# EDI Metadata Template (2019)[[1]](#footnote-1)

Data should be in csv text file. If starting with an Excel spreadsheet, please make sure it does not contain any formulas and comments on cells. If you need comments put them in their own column. If data were used in a database and major table linking is necessary to analyze, please de-normalize into a flat file, not just database table exports.

## Dataset Title

Lake chloride concentrations and model predictions for 49,432 lakes in the Midwest and Northeast United States.

## Short name or nickname you use to refer to this dataset:

LAGOS chloride predictions

## Abstract

Lakes in the Midwest and Northeast United States are at risk of anthropogenic chloride contamination, but we have little knowledge of the prevalence and spatial distribution of the problem. The majority of salt pollution in north temperate regions stems from road salt application but other chloride sources include water softeners, synthetic fertilizers, and livestock excretion. Although chloride contamination of lakes is well documented, it is unknown how many lakes are at risk of long-term salinization. We used a quantile regression forest to leverage information from 2,773 lakes to predict the chloride concentration of all 49,432 lakes greater than 4 ha in a 17-state area. The QRF used 22 predictor variables, which included lake morphometry characteristics, watershed land use, and distance to the nearest interstate and road. Model predictions had an r2 of 0.94 for all chloride observations, and 0.87 for predictions of the mean chloride concentration observed at each lake.

## Investigators

(list in order as for a paper with e-mail addresses, organization and preferably ORCID ID, if you don’t have one, get it, it’s easy and free: <http://orcid.org/>) add table rows as needed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) |
| Hilary | A | Dugan | University of Wisconsin-Madison | hdugan@wisc.edu | 0000-0003-4674-1149 |
| Nicholas | K | Skaff | University of California, Berkeley University | nskaff@berkeley.edu | 0000-0002-5929-3966 |
| Jonathan | P | Doubek | Lake Superior State University | jdoubek@lssu.edu | 0000-0003-2651-4715 |
| Samantha | M | Burke | University of Guelph | samantha.burke2@gmail.com | 000-0002-7941-2096 |
| Flora | E | Krivak-Tetley | Dartmouth College | fkt.gr@dartmouth.edu | 0000-0003-3521-2460 |
| Jamie | C | Summers |  | jamiecsummers@gmail.com | 0000-0002-7497-2326 |

## Other personnel names and roles

(dataset creators & contact, field crew, data entry etc. with e-mail addresses, organization and ORCID ID)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) | Role in project |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## License

(Select a license for release of your data. We have 2 recommendations: [CCO – most accommodating of data reuse](https://creativecommons.org/publicdomain/zero/1.0/), & [CCBY – requires attribution](https://creativecommons.org/licenses/by/4.0/))

CCO

## Keywords

(List keywords and separate with commas. Using keywords from a controlled vocabulary (CV) will improve the future discovery and reuse of your data. The LTER CV is effective at describing ecological and environmental data. [Access the LTER CV here](http://vocab.lternet.edu/vocab/vocab/index.php). [Try this text mining service to extract LTER CV keywords from your abstract or methods](http://vocab.lternet.edu/keywordDistiller/). Additionally, please determine one or two keywords that best describe your lab, station, and/or project (e.g., Trout Lake Station, NTL LTER). This will help others discover your data by site/project).

Chloride, lakes, reservoirs, LAGOS, limnology, road salt, salt, impervious surface, salinization

## Funding of this work:

Add rows to table if several grants were involved, list only the main PI, start with main grant first:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PI First Name | PI Middle Initial | PI Last Name | PI ORCID ID (optional) | Title of Grant | Funding Agency | Funding Identification Number |
| Kathleen | C | Weathers | 0000-0002-3575-6508 | **Collaborative Research: Building Analytical, Synthesis, and Human Network Skills Needed for Macrosystem Science: a Next Generation Graduate Student Training Model Based on GLEON** | NSF | EF-1137327 and EF-1702991 |

## Timeframe

* Begin date **1990-01-01**
* End date **2018-12-13**
* Data collection ongoing/completed **Completed**

## Geographic location

* Verbal description: **Midwest and Northeast USA**
* North bounding coordinates (decimals) 49.42
* South bounding coordinates (decimals) 36.56
* East bounding coordinates (decimals) -68.19
* West bounding coordinates (decimals) -96.73

## Taxonomic species or groups

## Methods

(please be specific, include instrument descriptions, or point to a protocol online, if this is a data compilation please specify datasets used, preferably their DOI or URL plus general citation information)

We leveraged publicly available land use, lake catchment and morphometry, and climate data across a 17-state area of the Midwest and Northeast United States, to predict chloride concentrations in 49,432 lakes. Our general methodology included: 1) Acquiring and geoprocessing lake water quality data and site characteristics. 2) Harmonizing training datasets. 3) Building a machine learning model for chloride prediction. Calculating model fit. 4) Building a prediction dataset for 49,432 lakes.

#### Training Dataset

Observational chloride measurements from lakes, reservoirs, and impoundments were downloaded from the US water quality portal (WQP). All results were converted to mg L-1, and only data with *ResultStatusIdentifier* as ‘Accepted’ or ‘Final’ noted in the dataset were retained. The initial search of 115,389 observations was then filtered to data collected after 1990, chloride concentrations < 10,000 mg L-1, and water samples less than 10 m deep or with depth not listed (where the assumption was an epilimnion measurement). These quality control steps were taken to limit inclusion of historical data that may not represent current conditions, remove naturally saline waterbodies (n =5, adjacent/connected to the Atlantic Ocean), and remove potentially meromictic lakes (n=0). Multiple observations collected on the same day were averaged. Lakes with missing watershed information were removed, resulting in 29,675 unique daily observations from 2,773 lakes. Three states (Illinois, Iowa, and Rhode Island) had no chloride data, and three states (Pennsylvania, Connecticut, and New Hampshire) had chloride data from only one lake. 2,773 lakes represent 5% of the region’s lakes.

WQP site identification numbers (IDs) from the dataset were linked to the high-resolution National Hydrography Dataset (NHD) that accessed bounding box information of each NHD shapefile and ran a spatial join. The resulting relational table linked each chloride observation to an individual lake through an NHD ID. For every NHD lake ID, geospatial lake data were obtained from the LAGOS-NE database (Soranno et al. 2017), which provides watershed ecological context for all lakes greater than 4 ha in the 17-state area. Additional site characteristics were extracted from GIS line files of US interstates, US primary roads, and gridded winter severity data. Across all predictor variables in the training dataset, minimum values were >= 0.01. After converting zero values to 0.001, all data were log-transformed.

#### Machine Learning Model

A quantile regression forest (QRF) was used to model the relationship between observed chloride concentrations and lake and watershed characteristics. This model was chosen to accommodate a large number of correlated predictor variables, the presence of non-linear responses, and the potential importance of interactions among predictor variables. The QRF was implemented with 1,000 trees using the ranger package in R, with *mtry* set to 4 (Wright and Ziegler 2017).

To avoid overfitting the QRF to lakes with a greater number of chloride observations, we developed a customized sampling routine that constructed individual trees using the observations from a random subset of the study lakes (95% subset: the ‘in-bag samples’). Each resulting tree was used to make out-of-bag predictions on the remaining observations from the 5% of excluded lakes. All predictions are reported as the median of the terminal node values from each tree, with the corresponding 90%-prediction interval calculated from the .05 and 0.95 quantiles of the estimated conditional distribution of the response variable (Meinshausen 2006). Median terminal node values were chosen over mean values because they had superior predictive performance on out-of-bag observations.

#### Prediction Dataset

A prediction dataset was constructed for the full LAGOS-NE dataset, which contained 51,102 lakes and reservoirs greater than 4 ha in the 17-state area. After removing lakes with no available land-use data because the watersheds crossed the US/Canada border, 49,432 lakes remained, of which 2,773 were used for training the model. The prediction dataset was identical in structure to the training dataset, but contained no observational chloride data.

Meinshausen, N. 2006. Quantile Regression Forests. Journal of Machine Learning Research 7:983–999.

Soranno, P. A., L. C. Bacon, M. Beauchene, K. E. Bednar, E. G. Bissell, C. K. Boudreau, M. G. Boyer, M. T. Bremigan, S. R. Carpenter, J. W. Carr, K. S. Cheruvelil, S. T. Christel, M. Claucherty, S. M. Collins, J. D. Conroy, J. A. Downing, J. Dukett, C. E. Fergus, C. T. Filstrup, C. Funk, M. J. Gonzalez, L. T. Green, C. Gries, J. D. Halfman, S. K. Hamilton, P. C. Hanson, E. N. Henry, E. M. Herron, C. Hockings, J. R. Jackson, K. Jacobson-Hedin, L. L. Janus, W. W. Jones, J. R. Jones, C. M. Keson, K. B. S. King, S. A. Kishbaugh, J.-F. Lapierre, B. Lathrop, J. A. Latimore, Y. Lee, N. R. Lottig, J. A. Lynch, L. J. Matthews, W. H. McDowell, K. E. B. Moore, B. P. Neff, S. J. Nelson, S. K. Oliver, M. L. Pace, D. C. Pierson, A. C. Poisson, A. I. Pollard, D. M. Post, P. O. Reyes, D. O. Rosenberry, K. M. Roy, L. G. Rudstam, O. Sarnelle, N. J. Schuldt, C. E. Scott, N. K. Skaff, N. J. Smith, N. R. Spinelli, J. J. Stachelek, E. H. Stanley, J. L. Stoddard, S. B. Stopyak, C. A. Stow, J. M. Tallant, P.-N. Tan, A. P. Thorpe, M. J. Vanni, T. Wagner, G. Watkins, K. C. Weathers, K. E. Webster, J. D. White, M. K. Wilmes, and S. Yuan. 2017. LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. GigaScience 6:1–22.

Wright, M. N., and A. Ziegler. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software:1–17.

## Data Table

* Column name: exactly as it appears in the dataset. Please avoid special characters, dashes and spaces.
* Description: please be specific, it can be lengthy
* Unit: please avoid special characters and describe units in this pattern: e.g. microSiemenPerCentimeter, microgramsPerLiter, absoptionPerMolePerCentimeter
* Code explanation: if you use codes in your column, please explain in this way: e.g. LR=Little Rock Lake, A=Sample suspect, J=Nonstandard routine followed
* Data format: please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy hh:mm:ss plus the time zone and whether or not daylight savings was observed.
* If a code for ‘no data’ is used, please specify: e.g. -99999

Please add rows as needed

**Table description:** Training dataset for QRF model, with 29,010 observed chloride concentrations

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Empty value code |
| lagoslakeid | Unique lake identifier developed for LAGOS-NE | string |  |
| nhdid | Unique lake identifier from National Hydrography dataset | string |  |
| gnis\_name | Lake Name | string | NA |
| nhd\_lat | Latitude | Deg |  |
| nhd\_long | Longitude | Deg |  |
| MaxDepth | Maximum depth of lake | Meters | NA |
| state\_name | Name of US state that lake is located in (or partially in) | string |  |
| ActivityStartDate | Date of sampling | YYYY-mm-dd |  |
| Chloride | Chloride concentration | mg L-1 |  |
| Month | Month of sampling | 1-12 |  |
| LakeArea | Surface area of the lake | hectares |  |
| WS.Area | Surface area of the watershed | hectares |  |
| WinterSeverity | Winter severity index obtained from ClearRoads (national research consortium, clearroads.org). Calculated from 2000 to 2010 as 0.50 × (average annual snowfall in inches) + 0.05 × (annual duration of snowfall in hours) + 0.05 × (annual duration of blowing snow in hours) + 0.10 × (annual duration of freezing rain in hours). | Unitless.  Range 4.9-185 |  |
| WS.OpenWater | % landuse classified as open water in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Dev.Open | % landuse classified as open space, developed in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Dev.Low | % landuse classified as developed, low intensity in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Dev.Med | % landuse classified as developed, medium intensity in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Dev.High | % landuse classified as developed, high intensity in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Barren | % landuse classified as barren/transitional in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.DeciduousForest | % landuse classified as deciduous forest in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.EvergreenForest | % landuse classified as evergreen forest in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.MixedForest | % landuse classified as mixed forest in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Schrub | % landuse classified as schrubland in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Grassland | % landuse classified as grassland in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.PastureHay | % landuse classified as pasture/hay in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.Crops | % landuse classified as row crops in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.WoodyWetlands | % landuse classified as woody wetlands in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.EmergentWetlands | % landuse classified as herbaceous wetlands in the watershed. Derived from the National Land Cover Dataset (NLCD). | percent |  |
| WS.RoadDensity | Road density in the watershed. Derived from the National Land Cover Dataset (NLCD). | Meters per hectare |  |
| InterstateDistance | Distance to the nearest interstate | Meter |  |
| RoadDistance | Distance to the nearest road | Meter |  |
| pred.05 | Prediction interval: 0.05 quantile | mg L-1 |  |
| pred.50 | Median prediction | mg L-1 |  |
| pred.95 | Prediction interval: 0.95 quantile | mg L-1 |  |

**Table description:** Lake chloride predictions for 49,432 lakes

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Empty value code |
| lagoslakeid | Unique lake identifier developed for LAGOS-NE | string |  |
| nhdid | Unique lake identifier from National Hydrography dataset | string |  |
| gnis\_name | Lake Name | string | NA |
| nhd\_lat | Latitude | Deg |  |
| nhd\_long | Longitude | Deg |  |
| LakeArea | Surface area of the lake | hectares |  |
| WS.Area | Surface area of the watershed | hectares |  |
| MaxDepth | Maximum depth of lake | Meters | NA |
| state\_name | Name of US state that lake is located in (or partially in) | string |  |
| prediction.05 | Prediction interval: 0.05 quantile | mg L-1 |  |
| prediction.50 | Median prediction | mg L-1 |  |
| prediction.95 | Prediction interval: 0.95 quantile | mg L-1 |  |

## Articles

(List articles citing this dataset)

|  |  |  |
| --- | --- | --- |
| Article DOI or URL (DOI is preferred) | Article title | Journal title |
| TBD | Lakes at risk of chloride contamination | TBD |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Scripts/code (software)

(List any software scripts/code you would like to archive along with your data. These may include processing scripts you wrote to create, clean, or analyze the data.)

|  |  |  |
| --- | --- | --- |
| File name | Description | Scripting language |
| QRF\_script.R | R code which builds a quantile regression forest model using observational chloride data and predictor variables found in lakeCL\_trainingData.csv | R |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Data provenance

(Were these data derived from other data? If so, you will want to document this information so users know where these data come from.)

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset title | Dataset DOI or URL | Creator (name & email) | Contact (name & email) |
| Water Quality Portal | <https://www.waterqualitydata.us/> | USGS/EPA |  |
| LAGOS-NE-GEO | doi:10.6073/pasta/16f4bdaa9607c845c0b261a580730a7a | Soranno, Patricia A; Michigan State University  Cheruvelil, Kendra S; Michigan State University |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Notes and Comments

1. This document liberally borrows from similar documents at SBC and GCE [↑](#footnote-ref-1)