto (Clauset et al., 2009; Seekell and Pace, 2011). The results of the present study confirm that an apparent change in the shape of the lower tail of a lake area distribution does not necessarily indicate a change in the data generating process. Note that the simulation dataset analyzed herein is a truly fractal generated dataset with a homogeneous data generating process, yet a cursory look would seem to indicate a change of data-generating process in the lower tail (Figure 2).

In addition to a sensitivity analysis of censoring, future work might consider more complex models that treat lake areas as a mixture of a Pareto distribution for small lakes and either a negative exponential or lognormal distribution for large lakes. Such an approach has been demonstrated by (Bonabeau et al., 1999) and (Scollnik, 2007). Both studies show that the point at which the distribution mixtures converge can provide valuable inference. In the case of lakes, such a convergence point may indicate a change in the data generating process such as the point at which lake areas are controlled by continent placement rather than fractal landscape morphology (Goodchild, 1988; Hamilton et al., 1992). Knowing such a convergence point would provide a data-driven estimate of what constitutes a “large lake” beyond ad-hoc cutoffs and supplement existing definitions based on hydrodynamics and circulation.

5 Data Availability Statement

All data and code associated with this manuscript are available at: https://doi.org/10.5281/zenodo.7459226

6 Acknowledgements

This work was supported by the U.S. National Science Foundation’s Harnessing the Data Revolution Program (OAC-1934633) and Los Alamos National Laboratory (LDRD-20210213ER, LDRD-20220697PRD1). I thank the following individuals for a friendly review of an earlier draft: P.C. Hanson, E. Zipkin.

References

Bonabeau, E., Dagorn, L., and Freon, P., 1999. Scaling in animal group-size distributions. *Proceedings of the National Academy of Sciences*, 96(8):4472–4477. doi:10.1073/pnas.96.8.4472.

Cheruvelil, K. S., Soranno, P. A., McCullough, I. M., Webster, K. E., Rodriguez, L. K., and Smith, N. J., 2021. Lagos-us locus v1. 0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous us. *Limnology and Oceanography Letters*, 6(5):270–292.

Clauset, A., Shalizi, C. R., and Newman, M. E., 2009. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703.

DelSontro, T., Beaulieu, J. J., and Downing, J. A., 2018. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change: GHG emissions from lakes and impoundments. *Limnology and Oceanography Letters*. doi:10.1002/lol2.10073.

Downing, J. A., Prairie, Y. T., Cole, J. J., Duarte, C. M., Tranvik, L. J., Striegl, R. G., McDowell, W. H., Kortelainen, P., Caraco, N. F., and Melack, J. M., 2006. The global abundance and size distribution of lakes, ponds, and impoundments. *Limnology and Oceanography*, 51(5):2388–2397.

Goodchild, 1988. Lakes on fractal surfaces: A null hypothesis for lake-rich landscapes. *Mathematical Geology*.

Hamilton, S. K., Melack, J. M., Goodchild, M. F., and Lewis, W., 1992. Estimation of the fractal dimension of terrain from lake size distributions. *Lowland floodplain rivers: Geomorphological perspectives. Wiley*, pages 145–163.

Keller, P. S., Marcé, R., Obrador, B., and Koschorreck, M., 2021. Global carbon budget of reservoirs is overturned by the quantification of drawdown areas. *Nature Geoscience*. doi:10.1038/s41561-021-00734-z.

Lehner, B. and Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, 296(1-4):1–22. doi:10.1016/j.jhydrol.2004.03.028.

McDonald, C. P., Rover, J. A., Stets, E. G., and Striegl, R. G., 2012. The regional abundance and size distribution of lakes and reservoirs in the United States and implications for estimates of global lake extent. *Limnology and Oceanography*, 57(2):597–606. doi:10.4319/lo.2012.57.2.0597.

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, 7:13603. doi:10.1038/ncomms13603.

Newman, M., 2005. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46(5):323–351. doi:10.1080/00107510500052444.

Pi, X., Luo, Q., Feng, L., Xu, Y., Tang, J., Liang, X., Ma, E., Cheng, R., Fensholt, R., Brandt, M., Cai, X., Gibson, L., Liu, J., Zheng, C., Li, W., and Bryan, B. A., 2022. Mapping global lake dynamics reveals the emerging roles of small lakes. *Nature Communications*, 13(1):5777. doi:10.1038/s41467-022-33239-3.

Scollnik, D. P. M., 2007. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, 2007(1):20–33. doi:10.1080/03461230601110447.

Seekell, D. A. and Pace, M. L., 2011. Does the pareto distribution adequately describe the size-distribution of lakes? *Limnology and Oceanography*, 56(1):350–356.

Shalizi, C. R., 2017. *Advanced Data Analysis from an Elementary Point of View*.

Slater, J. A., Garvey, G., Johnston, C., Haase, J., Heady, B., Kroenung, G., and Little, J., 2006. The srtm data “finishing” process and products. *Photogrammetric Engineering & Remote Sensing*, 72(3):237–247.

Stachelek, J., 2022. VeinsOfTheEarth/pareto_lake_area. `https://doi.org/10.5281/zenodo.7459226`. doi:10.5281/zenodo.7459226.

Stan Development Team, P., 2017. Stan modeling language users guide and reference manual, version 2.17. 0. *Technical report*.

Team, R. C. et al. *R: A Language and Environment for Statistical Computing*, 2018.

Verpoorter, C., Kutser, T., Seekell, D. A., and Tranvik, L. J., 2014. A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, 41(18):6396–6402. doi:10.1002/2014GL060641.

Winslow, L., Read, J., Hanson, P., and Stanley, E., 2015. Does lake size matter? Combining morphology and process modeling to examine the contribution of lake classes to population-scale processes. *Inland Waters*, 5(1):7–14. doi:10.5268/IW-5.1.740.