Predicting Lake Depths from Topography to Map Global Lake Volume

By
Nataniel M. Holtzman

Senior Honors Thesis
Department of Geological Sciences
University of North Carolina at Chapel Hill

2016

Approved by:
Tamlin M. Pavelsky

Abstract

Nataniel M. Holtzman: Predicting Lake Depths from Topography to Map Global Lake Volume
(Under the direction of Tamlin M. Pavelsky)

The depth of a lake affects its role in climate and biogeochemical cycling. There is a lack of lake depth data due to the difficulty of measuring bathymetry, which impedes the accurate inclusion of lakes in climate models and the assessment of global water resources and carbon storage. However, lake depths can be estimated from land topography, for which remotely-sensed DEM data is available. We develop a simple statistical model to predict lake depth from two explanatory variables: the mean relief above the lake surface of an area around the lake, and whether the lake's location was glaciated in the last ice age. The model is based on 328 lakes with known depths, located on all continents but Antarctica, and has an r^2 of 0.57. We then apply this model to a database of over 200,000 lakes to produce global gridded maps of predicted total lake volume and average depth. The realistic depth estimates provided by our model can improve the accuracy of future studies of climate and water resources.

Table of Contents

List of Figures

Figure

List of Tables

Table

Introduction

Lakes are important to climate, biogeochemistry, and ecology, and human activities. A lake's depth modulates its role in all of these systems. Depth affects the lake's residence time for nutrients and pollutants (Hollister et al., 2011), and what community of organisms can live in the lake. Knowledge of those factors is vital to models of biogeochemical cycles and of ecology. Data on residence time and water volume itself is also needed for regional water resources planning (Sobek et al., 2011). Recently, remote sensing has allowed for better assessments of lake occurrence on a global scale (Verpoorter et al, 2014). Such products are a step towards quantifying the total contribution of lakes to biogeochemical cycles, but without depth data we do not have a complete picture. Unknown lake depths produce significant uncertainty in regional and global estimates of carbon cycling (Sobek et al., 2011).

Being filled with water, lakes have a greater heat capacity than land. Large lakes can delay the seasonal temperature changes on nearby land. Inclusion of the Great Lakes and other large lakes makes regional climate models more accurate (Long et al., 2007). The climate impact of small lakes is not well-studied, although it is known to be significant at least in Arctic areas, where abundant lakes can cover more than one-third of a region's land surface (Martynov et al., 2012; Rouse et al., 2005).

Lakes are ignored in many global climate models that have coarse horizontal resolution (Subin et al., 2012). Regional models that do include lakes usually treat them in a very simplified manner, assigning all lakes the same depth. In an effort to improve on this representation, recent research has examined coupling RCMs to simple models of the internal thermal structure of lakes (Subin et al., 2012). Lake depth data becomes vital when using these coupled models. Even

if individual lakes are not resolved in a climate model, gridded values of lake depth and volume could be used in parameterizing the thermal properties of the land surface.

The only way to measure the depths – and thus volumes – of lakes is to use sonar or other methods that require visiting the lake and surveying it from a boat. Lake bathymetry cannot be reliably measured remotely except in the case of very clear lakes. In this study, we explore a method to approximate a lake's depth and volume, based on remotely sensed measurements of nearby land surface and the lake's location. Previous studies of this sort of depth prediction have focused on single regions such as Sweden, Quebec, or the eastern United States (Sobek et al., 2011; Heathcote et al., 2015; Hollister et al., 2011). In this paper we show that such a method works on a worldwide scale. We then apply this model to a database of over 200,000 lakes to produce global gridded maps of predicted average lake depth and total lake volume.

Data and Methods

Data

In order to develop statistical relationships to predict lake depth, we require a substantial amount of training data. We wrote a script to extract relevant information on lake depth and other variables from the World Lake Database website of the International Lake Environment Committee Foundation (ILEC), visiting each lake page and recording the lake name, area, latitude, longitude, maximum depth, and mean depth. The 481 lakes that had a mean depth were included in the training dataset. To obtain lake boundary information, we selected all the features of type "LAKE" from the Global Lakes and Wetlands Database (GLWD, Lehner and Doll, 2004). This selection excludes reservoirs and wetlands. We performed a spatial join of the GLWD lakes with the ILEC data, joining each ILEC lake point to the closest GLWD polygon within 5 km. For lakes that did not match GLWD polygons, we manually corrected their

coordinates where possible. The resulting dataset of 328 lakes constitutes our training dataset. Figure 1 shows the locations of these lakes, and Figure 2 (left column) compares their areas to the areas of all the GLWD lakes. The median training lake is larger in area than the median GLWD lake, with few training lakes on the scale of 1 km^2 that is approximately the modal area among all GLWD lakes. Geographically, the training lakes are especially concentrated in the US, southern Canada, Argentina and Chile, Europe, and Japan. There is poor coverage of the Arctic, especially Siberia, and of mainland Asia in general, which is unfortunate as those are some of the regions with the densest concentrations of lakes.

Methods

We set out to study the relationship between a lake's depth and the topographic relief in an area nearby the lake, obtained from a digital elevation model (DEM). We performed multiple linear regression in R with log(depth) as the response/dependent variable. The explanatory variables considered were mean, maximum, and minimum elevation near the lake; lake elevation; latitude; and lake area. We used the differences between topographic variables because it is the local relative topography that should affect lake depth. For each pair of topographic measurements m_1, m_2, we took $\log(|m_1 - m_2| + 1)$. The logarithm is used to make the relationship with *log(depth)* more linear, and the absolute value and plus one are included to make sure the inside of the logarithmic expression is positive. We chose the best model, which contained two variables.

To delineate the nearby area of consideration for measuring topographic variables, we constructed buffers around each lake polygon, with a width given by the formula identified by Heathcote et al., (2015):

$$Buffer\ width = \frac{1}{2}\sqrt{\frac{Lake\ area}{\pi}}$$

Calculations involving elevation were done in Google Earth Engine, using two DEMs: ASTER GDEM from Global Emissivity Database version 3, and SRTM (Table 1) (Hulley et al., 2015; Jarvis et al., 2008). For each DEM, we calculated the mean, maximum, and minimum elevations in each buffer, and the elevation of each lake was estimated as the modal elevation within the lake polygon.

The buffering and DEM calculations were repeated for all lakes in GLWD (not including reservoirs or lakes greater than 5000 km^2 in area, whose mean depths and volumes were taken from ILEC). For each DEM, we applied the corresponding model to predict the log depths all lakes and then their depths, and the predicted mean depth was multiplied by the known area to calculate a predicted volume. The lakes were aggregated into 1 degree by 1 degree square grid cells. For each cell, total lake area and total predicted lake volume were calculated. Total volume divided by total area gives an area-weighted average of lake depth. Since the prediction error in log(depth) is approximately normal for any lake, the error in depth is approximately lognormal. The error in total volume or average depth, then, is the sum of many lognormal variables, which cannot be expressed mathematically in a simple way. If there are a large enough number of lakes in a cell, the error in total volume will be approximately normal due to the Central Limit Theorem, but many grid cells contain only a few lakes. For this reason, a Monte Carlo method was used to assess errors in the gridded product. For each cell, the simulated log depth of each lake was sampled from a normal distribution centered at the model-predicted value, with standard deviation given by the model standard error. Then total volume was calculated based on these simulated log depths. This simulation was performed 2000 times for each cell to obtain a

distribution of predicted total volume. Several quantiles of this distribution were recorded in order to describe the error in predicted total volume.

Results

Of the 8 explanatory variables tested here, we developed a model with 2 variables. Table 2 summarizes the relationship between pairs of topographic measurements and *log(depth)*. The best predictor is $\log(|mean - elev| + 1)$, the mean relief of the buffer above the lake surface, which we will henceforth call *meanelev*. To assess whether the model should contain more variables than *meanelev* and an intercept, we found the correlations between the other variables and the residuals of the *meanelev* model (Table 3). The other topographic variables were not significant, as many of them are highly correlated with *meanelev* itself. The variable most strongly correlated with the residuals was latitude. However, the latitude effect is not large and becomes even smaller (although still statistically significant) when we include the dummy variable *isglacial*, which is 1 if a lake was covered by the ice sheet at the Last Glacial Maximum and 0 otherwise, using LGM ice extent shapefiles from Ray and Adams (2001). The increased depth of lakes in high latitudes (controlling for topography) can be explained by glacial lakes being deeper. Also, we were not confident that a possible latitude effect would be linear, as variables such as precipitation change non-linearly with latitude. More data on low-latitude lake depths are needed to assess whether there is a latitude effect on depth. There was no significant interaction between *meanelev* and either latitude or *isglacial*.

The 2-variable model (*meanelev* and *isglacial*) is summarized in Table 4 and Figure 3. To determine whether to include more variables, we performed all-possible-subsets regression with a maximum of 8 variables, using the *leaps* R package. Table 5 compares the fit of the 2-variable model with the model from all-possible-subsets regression with the best (lowest)

Bayesian Information Criterion (BIC), for both DEMs. The models with more variables are definitely better fits, but only by a small amount. (Adding more variables raised adjusted r^2 from 0.57 to 0.61 for ASTER, and 0.58 to 0.64 for SRTM.) The more variables the model includes, the more susceptible it is to overfitting and the less scientifically interpretable it is. We were particularly wary of overfitting because the distributions of the predictors and of lake area in the training data are quite different from those of the total GLWD lakes whose depths we aim to predict (Figure 2). For these reasons we chose to focus on the two-variable model. We also tried fitting the same two-variable model on each continent individually (Figure 4). The coefficient of *meanelev* was similar between Asia, Europe, and North America, and higher in South America. Africa has a lower value, but a large standard error of the coefficient as there are relatively few African lakes in the dataset. More depth data is needed to study regional differences in the relationship between topography and lake depth.

The gridded maps of our lake volume predictions are shown in Figures 5 and 6. The regions with high lake volume per land area, aside from very large lakes such as the Caspian Sea, include Northern Canada, the west coast of North America the southern Andes, Scandinavia, the southern Tibetan Plateau, and northeastern Siberia. When we divide by lake area to get average depth, the maps are even more closely associated with terrain, largely following mountain ranges as one would expect from our elevation-based model. Because of the glacial component of the model, high-latitude mountains such as the Putorana Plateau in central Siberia stand out more. Errors in the gridded predictions are closely related to the number of lakes in a grid cell and whether all those lakes have similar volumes. The lowest errors are found where there are many small-to-medium sizes lakes per cell, mostly in northern latitudes. The highest errors are found where there is just one lake in a cell, which is true of 15% of ASTER-predicted cells. ASTER-

based gridded volume predictions tended to be slightly higher than their SRTM-based counterparts (Figure 6). The mean difference was 0.013 *log10*(km^3), which corresponds to ASTER predictions being 3% higher than their SRTM equivalents. That is statistically significant in a paired t-test with $p < 0.001$. The latitudinal distributions of total lake area and volume (figure 8) are dominated by the Caspian Sea and North American Great Lakes (between 40 and 50 degrees north) and the African Great Lakes (between 0 and 10 south). There is a zone of roughly constant high lake area between 50 and 70 degrees north, but volume is much higher in its southern part than its northern part. There are zones of large volume relative to area at 20-10 S and 40-50 S, indicating that lakes in those latitudes are predicted to be especially deep, probably due to the presence of mountains there.

Discussion and Conclusion

Our results suggest that it is possible to estimate lake depth for individual lakes around the world with relatively high accuracy, using a small number of variables in a simple statistical model based on land topography. All that is needed to predict a lake's depth with our model is an outline of the lake's approximate extent, publicly available elevation data from the ASTER or SRTM DEM, and a map of LGM ice sheet extent. The model can be applied to better constrain estimates of water resources in areas with known lake extent but unknown lake depth. It can also be used to produce lake depth estimates for use in regional climate models, which should increase RCM accuracy compared to using the same depth for all lakes.

The global gridded estimates of lake water volume and average depth that we produced might be used in global climate models and global studies of carbon cycle dynamics. However, they are definitely an underestimate, because they are limited to the lakes in GLWD. GLWD intends to provide a global inventory of lakes with areas greater than 1 km^2. According to

Verpoorter et al. (2014), such medium and large lakes make up a minority of the world's lakes, but about 60% of total lake area. Also, GLWD undercounts lakes in its size range relative to Verpoorter et al.'s GLOWABO inventory, with about 250,000 lakes compared to about 350,000. As such, future advances could include applying our model to the Verpoorter et al. dataset or a similar product.

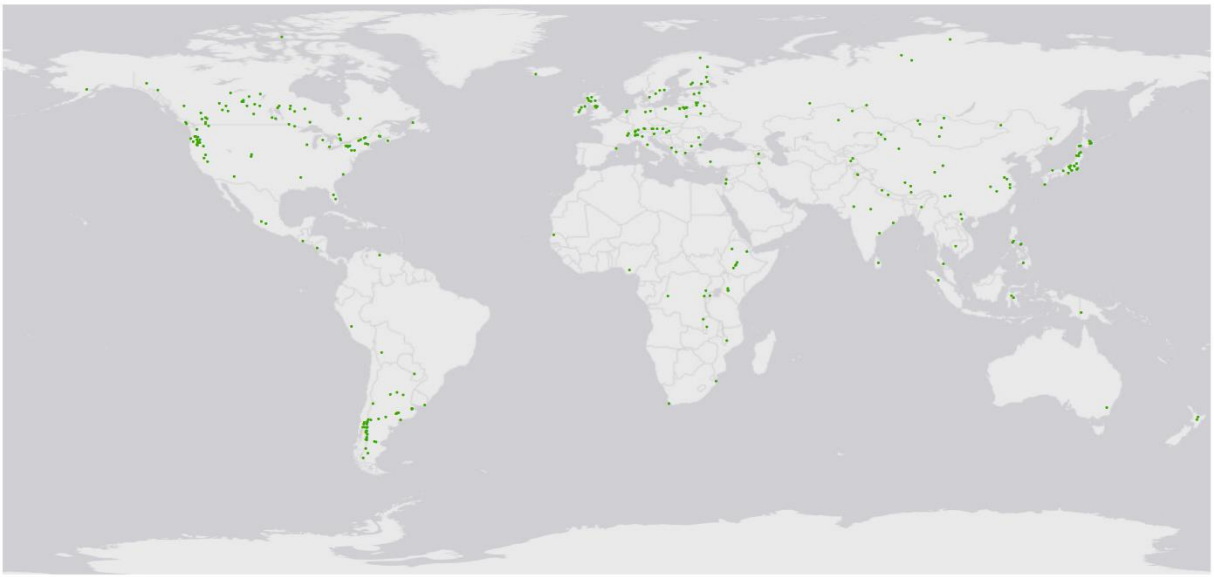

Figure 1. Geographic distribution of lakes in training dataset.

Figure 2. Comparison of training dataset with full GLWD.

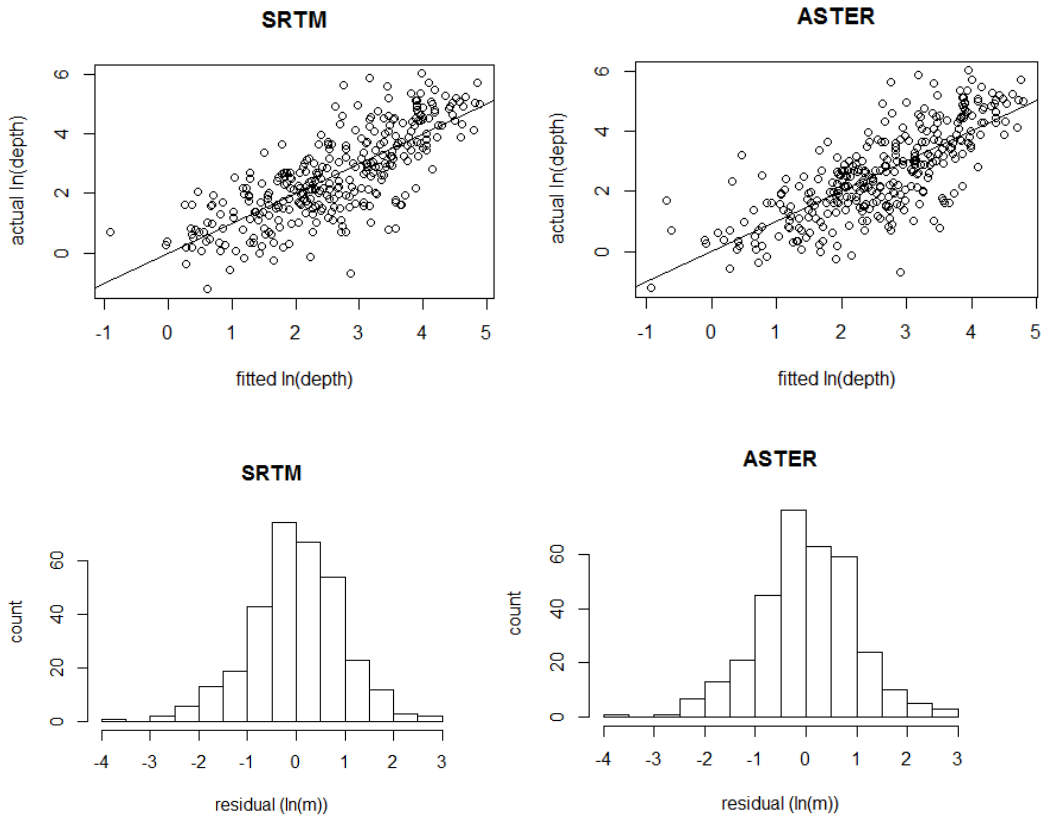

Figure 3. Plots of actual vs fitted ln(depth), and histograms of residuals from regression.

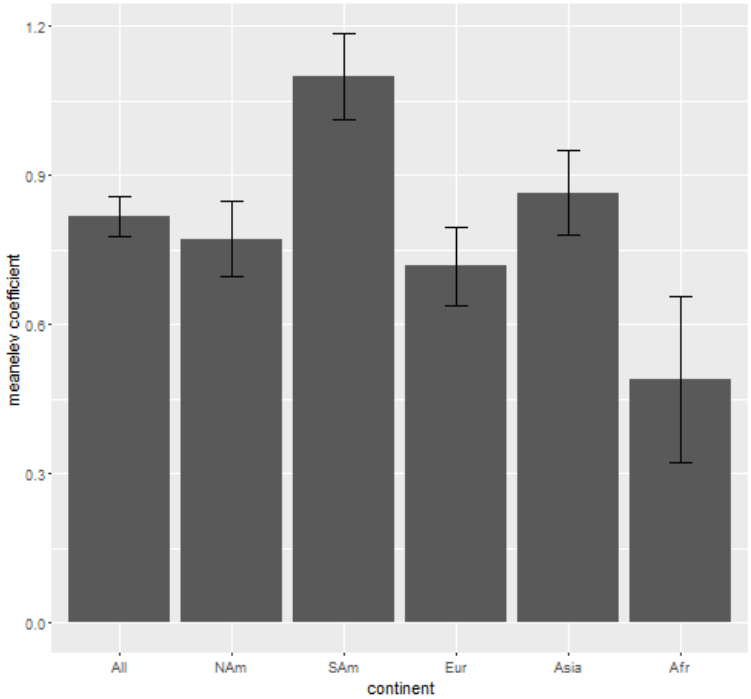

Figure 4. Variation in the *meanelev* coefficient of the model between continents. Error bars are 1 standard error.

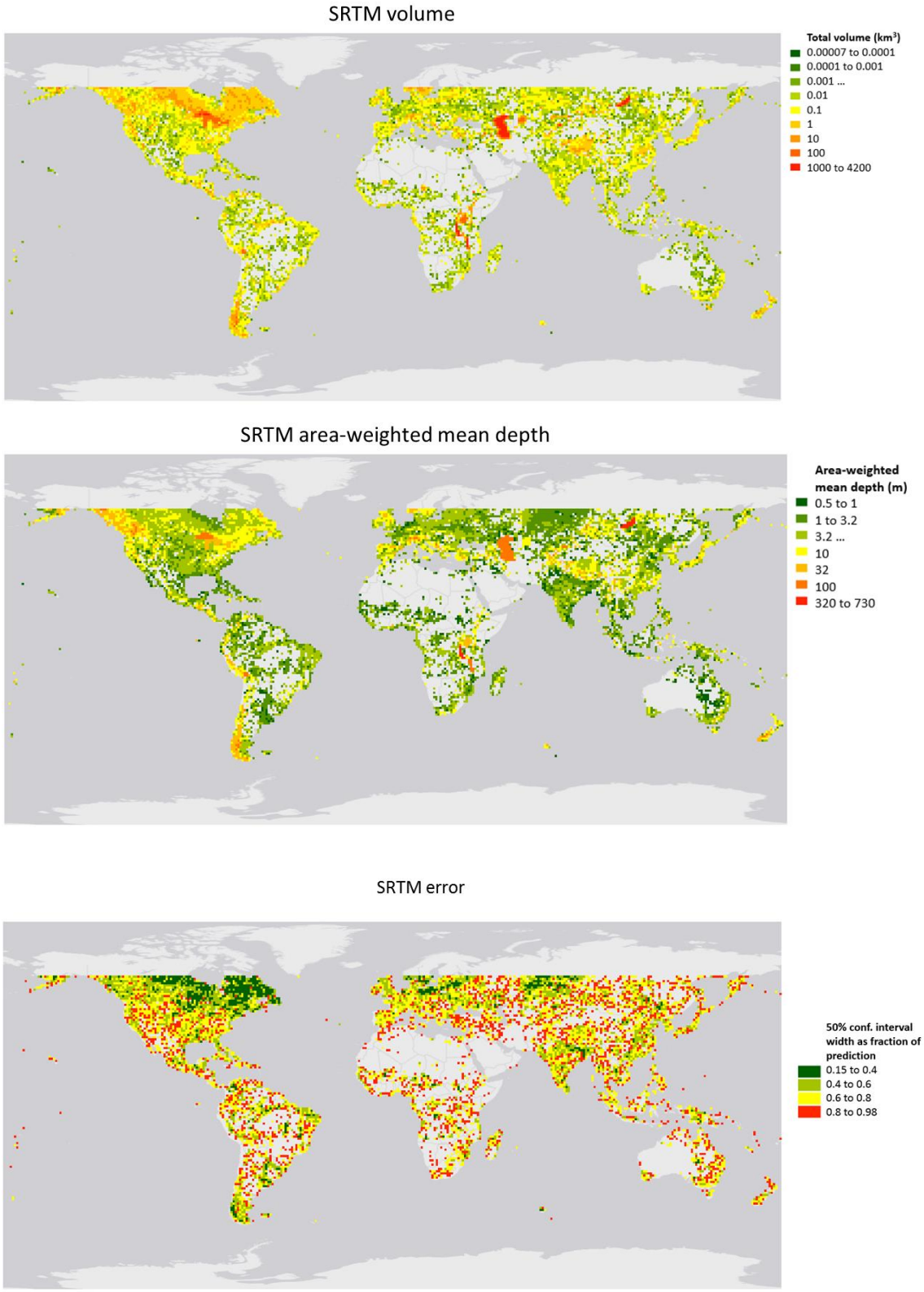

Figure 5. 1 degree by 1 degree gridded maps of model-predicted lake water storage using SRTM DEM.

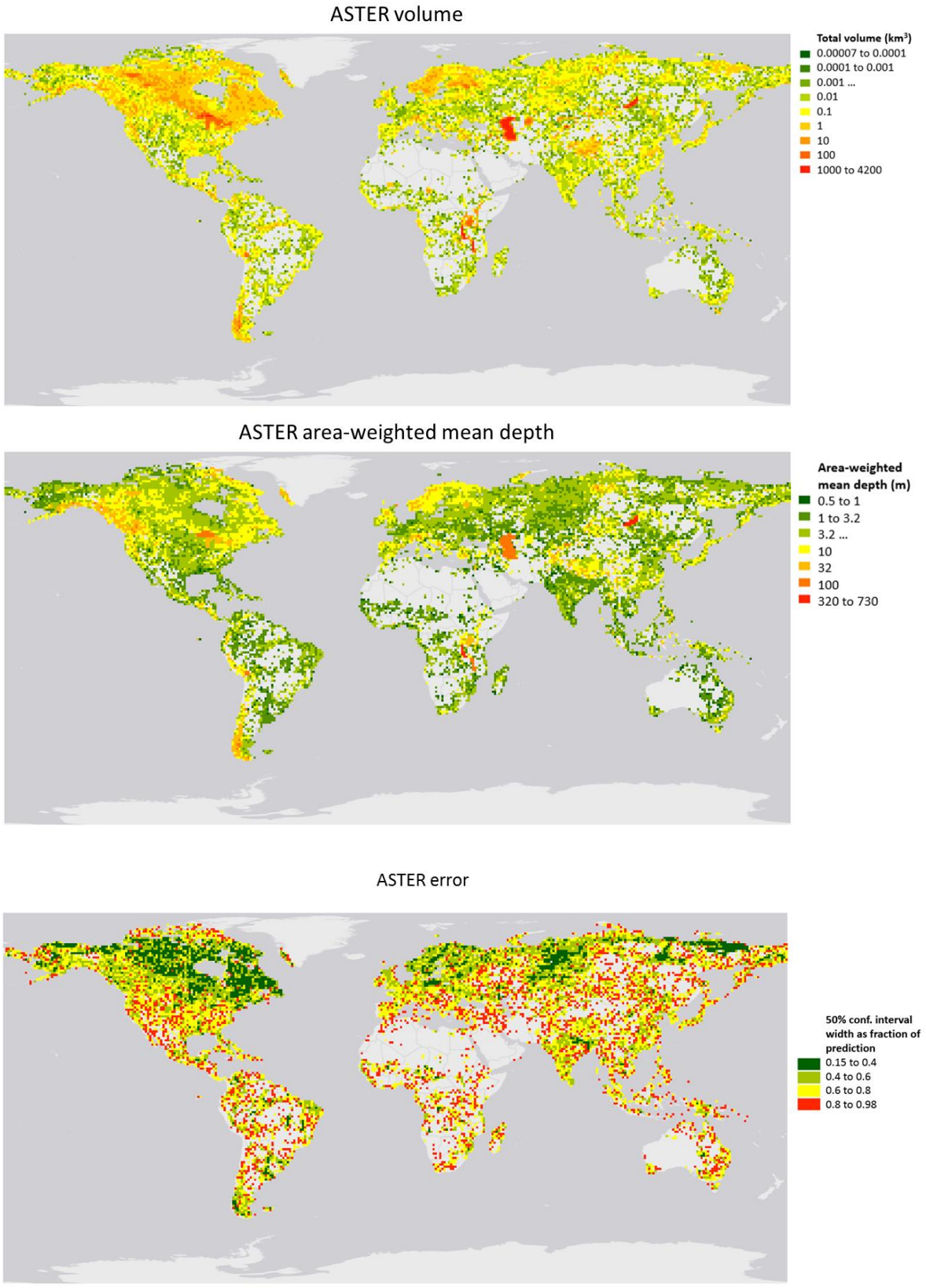

Figure 6. 1 degree by 1 degree gridded maps of model-predicted lake water storage using ASTER DEM.

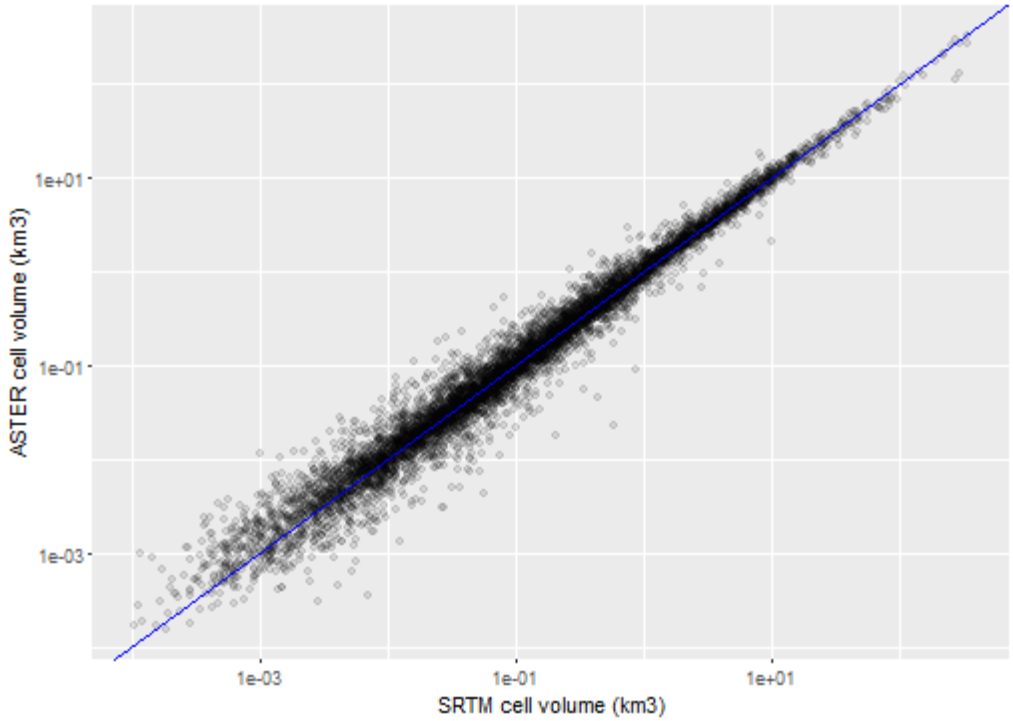

Figure 7. Relationship between ASTER-based and SRTM-based gridded lake water volume estimates

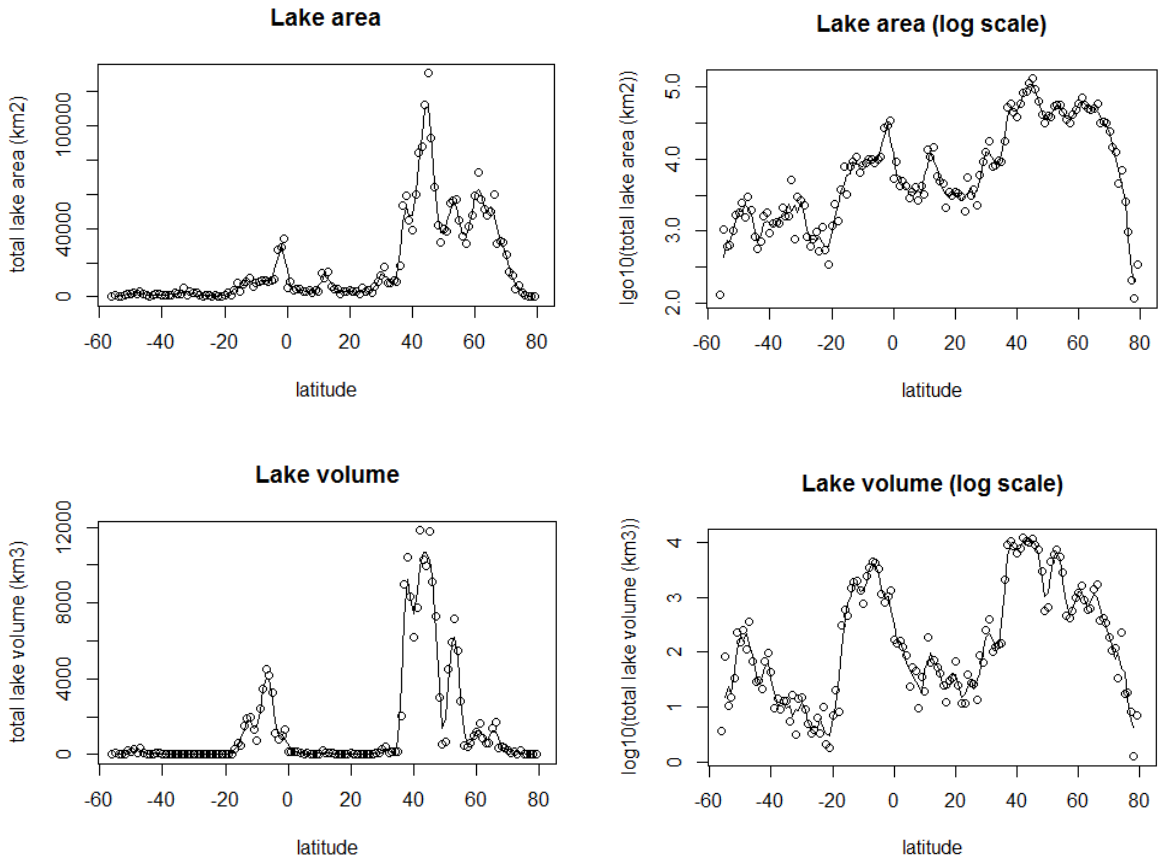

Figure 8. GLWD lake area and predicted total volume by latitude. Lines are 3-degree moving averages.

Table 1. DEM comparison.

DEM	SRTM	ASTER
Elevation measurement method	Interferometric synthetic aperture radar	Stereo pairs of IR/Visible images
Regions not covered	North of 60 degrees latitude	New Zealand and other Pacific islands
Spatial resolution (m)	90	100

Table 2. Correlations with *log(depth)* of differences between pairs of topographic variables, transformed with $f(x) = log(|x|+1)$. (ASTER shown here; SRTM is similar.)

	Max	Mean	Min
Elev	0.62	0.74	0.26
Max		0.58	0.62
Mean			0.69

Table 3. Correlations of predictor variables with residuals of 1-variable model. All other variables had correlations less than 0.15 in magnitude.

DEM used	Area	abs(Latitude)	isGlacial
ASTER	-0.18	0.23	0.21
SRTM	-0.13	0.28	0.25

Table 4. Summary of final models.

DEM used	*meanelev* coef	*isglacial* coef	Intercept	r2	Number of lakes
ASTER	0.82	0.42	-0.99	0.57	328
SRTM	0.83	0.49	-1.06	0.59	319

Table 5. Model fit comparison. RMSE was calculated using leave-one-out cross validation. Everything is comparable between the two DEMs except for the BIC (as they have different numbers of lakes).

ASTER

	2-var model	Best-BIC model
# of vars	2	6
Adj. r2	0.57	0.61
BIC	933	917
Resid SE	0.97	0.92
Xval RMSE	0.98	0.93

SRTM

	2-var model	Best-BIC model
# of vars	2	4
Adj. r2	0.58	0.64
BIC	901	866
Resid SE	0.96	0.90
Xval RMSE	0.97	0.91

References

Heathcote AJ, del Giorgio PA, Prairie YT, Brickman D. 2015. Predicting bathymetric features of lakes from the topography of their surrounding landscape. Canadian Journal of Fisheries and Aquatic Sciences 72:643–650.

Hollister JW, Milstead WB, Urrutia MA. 2011. Predicting Maximum Lake Depth from Surrounding Topography. Schumann GJ-P, editor. PLoS ONE 6:e25764.

Hulley GC, Hook SJ, Abbott E, Malakar N, Islam T, Abrams M. 2015. The ASTER Global Emissivity Dataset (ASTER GED): Mapping Earth's emissivity at 100 meter spatial scale. Geophysical Research Letters 42:7966–7976.

Jarvis A, Reuter HI, Nelson A, Guevara E. 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).

Lehner B, Döll P. 2004. Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296:1–22.

Long Z, Perrie W, Gyakum J, Caya D, Laprise R. 2007. Northern Lake Impacts on Local Seasonal Climate. Journal of Hydrometeorology 8:881–896

Martynov A, Sushama L, Laprise R, Winger K, Dugas B. 2012. Interactive lakes in the Canadian Regional Climate Model, version 5: the role of lakes in the regional climate of North America. Tellus A 64.

Ray N, Adams JM. 2001. A GIS-based Vegetation Map of the World at the Last Glacial Maximum (25,000-15,000 BP). Internet Archaeology 11.

Rouse WR, Oswald CJ, Binyamin J, Spence C, Schertzer WM, Blanken PD, Bussières N, Duguay CR. 2005. The Role of Northern Lakes in a Regional Energy Balance. Journal of Hydrometeorology 6:291–305.

Sobek S, Nisell J, Fölster J. 2011. Predicting the depth and volume of lakes from map-derived parameters. Inland Waters 1:177–184.

Subin ZM, Riley WJ, Mironov D. 2012. An improved lake model for climate simulations: Model structure, evaluation, and sensitivity analyses in CESM1. Journal of Advances in Modeling Earth Systems 4.

Verpoorter C, Kutser T, Seekell DA, Tranvik LJ. 2014. A global inventory of lakes based on high-resolution satellite imagery. Geophysical Research Letters 41:6396–6402.