**Title:** Lake Photic Zone Temperature across the Conterminous United States

*1* US Environmental Protection Agency, Office of Research and Development, Atlantic Coastal Environmental Sciences Division (ACESD), Narragansett, RI 02882

*2* ORISE, Narragansett, RI 02882

**Authors:** Kreakie, B. J. \* *1*, Shivers, S. *2*, Hollister. J. W. *1*, Milstead, W. Bryan. *1*

\* \*corresponding author: [kreakie.betty@epa.gov\*](mailto:kreakie.betty@epa.gov*)

**Abstract:** As the average global air temperature on Earth increases, surface temperatures of lakes are also increasing globally (0.34 °C per decade from 1985 to 2009). The influence of this increased temperature is known to impact lentic ecosystems across local to broad scales. For example, warming lake temperature is linked to disruptions in trophic linkages, changes in thermal stratification, and cyanobacteria bloom dynamics. Thus understanding broad trends in lake temperature is important to understanding the changing ecology of lakes and the potential human health impacts of these changes. To help address this, we developed a simple yet robust model of lake photic zone temperature using the 2007 and 2012 US EPA’s National Lakes Assessment data for the conterminous United States. The final model has a root mean square error of 1.48 and an adjusted R2 of 0.88. The sampling date, that day’s average ambient air temperature, data obtained from the PRISM Climate Group, and longitude are the most important variables impacting the final model’s accuracy. Given the importance of temperature to a lake ecosystem, this model can be a valuable tool for researchers and lake resource managers. Daily predicted lake photic zone temperature for all lakes in the conterminous US can now be estimated based on basic ambient temperature and location information.  
**Keywords:** Random forest, National Lakes Assessment, Limnology, water temperature, warming lakes

# Introduction

During a time of unprecedented environmental and climatic variability, lakes can serve as sentinels and integrators in a changing world (Schindler 2009, Williamson et al. 2009). As the average global air temperature on Earth increases (0.15-0.20 °C per decade since 1975)(Hansen et al. 2010), surface temperatures of lakes are also increasing globally (0.34 °C per decade from 1985 to 2009) (O’Reilly et al. 2015). The influence of this increased temperature touches all biotic and abiotic components of lentic ecosystems.

For example, warming lakes are linked to a disruption in trophic linkages between phytoplankton and zooplankton (Winder and Schindler 2004). Specifically, Winder and Schindler (2004) report a spring diatom bloom occurring 20 days earlier and a long-term decline in Daphnia populations in Lake Washington, a lake near Seattle, Washington. Second, warming has been shown to result in changes in thermal stratification across different lake types [Hampton et al. (2018); O’Reilly et al. (2003); Verburg et al. (2003); Cohen et al. (2016); michelutti2016equatorial]. In particular, lakes that typically stratify can exhibit a strengthening of the stratification as well as a reduction in the depth of the thermocline or previously polymictic lakes can begin to stratify as a result of warming lake temperatures. Temperature, in addition to nutrients, is also a key driver to cyanobacteria bloom dynamics (Paerl and Paul 2012). During periods of higher temperature, cyanobacteria species dominate the phytoplankton community (Peperzak 2003, O’neil et al. 2012, Lürling et al. 2013). As lake temperature increase (typically above 25°C), cyanobacteria have a competitive advantage over phytoplankton and can proliferate quickly (Paerl and Huisman 2008). Moreover, experimentally enhanced water temperatures yielded significantly increased growth rates of higher toxic Microcystis, but not the non-toxic strains (Davis et al. 2009). Thus, our ability to understand and predict toxic cyanobacteria blooms will be deeply dependent on our ability to forecast lake photic zone temperature. Ultimately, temperature changes will greatly impact every aspect of lake ecology and lake resource management.

Because of the ecological significance, it is not surprising that modelling near-surface lake temperature has been broadly investigated. Models typically vary in number of lakes studied, complexity of modelling approach, and study interval. Modelling efforts include efforts to model a single lake over relatively small (e.g. hourly) time intervals (Peeters et al. 2002, Saeed et al. 2016, Zhong et al. 2016). Additionally, there are numerous studies that model temperature for a small number of lakes while attempting to limit the number of predictor variables (Matuszek and Shuter 1996, Kettle et al. 2004 p. @livingstone1998relationship, Toffolon et al. 2014, Piccolroaz 2016). In these efforts, air temperature and lake size are often the only selected predictor variables. There have also been some efforts aimed at modelling and measuring lake temperature across large spatial extents for large number of lakes (O’Reilly et al. 2015, Minns et al. 2017, Wan et al. 2017). These past efforts use a wide variety of approaches and have been rather successful at modelling lake photic zone temperature. One limitation, however, is that the majority of these lakes are large and few studies document or model smaller lakes. Considering that the number and importance of small lakes (< 1km2) has been historically underestimated, it is important to also predict temperatures in these smaller waterbodies (Downing et al. 2006, Winslow et al. 2014).

In spite of the need to understand near-surface temperature for all lakes across a large spatial extent, there have been several challenges that have slowed our progress. Modelling lake temperature requires a large amount of data, and therefore study lakes are often selected opportunistically, which may introduce a spatial bias. Frequently, the study lakes have a high regional resource value (e.g. the Great Lakes) and commonly have an extensive monitoring history due to the vested interest of the public. When model efforts do attempt to cross large spatial extents, these efforts often rely on satellite data. While these models predict over large areas, they are restricted by the size of the lake captured by satellite (typically 3km2 for 1 km MODIS pixels).

Our modelling effort takes advantage of the relatively recent availability of broad scale field data for lakes and uses the US Environmental Protection Agency’s (EPA’s) National Lake Assessment (NLA). The NLA is a stratified random sample of all lakes in the conterminous United States repeated every five years beginning in 2007. Even though this is a large effort involving numerous agencies, the sampling methods are standardized and have a comprehensive quality assurance plan. The uniqueness of this data set allows us to build a robust lake photic zone temperature model for all US lakes. By using lakes across the US (i.e. at a large spatial extent), we included lakes with different morphologies, in different climates with diverse geologies, and surrounding landscapes. Additionally, small lakes were well represented with more than 50% of the lakes in the 2012 survey less than 0.5 km2.

The main goals of this work are to 1) develop a simple yet robust lake photic zone temperature model and 2) develop a model that is applicable to all lakes in the conterminous US that capture key drivers of near-surface lake temperature. Additionally, it is our practice conduct our work as openly as is feasible. Towards that end we provide access to all code and data used to develop these models with the active repository available at <https://github.com/usepa/lake_photic_zone> and the code, data, and documents are archived at <https://link_to_zenodo_right_before_submission>.

# Methods

## Data

We relied on the in situ temperature data provided in the US EPA’s National Lake Assessment (U. S. Environmental Protection Agency 2009, 2016). The NLA is a stratified random sample of lakes (great than 1 ha) across the United States. NLA sampling took place in 2007, 2012, and 2017. This research effort used the 2007 and 2012 sample years. The 2017 data are currently undergoing quality control before being released the public. Upon release, the data can easily be included to improve this model. For both sample years, we have included over 1,000 lakes across the conterminous US excluding the Great Lakes (Figure 1). For each sampled lake, we used the mean temperature for all sampled depths of less than 2m (this being the depth typically considered the photic zone).

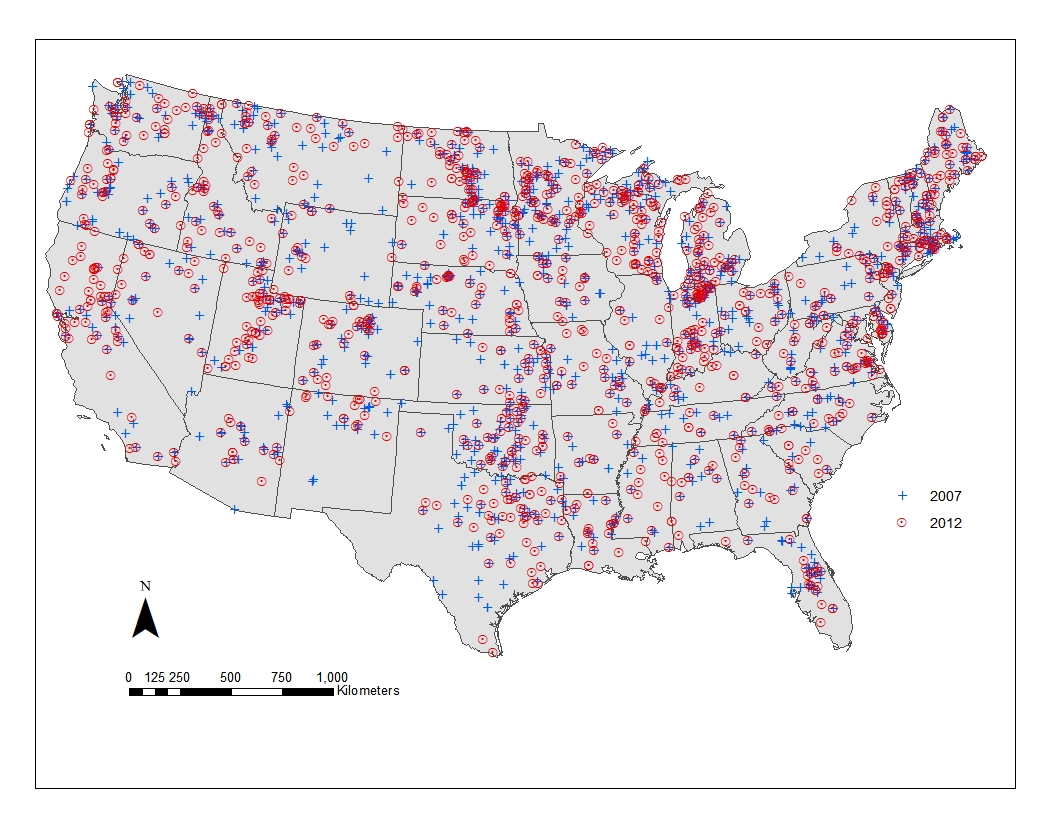


Figure 1: Map of 2007 and 2012 US EPA’s National Lake Assessment (NLA) Lakes

We included numerous predictor variables that are hypothesized to impact lake photic zone. As a proxy for directly measured ambient air temperature, we used the PRISM AN81d dataset (PRISM Climate Group, Oregon State University, <http://prism.oregonstate.edu>, created 07 Nov 2018). This dataset provides interpolated daily temperature estimates (mean, maximum, and minimum) for 4km grids in the conterminous United States for 1981 to the present (see: <http://www.prism.oregonstate.edu/documents/PRISM_datasets.pdf>). PRISM takes advantage of measured climate variables to interpolate point data to spatially defined grids using regression techniques and expert knowledge (Daly et al. 2008). For our study, we used the “prism” R package to download the mean daily temperatures for the PRISM grid cells corresponding to the centroids of all NLA lakes included in this study for dates between 01 May and 30 September of 2007 and 2012 (Hart and Bell 2015).

In addition to climate, there are additional factors clearly impacting lake temperature (e.g., surrounding land use, lake depth, size, and configuration, and elevation). To test for the relative importance of lake morphometry and surrounding landscape, we used the R packages lakemorpho to calculate a suite of lake morphometry metric and elevatr to access digital elevation models for each lake (Hollister and Stachelek 2017, Hollister and Tarak Shah 2017). Additionally, the National Land Cover Database (NLCD) was our source for land cover data. Specifically, we calculated the percent impervious surface of a 3000m buffer for each lake. Lakes with partials buffers falling outside the US were excluded.

To test the relative influence of both short and long-term temperature, we derived several measures for a lake’s local air temperature. Mean air temperatures for day of and the day before the sample date were extracted directly from the prism data. To understand longer term influences we calculated average mean air temperatures for periods 3, 7, and 30 days prior to the sample date.

## Random Forest Modelling

Random forest modelling was used to not only develop a predictive model of photic zone temperature, but also used to aid in variable selection and to calculate relative variable importance. Random forest is a machine learning method that builds a consensus prediction from the assemblage of multiple tree models (here specifically 10,000 trees for the final model and 1,000 trees for the variable selection models). Each individual tree model is constructed from a subset of the full data set and a subset of all predictor variables. All random forest modelling was conducted in R, version 4.0.2 with the randomForest package (Liaw and Wiener 2002, R Core Team 2020). Model performance is reported as mean square error and adjusted R2.

Random forest does not require that users reduce the number of predictor variables because the random forest algorithm prevents overfitting and is not impacted by correlated predictor variables (Cutler et al. 2007). It is unlikely that random forest models constructed with reduced numbers of predictor variables perform any better than models constructed with a full suite of available variables (Fox et al. 2017). However, reducing the number of predictor variables eases interpretation and can reduce potentially unneeded computation time while also not severely impacting prediction accuracy. Several of our climatic predictor variables are computationally intensive to create for the entire conterminous United States. So, in order to use our final model for future forecasting or historical backcasting, we strove to create a robust predictive model while minimizing computational demands. To determine the optimal number and set of variables, we followed the variable selection method presented in (Hollister et al. 2016).

In addition to a measure of overall model performance, we used percent increase of mean-squared error to assess variable importance. The percent increase in mean-squared error is a comparison between the true values of a variable and randomly permuted values of a specific variable on overall model performance.

# Results

We used 1185 data points from the 2007 NLA and 1097 from 2012 NLA and these points are a spatially representative sample of lakes in the United States (Figure 1)

Using the average temperature for the upper two meters as the response variable, we initially began the variable selection process with 16 predictor variables. The variable selection process identified a reduced model with ~8 variables showing minimal model error (Figure 2). The selected variables were average ambient air temperature for the sample date, sample date, longitude, average ambient air temperature for 30 day proceeding the sample date, elevation, latitude, length of lake shoreline, and the lake surface area.

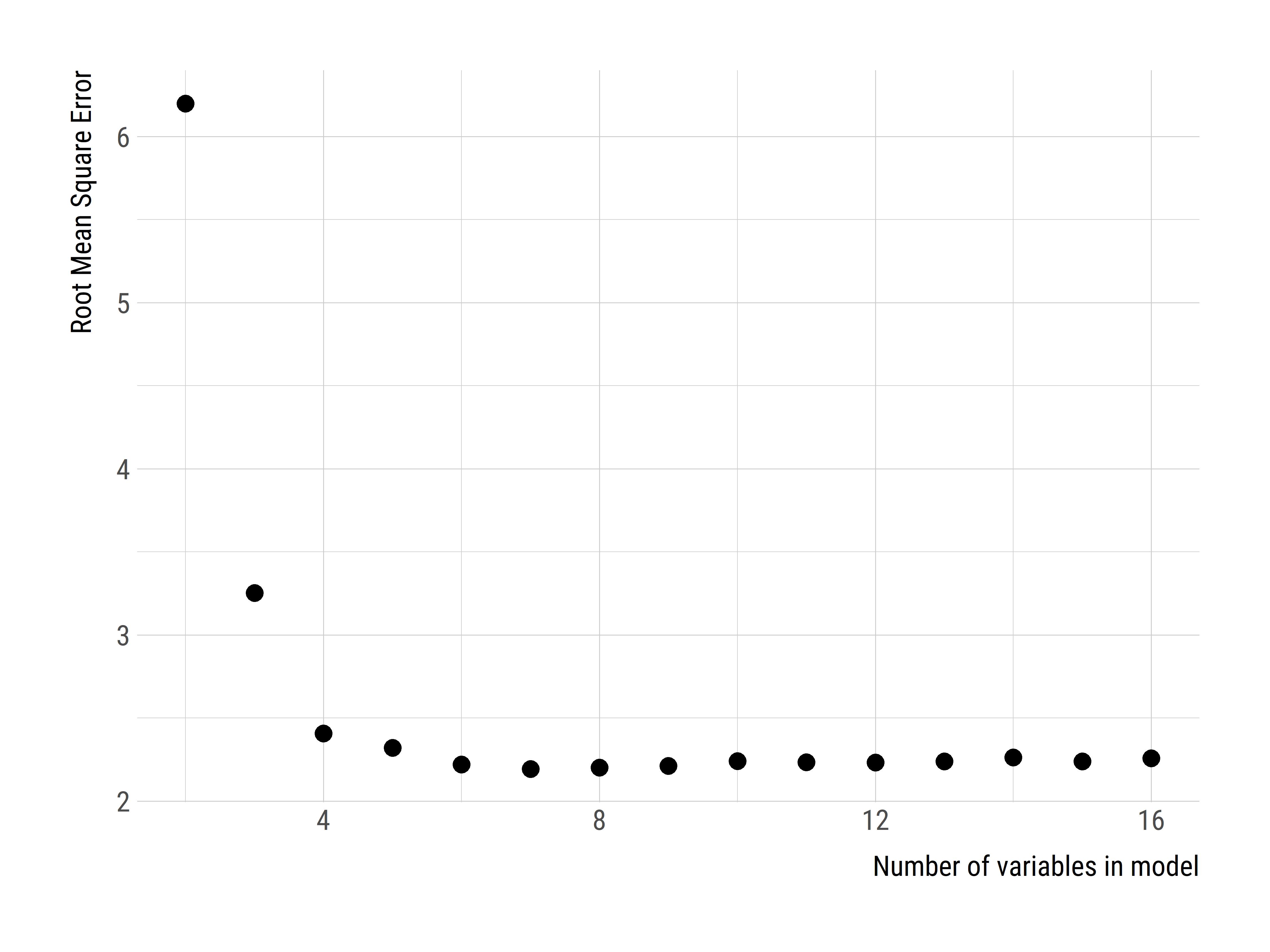


Figure 2: Variable selection plot for all variables. Shows percent increase in mean square error as a function of the number of variables.

The final model built with the eight selected variables has a root mean-squared error of 1.48 and adjusted R2 of 0.88. The model performs well across a wide range of temperatures (Figure 3); however, at the higher and lower temperature (i.e. less than 15 °C and greater than 32 °C) the model does not perform as well but lakes with growing season temperatures at these extremes were rare and represented only 3.86 of lakes sampled in the 2007 and 2012 NLA.

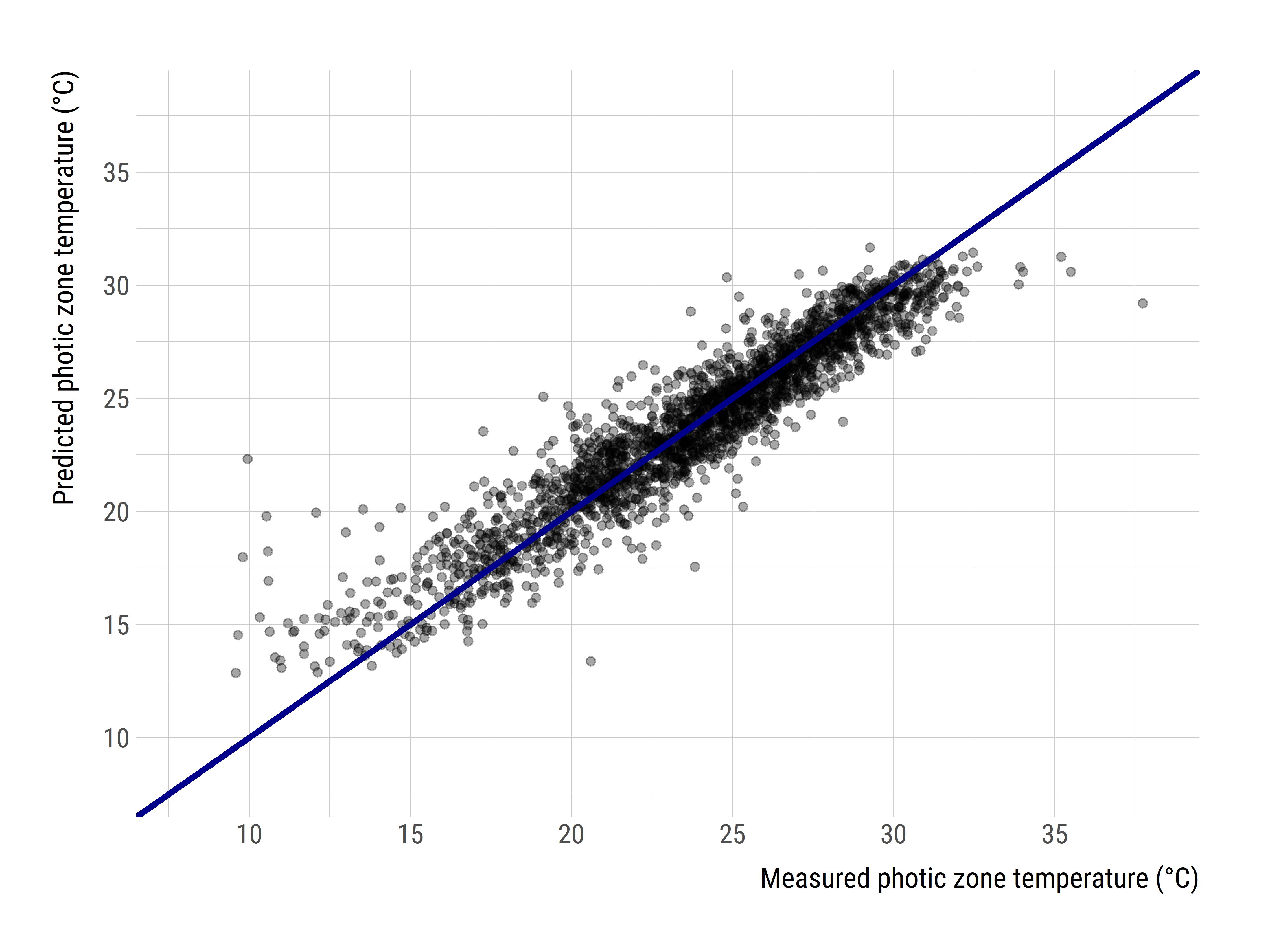


Figure 3: Predicted versus measured photic zone temperature

The variables ranked in order of importance were date, average temperature, longitude, 30-day average temperature, elevation, latitude, surface area, and shoreline length (Figure 4). The partial dependency plots illustrate how the predicted photic zone temperature changes over the range of values for all predictor variables (Figure 5).

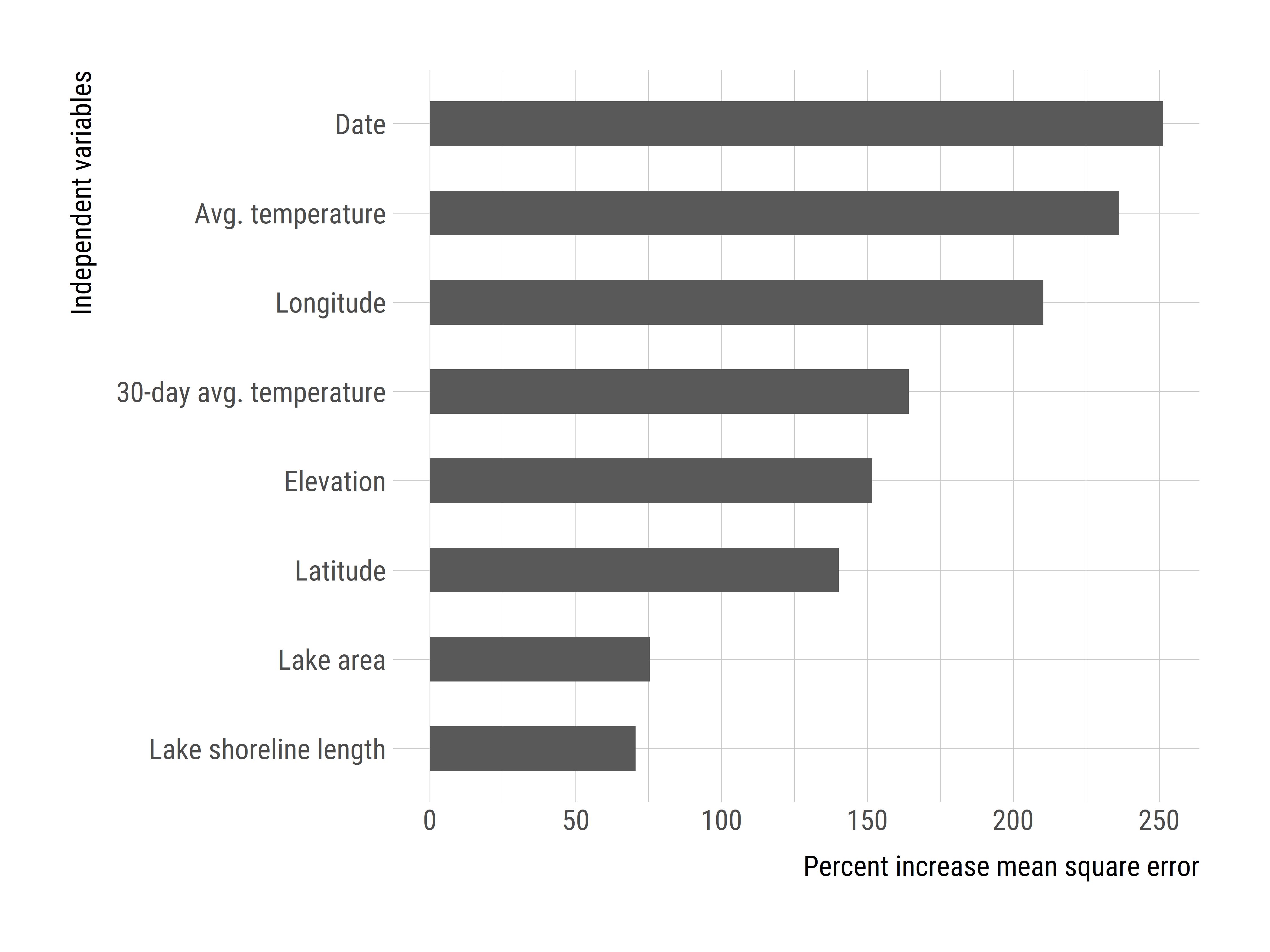


Figure 4: Variable importance plot for selected variables. Shows precent increase in mean square error. Higher values indicate a higher impact on overall model accuracy.

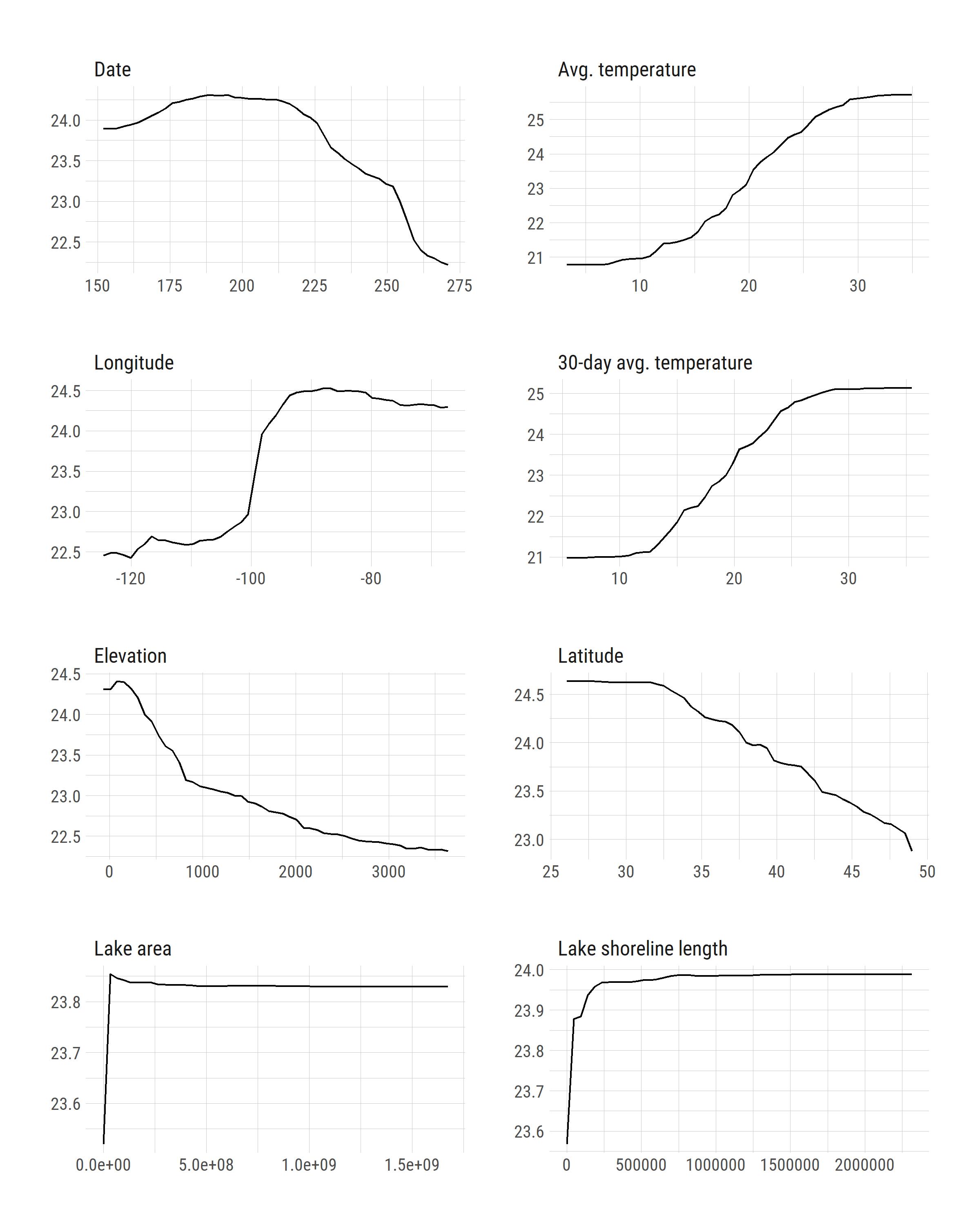


Figure 5: Partial dependence plots for selected variables.

# Discussion and conclusions

Here we present a simple yet robust model of lake photic zone temperature using the 2007 and 2012 NLA data for the conterminous United States. The final model has a root mean square error of 1.48 and an adjusted R2 of 0.88. The sampling date, that day’s average ambient air temperature, data obtained from the PRISM Climate Group, and longitude are the most important variables impacting the final model’s accuracy. Given the importance of temperature to a lake ecosystem, especially to cyanobacteria bloom dynamics, this model can be a valuable tool for researchers and lake resource managers. Daily predicted lake photic zone temperature for all lakes in the conterminous US can now be estimated based on basic ambient temperature and location information.

Using only easily obtainable metrics to estimate lake temperature was a primary focus in the creation of this model. Several studies (e.g. (Piccolroaz et al. 2018)) have used similar metrics with good results (RMSE = 0.5 °C) to estimate lake temperature but were applied to a single, or a small number of lakes. Other studies have used more complex models (e.g. General Lake Model (GLM)) to yield good results (RMSE = 1.62 °C for the epilimnion) at varying scales (Bruce et al. 2018, Hipsey et al. 2019). However, these models require more complex input variables which are not available for all lakes and may be biased towards large lakes. Using the NLA data and simple metrics allowed this model to be applicable to any lake within the conterminous United States using minimal input data with a comparable RMSE.

The final model included, average ambient air temperature and the average temperature of the prior 30 days. Yet the 3-day and 7-day prior to the sampling date averages were not selected in the final model. It is likely that the three and seven day averages are not providing the model with unique information as the three and seven day averages are likely to be similar to the day-of average. Ambient air temperature does not change drastically over such a short time scale [CITATION??]. Even if a big swing in temperature did happen during this short time period, it would be too rare to be a significant factor in the model. Yet the 30-day average does have an impact on lake photic zone temperature as it provides information about the longer term temperature intensity leading up to the sample date. Not just how far along into summer; that information would be captured in the date parameter. But more specifically, this measures the long term thermal heating happening at a site.

In addition to the several derived air temperature variables, we included land-use/land cover variables in our initial variable selection process. Specifically, we calculated the percent impervious surface for a 3km lake buffer and a measure of shoreline development. These variables were included based on the hypothesis that higher amounts of development and therefore impervious surface surrounding a lake would lead to higher temperatures in lakes. Yet neither of these variables were selected in the final model. Even though the land-use variables were not selected that does not mean development and impervious are not impactful. This urban-heat effect on lakes may have been adequately captured in the average ambient air temperature. Therefore, making the land-use variables redundant. Regardless these variables did not independently contribute to the model’s accuracy.

Despite being one of the most common measurements collected by limnologists, lake temperature datasets that cover long periods of time are very difficult to obtain. Sharma et al (2015) have compiled summer lake temperature data for 291 lakes for the period 1985-2009. This may be the largest lake temperature database to date; however, the data are only available to members of their research group and, realistically, the number of lakes included is very small. One of the reasons we chose to model lake photic zone temperature was to develop a database of lake temperatures for the 48 conterminous United States. The model we present has proven to be accurate and will allow us to backcast lake temperatures for all the > 300,000 lakes included in NHDplus for the period of time covered by the PRISM climate predictions (1981 to present). This dataset will allow us to investigate how photic zone temperatures vary both spatially and temporally across the United States. This database is being developed and, when complete, will be made available as an open source data set.

# Bibliography

Bruce, L. C., M. A. Frassl, G. B. Arhonditsis, G. Gal, D. P. Hamilton, P. C. Hanson, A. L. Hetherington, J. M. Melack, J. S. Read, K. Rinke, and others. 2018. A multi-lake comparative analysis of the general lake model (glm): Stress-testing across a global observatory network. Environmental Modelling & Software 102:274–291.

Cohen, A. S., E. L. Gergurich, B. M. Kraemer, M. M. McGlue, P. B. McIntyre, J. M. Russell, J. D. Simmons, and P. W. Swarzenski. 2016. Climate warming reduces fish production and benthic habitat in lake tanganyika, one of the most biodiverse freshwater ecosystems. Proceedings of the National Academy of Sciences 113:9563–9568.

Cutler, D. R., T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. International Journal of Climatology: a Journal of the Royal Meteorological Society 28:2031–2064.

Davis, T. W., D. L. Berry, G. L. Boyer, and C. J. Gobler. 2009. The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of microcystis during cyanobacteria blooms. Harmful algae 8:715–725.

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

Fox, E. W., R. A. Hill, S. G. Leibowitz, A. R. Olsen, D. J. Thornbrugh, and M. H. Weber. 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental monitoring and assessment 189:316.

Hampton, S. E., S. McGowan, T. Ozersky, S. G. P. Virdis, T. T. Vu, T. L. Spanbauer, B. M. Kraemer, G. Swann, A. W. Mackay, S. M. Powers, M. F. Meyer, S. G. Labou, C. M. O’Reilly, M. DiCarlo, A. W. E. Galloway, and S. C. Fritz. 2018. Recent ecological change in ancient lakes. Limnology and Oceanography 63:2277–2304.

Hansen, J., R. Ruedy, M. Sato, and K. Lo. 2010. Global surface temperature change. Reviews of Geophysics 48.

Hart, E. M., and K. Bell. 2015. Prism: Download data from the oregon prism project.

Hipsey, M. R., L. C. Bruce, C. Boon, B. Busch, C. C. Carey, D. P. Hamilton, P. C. Hanson, J. S. Read, E. De Sousa, M. Weber, and others. 2019. A general lake model (glm 3.0) for linking with high-frequency sensor data from the global lake ecological observatory network (gleon).

Hollister, J., and J. Stachelek. 2017. Lakemorpho: Calculating lake morphometry metrics in r. F1000Research 6.

Hollister, J., and Tarak Shah. 2017. Elevatr: Access elevation data from various apis.

Hollister, J. W., W. B. Milstead, and B. J. Kreakie. 2016. Modeling lake trophic state: A random forest approach. Ecosphere 7:e01321.

Kettle, H., R. Thompson, N. J. Anderson, and D. M. Livingstone. 2004. Empirical modeling of summer lake surface temperatures in southwest greenland. Limnology and Oceanography 49:271–282.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Livingstone, D. M., and A. F. Lotter. 1998. The relationship between air and water temperatures in lakes of the swiss plateau: A case study with palsgmaelig; olimnological implications. Journal of Paleolimnology 19:181–198.

Lürling, M., F. Eshetu, E. J. Faassen, S. Kosten, and V. L. Huszar. 2013. Comparison of cyanobacterial and green algal growth rates at different temperatures. Freshwater Biology 58:552–559.

Matuszek, J. E., and B. J. Shuter. 1996. An empirical method for the prediction of daily water temperatures in the littoral zone of temperate lakes. Transactions of the American Fisheries Society 125:622–627.

Minns, C. K., B. J. Shuter, A. Davidson, and S. Wang. 2017. Factors influencing peak summer surface water temperature in canada’s large lakes. Canadian Journal of Fisheries and Aquatic Sciences 75:1005–1018.

O’neil, J., T. Davis, M. Burford, and C. Gobler. 2012. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. Harmful algae 14:313–334.

O’Reilly, C. M., S. R. Alin, P.-D. Plisnier, A. S. Cohen, and B. A. McKee. 2003. Climate change decreases aquatic ecosystem productivity of lake tanganyika, africa. Nature 424:766–768.

O’Reilly, C. M., S. Sharma, D. K. Gray, S. E. Hampton, J. S. Read, R. J. Rowley, P. Schneider, J. D. Lenters, P. B. McIntyre, B. M. Kraemer, and others. 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophysical Research Letters 42:10–773.

Paerl, H. W., and J. Huisman. 2008. Blooms like it hot. Science 320:57–58.

Paerl, H. W., and V. J. Paul. 2012. Climate change: Links to global expansion of harmful cyanobacteria. Water research 46:1349–1363.

Peeters, F., D. M. Livingstone, G.-H. Goudsmit, R. Kipfer, and R. Forster. 2002. Modeling 50 years of historical temperature profiles in a large central european lake. Limnology and Oceanography 47:186–197.

Peperzak, L. 2003. Climate change and harmful algal blooms in the north sea. Acta Oecologica 24:S139–S144.

Piccolroaz, S. 2016. Prediction of lake surface temperature using the air2water model: Guidelines, challenges, and future perspectives. Advances in Oceanography and Limnology.

Piccolroaz, S., N. Healey, J. Lenters, S. Schladow, S. Hook, G. Sahoo, and M. Toffolon. 2018. On the predictability of lake surface temperature using air temperature in a changing climate: A case study for lake tahoe (usa). Limnology and Oceanography 63:243–261.

R Core Team. 2020. R: A language and environment for statistical computing, version 4.0.2. R Foundation for Statistical Computing, Vienna, Austria.

Saeed, S., K. Honeyeh, K. Ozgur, and L. Wen-Cheng. 2016. Water temperature prediction in a subtropical subalpine lake using soft computing techniques. Earth Sciences Research Journal 20:1–11.

Schindler, D. 2009. Lakes as sentinels and integrators for the effects of climate change on watersheds, airsheds, and landscapes. Limnology and Oceanography 54:2349–2358.

Sharma, S., D. K. Gray, J. S. Read, C. M. O’Reilly, P. Schneider, A. Qudrat, C. Gries, S. Stefanoff, S. E. Hampton, S. Hook, and others. 2015. A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. Scientific data 2:150008.

Toffolon, M., S. Piccolroaz, B. Majone, A.-M. Soja, F. Peeters, M. Schmid, and A. Wüest. 2014. Prediction of surface temperature in lakes with different morphology using air temperature. Limnology and Oceanography 59:2185–2202.

U. S. Environmental Protection Agency. 2009. National lakes assessment: A collaborative survey of the nation’s lakes. EPA 841-R-09-001. Office of Water and Office of Research and Development, US Environmental Protection Agency, Washington, DC.

U. S. Environmental Protection Agency. 2016. National lakes assessment 2012: A collaborative survey of the nation’s lakes. EPA 841-R-16-113. Office of Water and Office of Research and Development, US Environmental Protection Agency, Washington, DC.

Verburg, P., R. E. Hecky, and H. Kling. 2003. Ecological consequences of a century of warming in lake tanganyika. Science 301:505–507.

Wan, W., H. Li, H. Xie, Y. Hong, D. Long, L. Zhao, Z. Han, Y. Cui, B. Liu, C. Wang, and others. 2017. A comprehensive data set of lake surface water temperature over the tibetan plateau derived from modis lst products 2001–2015. Scientific data 4:170095.

Williamson, C. E., J. E. Saros, W. F. Vincent, and J. P. Smol. 2009. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. Limnology and Oceanography 54:2273–2282.

Winder, M., and D. E. Schindler. 2004. Climate change uncouples trophic interactions in an aquatic ecosystem. Ecology 85:2100–2106.

Winslow, L. A., J. S. Read, P. C. Hanson, and E. H. Stanley. 2014. Lake shoreline in the contiguous united states: Quantity, distribution and sensitivity to observation resolution. Freshwater biology 59:213–223.

Zhong, Y., M. Notaro, S. J. Vavrus, and M. J. Foster. 2016. Recent accelerated warming of the laurentian great lakes: Physical drivers. Limnology and Oceanography 61:1762–1786.