Geometric models overestimate lake depth due to imperfect slope measurement

J. Stachelek[1], P. J. Hanly[1], and P. A. Soranno[1]

[1]Department of Fisheries and Wildlife, Michigan State University, 480 Wilson Rd., East Lansing, MI 48824, USA

Key Points:

- Geometric models to predict lake depth, which require in-lake slope, assume that nearshore land slope is a good proxy for in-lake slope.
- Using data from thousands of lakes, we show that nearshore land slope is a poor proxy for in-lake slope and increases prediction error.
- Prediction errors were systematic such that depth was overpredicted in concave and reservoir lakes.

Corresponding author: J. Stachelek, `stachel2@msu.edu`

Abstract

Lake depth is a critical characteristic that influences many important ecological processes in lakes. Unfortunately, lake depth measurements are labor-intensive to gather and are only available for a ~~small~~ tiny fraction of lakes globally. Therefore, scientists have tried to predict lake depth from characteristics that are easily obtained for all lakes such as lake surface area or the slope of the land surrounding a lake. One approach for predicting lake depth simulates lake basins using a geometric model where nearshore land slope is assumed to be a representative proxy for in-lake slope and the distance to the center of the lake is assumed to be a representative proxy for the distance to the deepest point of the lake. However, these assumptions have rarely been tested in a broad range of lakes. We used bathymetry data from approximately 5,000 lakes and reservoirs to test these assumptions and to examine whether differences in lake type or shape influences depth prediction error. We found that nearshore land slope was not representative of in-lake slope and using it for prediction increases error substantially relative to models using true in-lake slope. We also found that models using nearshore land slope as a proxy systematically overpredict lake depth in concave lakes (i.e. bowl-shaped; up to 18% of lakes in the study population) and reservoir lakes (up to 30% of lakes in the study population), suggesting caution in using geometric models for depth prediction in unsampled lakes.

1 Introduction

Lake depth is an important factor controlling lake physics, chemistry, and biota. ~~Deeper~~ For example, deeper lakes generally have higher water clarity and less complete mixing compared to shallow lakes (Fee et al., 1996; Read et al., 2014). These differences are reflected in variation among lakes in terms of biological productivity (Qin et al., 2020) and rates of greenhouse gas production (Li et al., 2020). However, because measured depth data is only available for a ~~small fraction (~~tiny fraction of lakes (including $<$ 25% ~~)~~ of all lakes in our study footprint), our ability to understand and predict depth-dependent processes is limited. The importance of lake depth, coupled with its limited availability, has led to numerous attempts to predict depth using measures available for all lakes such as lake surface area or the nearshore slope of the land surrounding a lake (Heathcote et al., 2015; Oliver et al., 2016; Sobek et al., 2011). Such efforts rely on a strategy of exploiting correlations between nearshore geomorphology and in-lake geometry, which at limited extents can be quite strong, while at larger extents can be dependent on geographic location and lake type (Oliver et al., 2016; Branstrator, 2009). Given the limited prediction accuracy of prior depth prediction efforts (± 6-7 m), a major focus has been on improving ~~prediction~~ accuracy using strategies such as employing more diverse covariates (Oliver et al., 2016), varying lake buffer sizes (Heathcote et al., 2015), or estimating hidden groupings (e.g. fitting different models for distinct size classes) among lakes ~~.~~(Cael et al., 2017; Sobek et al., 2011). Unfortunately, the predictive accuracy of these efforts has been limited.

One intuitive approach for predicting lake depth involves using a geometric model that assumes lake basins correspond to an idealized shape such as a cone, bowl, or an elliptic sinusoid (Getirana et al., 2018; Hollister et al., 2011; Neumann, 1959; Yigzaw et al., 2018). All such geometric models for lake depth prediction involve implicit assumptions about the terms of geometric formulae. In the simplest case, where lakes basins are treated as cones (Equation 1, Figure 1), two assumptions are required to make depth predictions for all lakes: 1) that nearshore land slope is a representative proxy for in-lake slope and 2) that the distance to the center of the lake is a representative proxy for the distance to the deepest point of the lake (Figure 1). This cone model imposes the following fixed (i.e. geometric) relationship between slope and horizontal distance:

$$depth_{geometric} = tan(slope) * distance \quad (1)$$

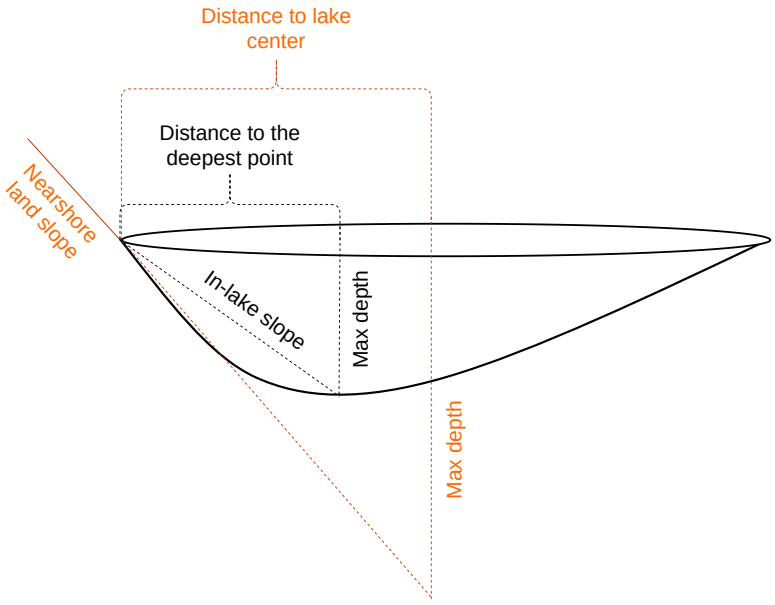

Figure 1. Diagram showing the relations between true (black) and proxy (orange) metrics of lake geometry. Geometric depth calculated via Equation 1 requires a single distance and slope metric.

where the product of slope and horizontal distance yields an exact geometric depth estimate ($depth_{geometric}$).

The assumptions of the cone model (as well as other geometric models) can be tested by comparing proxy measures of lake geometry against corresponding "true" (i.e. in-lake) values derived from bathymetric maps and by evaluating how lake cross-section shapes differ from that of an idealized cone (Johansson et al., 2007). For instance, lake cross-section shapes have been shown to vary from narrow "convex" forms to outstretched "concave" forms (Hakanson, 1977). Because tests of geometric model assumptions require bathymetric map data, which is only available for a tiny fraction of lakes (including about 15% of all lakes in our study footprint), existing evidence may not be applicable to all lakes. The few studies that have tested these assumptions have been limited to individual studies of very large (> 500 ha) lakes or studies on small numbers (< 100) of lakes (Johansson et al., 2007). Studies focused specifically on reservoirs (as opposed to the more typical case where reservoirs and natural lakes are combined), have been even more restricted to that of extremely large lakes > 1000 ha (Lehner et al., 2011; Messager et al., 2016).

As a result of this limited testing, we lack knowledge on both the predictive performance of geometric models, the effect of proxies on depth prediction, and whether depth predictions are more sensitive to measurement errors in the horizontal dimension (i.e. distance to the deepest point of the lake) or measurement errors in the vertical dimension (i.e. in-lake slope). Additionally it is unclear whether model prediction error is related to differences in lake type such those with different cross-section shapes or those classified as reservoirs versus natural lakes. Given these knowledge gaps, we asked three research questions: 1) How representative is nearshore land slope of in-lake slope; and how representative is the distance to the center of a lake compared to the distance to the deepest point of a lake? 2) How does the use of proxies for lake geometry affect lake depth prediction error? 3) How does lake cross-section shape (i.e. concave versus convex) and lake type (i.e. natural lake vs reservoir) affect depth prediction error? To answer these questions, we extracted maximum depth (hereafter referred to as "observed

maximum depth"), in-lake slope, cross-section shape (i.e., concave versus convex), and distance to the deepest point, of approximately 5,000 lakes from bathymetric map data and supplemented this data with classification estimates of whether lakes are reservoirs or natural lakes. We used this data to compute geometric depth estimates (Equation 1) and prediction "offsets" to these estimates using the random forest algorithm (Equation 3). Covariates used in offset modeling included a variety of lake, watershed, and hydrologic subbasin measures that are available for all lakes (Table S1).

By definition, the distance proxy (distance to the center of the lake) must always be greater or equal to the true distance value (distance to the deepest point of the lake). Therefore, we expect that the use of this proxy will lead to overestimation of lake depth (Figure 1). Furthermore we expect to see greater overestimation error in reservoirs as compared to natural lakes because many reservoirs are known to be drowned river valleys where the deepest point is close to the edge at the end of the reservoir (i.e. next to the dam) rather than in the center of the reservoir (Lanza & Silvey, 1985). In a similar fashion, we expect to see overestimation error associated with using a nearshore land slope proxy in lakes with differing cross-section shape such that the depth of U-shaped (i.e. concave) lakes will be overpredicted whereas the depth of V-shaped (i.e. convex) lakes will be underpredicted (Figure S1). Finally, we expect that depth predictions themselves will be strongly related to lake area and hydrologic subbasin variables as these measures have been influential in prior studies (Oliver et al., 2016).

By testing these expectations, we can establish whether barriers to increased depth prediction accuracy lie in lack of correspondence between true and proxy measures of lake geometry or in hidden groupings among lakes (such as lake cross-section shape or reservoir status). This information could help direct future research efforts to focus on particular dimensions of lake geometry (i.e. horizontal versus vertical) or to stratify model predictions based on specific lake types and cross-section shapes. Ultimately, achieving increased depth prediction accuracy would allow for more precise estimates of depth-dependent biotic and chemical processes across broad spatial extents.

2 Methods

2.1 Data description

We compiled bathymetry data on approximately 5,000 lakes in the Northeastern and Midwestern US from nine official state databases (Figure S2). These lakes represent a diverse cross section of lakes in terms of their surface areas (4 - 18500 ha) and span a wide geographic extent including glaciated and non-glaciated regions. The original data came in a variety of formats including pre-interpolated rasters (Minnesota), contour lines (Nebraska, Michigan, Massachusetts, Kansas, Iowa), contour polygons (New Hampshire, Connecticut), or point depth soundings (Maine). For the Minnesota data, we simply clipped the raster for each lake to its outline. For data from the remaining states, we processed each lake by converting its original representation to a point layer (if necessary), rasterizing these points, and creating an interpolated bathymetry "surface" using a simple moving window average in the `raster` R package (Hijmans, 2019). The size of the moving window was adjusted iteratively to ensure that each bathymetry raster contained no missing data.

All lake bathymetry was specifically calculated relative to high-resolution (1:24,000 scale) ~~NHD~~ National Hydrography Dataset (USGS, 2019) waterbodies such that source data and bathymetry surface outputs were clipped to the area of each lake polygon. We restricted the lakes in our study to those with an area of at least 4 ha and a maximum depth of at least 0.3 m (1 ft). The purpose of these restrictions was to ensure that lakes had enough contours (or points, or polygons) to generate adequately smooth interpolations with which to calculate in-lake geometry metrics.

We used our generated bathymetry surfaces to find the location of the deepest point in the lake and we resolved ties by choosing the deepest point that was closest to the center of the lake. We used the location of this deepest point to calculate "distance to the deepest point" as the minimum distance to the lake shoreline. To account for lakes where the centroid does not intersect lake bathymetry because it is located within an embedded island or peninsula, we calculated the center of the lake not as its centroid but rather by finding the point farthest from the lake shoreline (i.e. its "visual distance to lake center"). For these calculations, we used the `polylabelr` R package (Larsson, 2019), which interfaces with the Mapbox pole of inaccessibility algorithm (Agafonkin, 2019). We calculated in-lake slope as maximum lake depth divided by the distance to the deepest point ~~and we~~ where maximum lake depth is a point measurement derived from a bathymetric surface and distance to the deepest point is the smallest straight-line distance from that point to the waterbody perimeter. We calculated nearshore land slope for each lake by computing the slope within a 100-m buffer using data from a high resolution digital elevation model (~15x15m grain) accessed using the `elevatr` R package (Hollister & Shah, 2017)~~and computed using the terrain~~ . Slope computations proceeded by passing a 3x3 moving window over the 100-m buffer to calculate the slope at each point using Horn's algorithm via the `terrain` function in the `raster` R package (Hijmans, 2019). Reported nearshore land slope values are the mean of all points in the buffer. In addition to the aforementioned techniques of calculating in-lake (and nearshore) slopes and distances, we tried 7 alternate techniques which are described in Table S2.

We categorized lakes based on their cross-section shape and reservoir class. For cross-section shape, we categorized lakes as either convex or concave following the method of Hakanson (1977) by computing normalized lake depth-area relationships (i.e. hypsographic curves) and assigning class membership based on whether a lake's curve falls above or below that of a simple straight-sided cone (Figure S3).

We further classified lakes using the output of a ~~machine learning algorithm to assign a probability to each lake~~ deep convolutional neural network model trained on satellite images labeled according to whether there was visual evidence of a water control structure significantly impacting flow. This model had an overall validation accuracy of 81% and produced a probability for each waterbody as to whether it is a reservoir or a natural lake. For our purposes, we ~~determined a lake to be a reservoir if the classification probability was 0.75 or greater. Our reservoir classification data~~ set a conservative classification probability threshold of 0.75 to determine whether a lake would be considered a reservoir. Note that our reservoir classification defines reservoirs as any permanent waterbody that has a water control structure likely to significantly impact flow or pool water~~, beyond simply controlling water level~~. It makes no distinction between different dam types~~or dam heights.~~ , heights, or uses/purposes.

Covariates used in random forest modeling (Table S1, Equation 3) for lake elevation, area, island area, perimeter, shoreline development, watershed to lake area ratio, and hydrologic subbasin (i.e. HUC4s), were obtained from the LAGOS-US LOCUS database. One such measure, that of shoreline development, is a measure of lake perimeter shape defined as:

$$shoreline_{devel} = perimeter/(2 * \sqrt{(\pi * waterarea * 10000)}) \quad (2)$$

where sinuous lakes have larger values of shoreline development and circular lakes have smaller values of shoreline development. Watershed to lake area ratio is an approximation of water residence time and is defined as watershed area divided by lake area (Timms, 2009).

2.1.1 Proxy evaluation

We conducted a qualitative assessment of whether or not proxy measures of lake geometry are representative of their true values by visual inspection (i.e. plotting each proxy measure against its corresponding true value) and by computing coefficients of determination (R^2). We further tested proxy measures by examining their effect on lake depth prediction error. Our approach involved several steps. In the first step, we computed a geometric estimate of lake depth using only geometry information ($depth_{geometric}$, Equation 1). In the second step, we fit a random forest model to predict observed (i.e. true) depth as a function of geometric depth along with several covariates available for all lakes (Table S1). The purpose of this random forest "offset" modeling was to more rigorously test our expectations regarding prediction error among different formulations of $depth_{geometric}$ and among different lake types. Each of these steps were executed iteratively for each combination of true and proxy values of slope and distance (Table ??1).

2.2 Model description

2.2.1 Geometric model

We used a geometric model of lakes where basins are treated as cones with a fixed relationship between slope and distance (Equation 1). One reason that we used the cone model is that, unlike other idealized shapes, it is well suited for modeling maximum depth and does not require any knowledge of lake volume or mean depth. Note that Equation 1 is a geometric formula and has no intercept or ~~"coefficients "~~ coefficients and it produces a perfect depth estimate given true values of slope and distance. To use this model to predict the depth of all lakes, there is a necessary assumption that proxy slope and distance measures, which are available for all lakes, are representative of true slope and distance (Figure 1).

2.2.2 Random forest models

Prior studies using geometric models to predict lake depth include a statistical or machine learning model "layer" or "offset" to boost predictive accuracy (Hollister et al., 2011; Yigzaw et al., 2018). For our purposes, this offset modeling enabled us to test our expectations that prediction error would be different among different formulations of $depth_{geometric}$ and among different lake types. It also facilitated direct comparison against prior models of lake depth including those that are non-geometric. We generated an "offset" to geometric depth (sensu Hollister et al., 2011) using the random forest algorithm and the `ranger` R package (Wright & Ziegler, 2017) to predict observed maximum depth as a function of covariates including geometric maximum depth (from Equation 1) along with the lake elevation, area, perimeter, and ratio/index measures listed in Table 1:

$$depth_{observed} \sim depth_{geometric} + covariates \tag{3}$$

Neither cross-section shape nor reservoir class was used as a covariate in any random forest models. We used the random forest algorithm because it makes no assumptions about the distribution of model residuals, allows for non-linearity, and is insensitive to interactions (i.e. multicollinearity) among covariates (Prasad et al., 2006).

2.2.3 Model comparisons

We tested model sensitivity to slope and distance proxies by generating multiple "geometric maximum depth" estimates from 3 different model runs using each of the possible metric combinations for Equation 1 (true slope - proxy distance, proxy slope - true distance, proxy slope - proxy distance). Prior to entry into Equation 1, we standardized

proxy distances to have the same numeric range as their true counterpart. The purpose of this standardization was to prevent lakes with extremely long proxy distances from having an outsized impact on model evaluation metrics. In addition to comparing among model runs using different metric combinations, we compared among sets of model runs where slope and distance measures were calculated using different sets of calculation techniques (Table S2).

2.2.4 Model evaluations

We evaluated model fit and prediction error using root-mean-square error (RMSE) and coefficient of determination (R^2) metrics on a holdout set containing 25% of all lakes. We evaluated the residuals of each model relative to lake cross-section shape and reservoir classes to determine whether depth is consistently over or under predicted for some lake types relative to others.

3 Results

Lakes belonging to each cross-section shape and reservoir class were not evenly distributed across our study area (Figure S2, S3). For example, concave lakes were nearly absent from Michigan whereas Maine lakes had an overabundance of lakes categorized as neither concave nor convex. Lakes in the southern portions of our study area tended to be classified as reservoirs whereas lakes in the northern portions of our study area were a more even mix between reservoirs and natural lakes (Figure S2). Approximately 18%, 80%, and 2% of lakes were classified as having a concave, convex, or neither shape respectively whereas approximately 30% and 70% of lakes were classified as being a reservoir or a natural lake.

Although proxy distance to lake center was often ~~much~~ larger in magnitude compared to the true distance to the deepest point of lakes', they were strongly related (R^2 = 0.8). ~~In contrast~~Note that the coefficient of determination for this relationship is not strictly correct given that distance to lake center is an upper bound on distance to the deepest point of lakes. In contrast to distance metrics, proxy nearshore land slope and true in-lake slope were more weakly related (R^2 = 0.17). For slope measures, most lakes had lower magnitude (i.e. shallower) nearshore land slope compared to true in-lake slope (Figure ~~3~~2). Taken together, these results suggest that proxy distance to the center of lakes is representative of true distance to the deepest point of lakes whereas proxy nearshore land slope is not representative of true in-lake slope. The strong relationship between distance to the center of lakes and distance to the deepest point means that it is possible to compute a "correction factor" to convert between the two measures (See best-fit equations in Figure 2).

In addition to overall differences between slope and distance measures, we found differences in these relationships among lake shape classes. For example, in-lake slope and distance to the deepest point of the lake metrics were consistently larger in magnitude for convex lakes as compared to concave lakes (Figure S4). ~~However, there~~ We found evidence that this difference was at leat partly explained by the fact that convex lakes are deeper than concave lakes (Figure S5). There were not similar differences among slope and distance metrics for natural lakes versus reservoirs~~(Figure S4)~~.

Model fit and prediction error differed depending on the technique used to calculate in-lake and nearshore geometry metrics (Table S2). We found that the best model fit and lowest model error occurred when in-lake slope was calculated as the average point-wise slope of all points at maximum lake depth rather than a single point of maximum depth. However, given the small difference in the fit of models using either of these techniques and the significant cost in terms of computational load and complexity, we limit our discussion hereafter to the simpler case involving only a single deepest point.

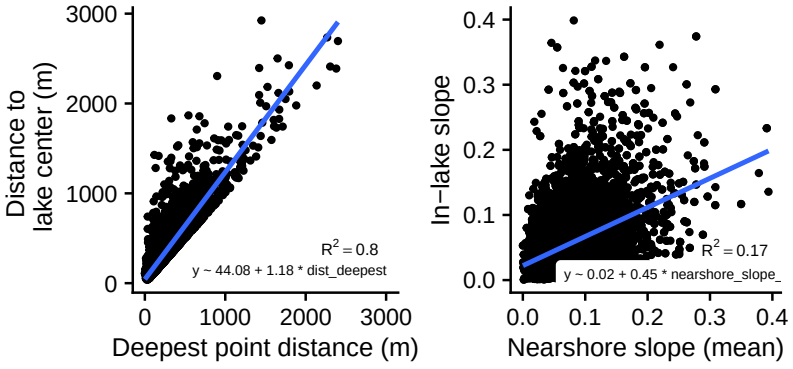

Figure 2. Comparison among proxy and true values of lake geometry for A) distance to deepest point versus distance distance to lake center and B) nearshore land slope versus in-lake slope. A best-fit line and coefficient of determination is shown to illustrate representativeness.

The use of proxy nearshore land slope had a larger effect on model fit and prediction error than the use of proxy distance to lake center (Table ~~??~~1). More specifically, the true slope - proxy distance model had a better fit (R^2 = ~~0.73~~0.7) and lower prediction error (RMSE = ~~4.23m~~4.4m, MAPE = 29%) compared to the proxy slope - true distance model (R^2 = ~~0.26~~0.32, RMSE = ~~6.87m~~6.6m, MAPE = 60%). Furthermore, analysis of model residuals showed overestimation of lake depth for concave lakes when models included a proxy slope measure (Figure ~~??~~3). We observed similar but smaller overestimation depending on if a lake was classified as a reservoir rather than a natural lake (Figure ~~??~~3).

Table 1. Model fit and predictive accuracy metrics (RMSE = root mean square error, R^2 = coefficient of determination, MAPE = mean absolute percent error) for all combinations of true (in-lake slope, distance to the deepest point of the lake) and proxy (nearshore land slope, distance to lake center) metrics.

slope	distance	RMSE	R^2	MAPE
true	true	-	-	-
true	proxy	~~4.2~~ 4.4 m	~~0.73~~0.70	29 %
proxy	true	~~6.9~~ 6.6 m	~~0.26~~0.32	60 %
proxy	proxy	~~6.6~~ 6.4 m	~~0.31~~0.35	59 %

The most important covariates in offset models were those relating to spatial location, lake area, and perimeter (Figure ~~S5~~S7). Conversely, watershed metrics and lake elevation had little contribution to random forest model fit (Figure ~~S5~~S7). The spatial location (i.e. HUC4) covariate was notably less importance in the true slope model compared to the two proxy slope models. Model importance calculations indicated that omitting a geometric max depth measure results in a 130%, 60%, or 50% increase in mean square error depending on the formulation of geometric max depth in Equation 1 (Figure ~~S5~~S7).

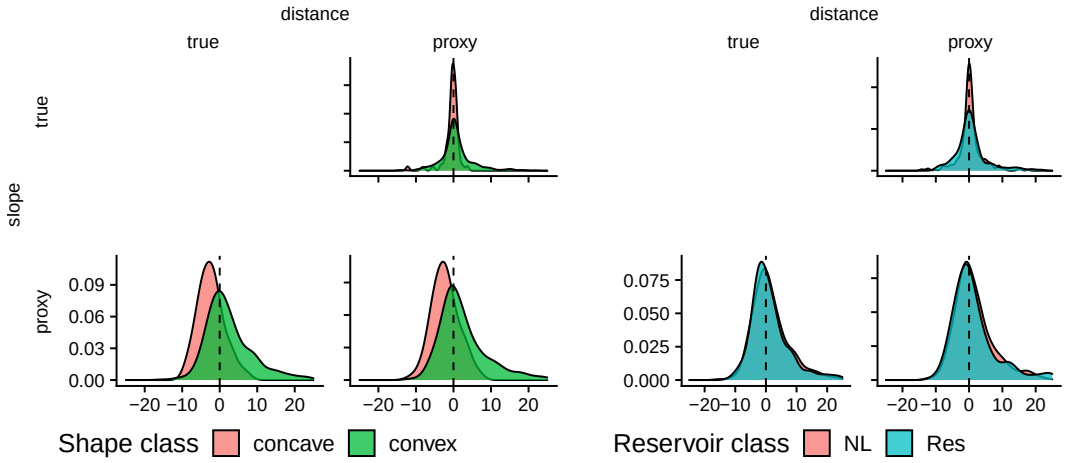

Figure 3. Depth model residuals (residual = observed - predicted) in meters by cross-section shape and reservoir class indicating overprediction of concave and reservoir lakes.

4 Discussion

Our tests of geometric lake depth models show that specific proxy measures of lake geometry are not representative of true geometry measures across a broad array of lakes. Models using non-representative proxies showed increased error and systematic overestimation of depth in concave and reservoir lakes. Although our analysis was limited to lakes with available bathymetry data, these lakes did not have characteristics that differed from that of the overall lake population (Figure ~~S6,~~ S8, S10). Although there is a possibility that there is some hidden bias not explored for in our analyses, this lack of difference suggests that our results are likely to be broadly applicable to all lakes.

4.1 Representativeness of proxy measures of lake geometry

In comparing among lake geometry measures, our analysis suggests that proxy distance to lake center is representative of true distance to the deepest point of the lakes but that proxy nearshore land slope is not representative of true in-lake slope. A simple indication of this non representativeness is that proxy nearshore land slope was often (in $> 74\%$ of cases) steeper than true in-lake slope. This finding is consistent with Heathcote et al. (2015) whos results suggest that in-lake slopes are shallower compared to the surrounding land. ~~The shallow nature of~~ Furthermore, the fact that in-lake slopes ~~is likley a function of erosion and sediment transport processes~~ were shallower compared to the surrounding land even after controlling for differences in area (Figure S5) is consistent with the idea of topographic scaling (i.e. scale invariance) explored in previous work and detailed by (Cael et al., 2017). The underlying reason for these shallow in-lake slopes may be related to slope-induced turbidity currents which distribute sediment from shallow high-energy areas of lakes to deep low-energy areas (Håkanson, 1981; Johansson et al., 2007). The strength of such sediment focusing is likely greater in "younger" lakes with steeper slopes leading to a smoothing of their bathymetry over time (Blais & Kalff, 1995).

One surprising finding with respect to the relationship between true and proxy geometry measures when examined by lake class was the fact that there was no greater difference between proxy and true distances in reservoirs compared to natural lakes. This is contrary to the idea that most reservoirs are drowned river valleys where the deepest point is close to the edge at the end of the reservoir (i.e. next to the dam) rather than in the center of the reservoir (Lanza & Silvey, 1985). One possible explanation is that

our reservoir classification data uses a more general definition of a reservoir (i.e. any permanent waterbody that has a water control structure likely to significantly impact flow or pool water) compared to that of conventional classifications that are tied to specific dam types or dam heights. Another possible explanation is that conventional reservoir classifications are conceptually biased towards more southern areas with few natural lakes (Figure S2).

We found other differences among lake geometry measures according to lake cross-section shape. One finding was that convex lakes, when compared to concave lakes, had longer distances to lake centers relative to corresponding distances to the deepest point of lakes. In addition, convex lakes often had steeper in-lake slopes relative to nearshore land slopes as compared to concave lakes. Finally, it was notable that convex lakes were deeper than concave lakes despite having similar distributions of lake surface area (Figure ~~S7~~S9). The underlying cause of these differences is unknown but one possibility is that geometry is tied to the circumstances of lake formation whereby the formation of concave lakes were a result of more intense glacial scouring compared to that of convex lakes (Gorham, 1958). While our findings provide some evidence in support of this idea, namely that there is a geographic hotspot of concave lakes associated with the glaciated "prairie pothole region" (see Hayashi & van der Kamp, 2000), the overall geographic distribution of lake cross-section shapes does not support this idea. Instead of a concentrated area of concave lakes in formerly glaciated regions, there appears to be a fairly even mix of concave and convex lakes distributed amongst the northern (i.e. glaciated) and southern (non-glaciated) portions of our study area (Figure S2).

4.2 Effects of proxy measures of lake geometry depth prediction error

Models using only proxy variables had prediction error rates (RMSE = ~~6.6~~6.4m) of a similar magnitude as that of prior studies (RMSE = 6 - 7.3m) predicting lake depth at broad geographic extents (Hollister et al., 2011; Oliver et al., 2016; Messager et al., 2016). When only a single proxy measure was used, there was a difference in model sensitivity depending on if it was a horizontal distance measure or a vertical slope measure. In the case of a true slope and proxy distance combination, models were more accurate (± ~~4.2m~~4.4m, 29%) than even the most accurate of prior studies (Hollister et al., 2011; Oliver et al., 2016; Messager et al., 2016). Conversely, models using a proxy slope and true distance combination had prediction error rates (± ~~6.9m~~6.6m, 60%) of a similar magnitude as that of the baseline proxy-proxy model (± ~~6.6m~~6.4m, 59%). The greater sensitivity of depth predictions to proxy slope measures relative to proxy distance measures may be explained by the fact that proxy slope measures were a more imperfect representation of true in-lake slopes relative to proxy versus true distances. ~~In addition, these results~~ We did not find evidence that the outsized sensivity of depth predictions to slope was dependent on variations in how these measures were calculated (Table S2). In a general sense, the sensitivity of depth predictions to slope help explain the relatively poor predictive performance of prior non-geometric lake depth models given that they rely heavily on lake area as a predictor (Messager et al., 2016; Oliver et al., 2016; Sobek et al., 2011) and both horizontal distance measures and vertical slope measures appear to be decoupled from lake area (Figure ~~S7~~S9).

4.3 Effects of lake shape and lake type on depth prediction error

As expected, we found that the maximum depth of concave lakes was systematically overpredicted by a simple geometric model using proxy nearshore land slope (Figure S1). However, contrary to our expectation, we did not observe underprediction of depth in convex lakes. The reason we did not observe underprediction of the depth of convex lakes is likely because geometric depth itself was always greater than observed maximum depth owing to the fact that proxy distance is constrained to be greater than true distance. This

suggests that depth estimates in prior studies may be overestimated when they encompass large numbers of lakes with diverse cross-section shapes.

4.4 Future research

One of our models (true slope, proxy distance) was more accurate than even the most accurate of prior studies. However, parameterization of this model requires data on bathymetry which is not available for all lakes. We propose that the error rate of this model (± ~~4.2m~~4.4m, 29%) be used as an out-of-sample prediction benchmark for future studies such that they should attempt to match it but not expect to exceed it.

Because this most accurate model requires bathymetry data, this suggests that it may not be possible with current data and models to produce depth predictions for all lakes with error rates below 6m. To achieve high prediction accuracy using data available for all lakes, future studies could explore alternative modeling approaches such as ordinal modeling, which would capture whether or not a lake crosses some important depth threshold but would not seek to predict a specific depth value, or emerging data types such as "topobathymetric" products that integrate both topographic and bathymetric data in a seamless fashion rather than treating them as separate entities. Topobathymetry would allow for more robust tests of the representativeness of geometric model inputs. Unfortunately, topobathymetric products are rare, have mostly been limited nearshore marine environments, and as such are not yet widely available for inland waters (Danielson et al., 2016).

Finally, our findings indicate that geometry measures differ according to lake cross-section shape. This makes it an attractive target for inclusion in depth prediction models. Unfortunately, identifying a lake's cross-section shape requires bathymetry data which is unavailable for most lakes. However, given the conceptual links between cross-section shape, glaciation, and sedimentation (Johansson et al., 2007) it may be advantageous for future studies to compile data on sedimentation to determine if this data can be used to predict cross-section shape and boost depth prediction accuracy.

5 Conclusion

To our knowledge, the present study is the largest and most comprehensive test to date of geometric models of lake depth. Using bathymetry data on approximately 5,000 lakes, we show that proxy slope measures are not representative of true in-lake slope and this leads to inaccuracies in predicting the depth of concave and reservoir lakes. These ~~innaccuracies~~ inaccuracies suggest that caution is warranted in using geometric models for depth prediction in unsampled lakes. Despite these apparent biases, overall prediction accuracy was equivalent to that of prior depth prediction studies (± 6-7m). Only our models using a true measure of in-lake slope, which is limited in availability to lakes with bathymetry data and where we already know lake depth, had greater accuracy than that of prior studies (± ~~4.2m~~4.4m, 29%). Lack of improved prediction accuracy (short of including data that is unavailable for most lakes) suggests that improved prediction may require new types of data or novel analysis techniques.

Acknowledgments

All data as well as code for data processing, model fitting, and model evaluation is available at [Zenodo DOI]. Funding was provided by the US NSF Macrosystems Biology Program grants, DEB-1638679; DEB-1638550, DEB-1638539, DEB-1638554. PAS was also supported by USDA National Institute of Food and Agriculture Hatch Project, Grant Number: 176820. Author contributions: JS conceived of the study, built models, analyzed data, and wrote the paper. PJH and PAS provided interpretation of results and edited the paper. This work benefited from participation in the Global Lake Ecological

Observatory Network (GLEON). We thank K.S. Cheruvelil for a friendly review of an earlier draft.

References

Agafonkin, V. (2019). *A JS library for finding optimal label position inside a polygon*. Retrieved from `https://github.com/mapbox/polylabel`

Blais, J. M., & Kalff, J. (1995). The influence of lake morphometry on sediment focusing. *Limnology and Oceanography*, *40*(3), 582–588. doi: 10.4319/lo.1995.40.3.0582

Branstrator, D. K. (2009). Origins of types of lake basins. In *Encyclopedia of inland waters* (pp. 613–624). Elsevier Inc.

Cael, B. B., Heathcote, A. J., & Seekell, D. A. (2017). The volume and mean depth of Earth's lakes. *Geophysical Research Letters*, *44*(1), 209–218. doi: 10.1002/2016GL071378

Danielson, J. J., Poppenga, S. K., Brock, J. C., Evans, G. A., Tyler, D. J., Gesch, D. B., ... Barras, J. A. (2016). Topobathymetric Elevation Model Development using a New Methodology: Coastal National Elevation Database. *Journal of Coastal Research*, *76*, 75–89. doi: 10.2112/SI76-008

Fee, E. J., Hecky, R. E., Kasian, S. E. M., & Cruikshank, D. R. (1996). Effects of lake size, water clarity, and climatic variability on mixing depths in Canadian Shield lakes. *Limnology and Oceanography*, *41*(5), 912–920. doi: 10.4319/lo.1996.41.5.0912

Getirana, A., Jung, H. C., & Tseng, K.-H. (2018). Deriving three dimensional reservoir bathymetry from multi-satellite datasets. *Remote Sensing of Environment*, *217*, 366–374. doi: 10.1016/j.rse.2018.08.030

Gorham. (1958). The Physical Limnology of Northern Britain: An Epitome of the Bathymetrical Survey of the Scottish Freshwater Lochs, 1897.–1909. *Limnology and Oceanography*, 11.

Hakanson, L. (1977). On Lake Form, Lake Volume and Lake Hypsographic Survey. *Geografiska Annaler. Series A, Physical Geography*, 31.

Håkanson, L. (1981). On lake bottom dynamics—the energy–topography factor. *Canadian Journal of Earth Sciences*, *18*(5), 899–909. doi: 10.1139/e81-086

Hayashi, M., & van der Kamp, G. (2000). Simple equations to represent the volume–area–depth relations of shallow wetlands in small topographic depressions. *Journal of Hydrology*, *237*(1-2), 74–85. doi: 10.1016/S0022-1694(00)00300-0

Heathcote, A. J., del Giorgio, P. A., Prairie, Y. T., & Brickman, D. (2015). Predicting bathymetric features of lakes from the topography of their surrounding landscape. *Canadian Journal of Fisheries and Aquatic Sciences*, *72*(5), 643–650. doi: 10.1139/cjfas-2014-0392

Hijmans, R. J. (2019). *Raster: Geographic data analysis and modeling*. Retrieved from `https://CRAN.R-project.org/package=raster`

Hollister, Milstead, W. B., & Urrutia, M. A. (2011). Predicting maximum lake depth from surrounding topography. *PloS one*, *6*(9), e25764. doi: 10.1371/journal.pone.0025764

Hollister, J., & Shah, T. (2017). *Elevatr: Access elevation data from various APIs*. Retrieved from `http://github.com/usepa/elevatr`

Johansson, H., Brolin, A. A., & Håkanson, L. (2007). New Approaches to the Modelling of Lake Basin Morphometry. *Environmental Modeling & Assessment*, *12*(3), 213–228. doi: 10.1007/s10666-006-9069-z

Lanza, G. R., & Silvey, J. (1985). Interactions of reservoir microbiota: Eutrophication—related environmental problems. In *Microbial processes in reservoirs* (pp. 99–119). Springer.

Larsson, J. (2019). *Polylabelr: Find the pole of inaccessibility (visual center) of a polygon*. Retrieved from `https://github.com/jolars/polylabelr`

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., ... Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*(9), 494–502. doi: 10.1890/100125

Li, M., Peng, C., Zhu, Q., Zhou, X., Yang, G., Song, X., & Zhang, K. (2020). The significant contribution of lake depth in regulating global lake diffusive methane emissions. *Water Research*, 115465. doi: 10.1016/j.watres.2020.115465

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & 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

Neumann, J. (1959). Maximum depth and average depth of lakes. *Journal of the Fisheries Board of Canada*, *16*(6), 923–927. doi: 10.1139/f59-065

Oliver, S. K., Soranno, P. A., Fergus, C. E., Wagner, T., Winslow, L. A., Scott, C. E., ... Stanley, E. H. (2016). Prediction of lake depth across a 17-state region in the United States. *Inland Waters*, *6*(3), 314–324.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. doi: 10.1007/s10021-005-0054-1

Qin, B., Zhou, J., Elser, J. J., Gardner, W. S., Deng, J., & Brookes, J. D. (2020). Water Depth Underpins the Relative Roles and Fates of Nitrogen and Phosphorus in Lakes. *Environmental Science & Technology*, *54*(6), 3191–3198. doi: 10.1021/acs.est.9b05858

Read, J. S., Winslow, L. A., Hansen, G. J., Van Den Hoek, J., Hanson, P. C., Bruce, L. C., & Markfort, C. D. (2014). Simulating 2368 temperate lakes reveals weak coherence in stratification phenology. *Ecological Modelling*, *291*, 142–150. doi: 10.1016/j.ecolmodel.2014.07.029

Sobek, S., Nisell, J., & Fölster, J. (2011). Predicting the depth and volume of lakes from map-derived parameters. *Inland Waters*, *1*(3), 177–184. doi: 10.5268/IW-1.3.426

Timms, B. (2009). Geomorphology of Lake Basins. In *Encyclopedia of Inland Waters* (pp. 479–486). Elsevier. doi: 10.1016/B978-012370626-3.00024-7

USGS. (2019). *National hydrography Dataset* (Tech. Rep.). Retrieved 2018-04-30, from `https://nhd.usgs.gov/`

Wright, M. N., & Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi: 10.18637/jss.v077.i01

Yigzaw, W., Li, H.-Y., Demissie, Y., Hejazi, M. I., Leung, L. R., Voisin, N., & Payn, R. (2018). A New Global Storage-Area-Depth Data Set for Modeling Reservoirs in Land Surface and Earth System Models. *Water Resources Research*, *54*(12). doi: 10.1029/2017WR022040