Predictability of Shoreline Flooding

Owen A. Isley
oisley@u.rochester.edu
University of Rochester
Rochester, New York, U.S.A.

ABSTRACT

Are there meaningful predictors for dangerous shoreline flooding?
The objective of this report is to investigate whether environmental measurements can be valid predictors for flooding events.

KEYWORDS

neural network regression, correlating features, pattern recognition

ACM Reference Format

Owen A. Isley. 2020. Predictability of Shoreline Flooding. *In Proceedings of DSC 391 (Owen A. Isley)*. ACM, New York, N.Y., U.S.A., 8 pages.

1 INTRODUCTION

Floods are a serious and recurring hazard for shoreline communities. Residents' properties and lives are put at risk when dangerous flooding occurs. Variables that predict floods can arm shoreline communities with an ability to identify when dangerous flooding is liable sooner.

The American government agency N.O.A.A. (National Oceanic and Atmospheric Administration) has recorded every flood near Lake Ontario between 2005-2010 in a public data set *NOAA Storm Events Database*.[1] Each record contains the date and time on which a shoreline flood occured, its score on the Fujita-Pearson scale (F-scale), its magnitude, the number of fatalities and injuries it caused, and the amount of crop and property damage it caused (in dollars)—among other features.[1]

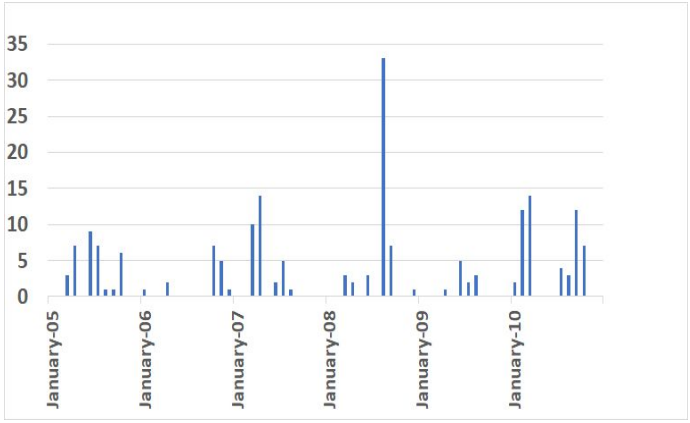

Figure 1: Number of Hazardous Shoreline Floods versus Time, Lake Ontario 2005-2010

Understandably, NOAA has also collected environmental measurements near Lake Ontario between 2005-2010. After investigating https://www.data.gov/ using the search term "Lake Ontario", the following measurements were found to have been collected by NOAA outside *Storm Events Database*:

- Over Basin Precipitation (cubic meters per second)[2]
- Overlake Precipitation (millimeters)[3]
- Overland Air Temperature (Celsius)[4]
- Overlake Air Temperature (Celsius)[5]
- Water Surface Temperature (Celsius)[6]
- Water Level (Feet)[7]

Combining the environmental data sets with *Storm Events* can provide insight into natural phenomena that associate with floods.

After combining and preprocessing the seven data sets, the dimension of the merged data set is 72 records (one for each month from 01-2005 through 12-2010) across 14 attributes:

- MONTH
  - (01-12)
- YEAR
  - (2005-2010)
- OVER_BASIN_PRECIP
  - (monthly total)
- OVERLAKE_PRECIP
  - (monthly total)
- OVERLAND_TEMP
  - (monthly average)
- OVERLAKE_TEMP
  - (monthly average)
- WATER_SURFACE_TEMP
  - (monthly average)
- WATER_LEVEL
  - (monthly average)
- NUMBER_PRECIP_STORMS
  - (per month)
- SUM_F_SCALE
  - (of all storms given month)
- SUM_MAGNITUDES
  - (of all storms given month)
- SUM_FATALITIES_INJURIES
  - (from all storms given month)
- SUM_DAMAGE
  - (from all storms given month)
- NUMBER_DANGEROUS_FLOODS
  - (per month)

The approach was as follows:

- *Exploratory Data Analysis* Present the composition of the datasets and the distribution of the variables
- *Data Preprocessing* Convert nonconforming units; calculate monthly totals for precipitation, number of storms, F-scales, magnitudes, fatalities and injuries, damage, and number of dangerous floods; bucket all records by month and truncate to 01-2005 through 12-2010; merge all data sets on primary key (MONTH, YEAR)
- *Analysis and Modeling* The techniques used include a pairwise correlation across the features, a scatterplot matrix of the features, a neural network regression with 1 hidden layer of 9 nodes, and a manual k-fold cross validation to determine the most appropriate train-test split

This report will detail each step of the aforementioned process and examine the final results.

2 EXPLORATORY ANALYSIS

To get an understanding of the composition of the data, several basic analyses were performed using the R programming language and the `ggplot2` visualization library.
The exploratory data analysis consists of:

- *Data Exploration*
- *Distribution of Numeric Features*

2.1 Data Exploration

2.1.1 As Numeric

Regressions require all features to be numbers which Figure 2 confirms:

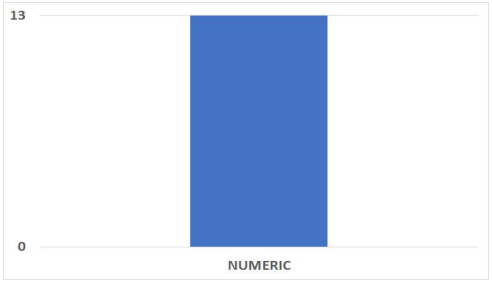

Figure 2: Features by Data Type

2.2 Distribution of Features

The following is a casual analysis of the features in question:

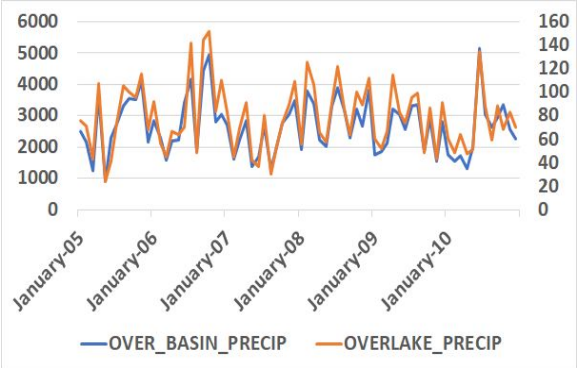

Figure 3: Precipitation Features over Time

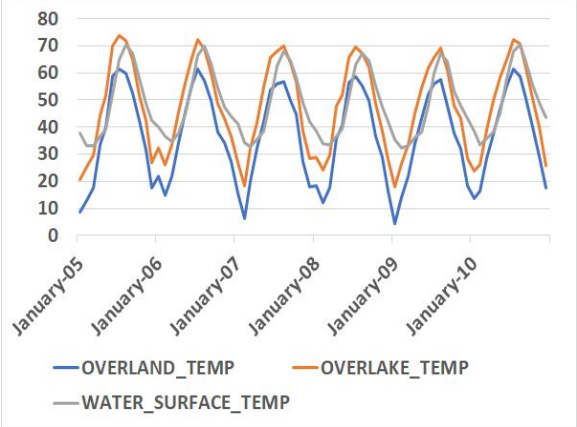

Figure 4: Temperature Features over Time

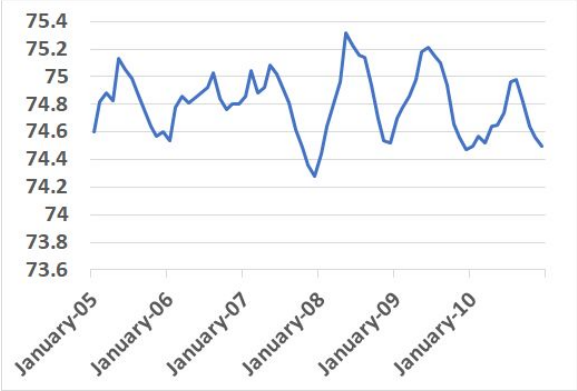

Figure 5: Lake Ontario Water Level over Time

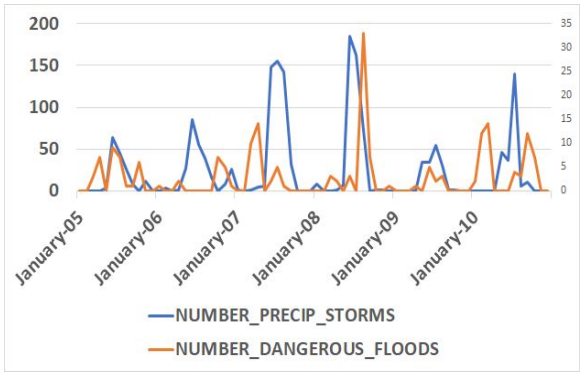

Figure 6: Cardinal Features over Time

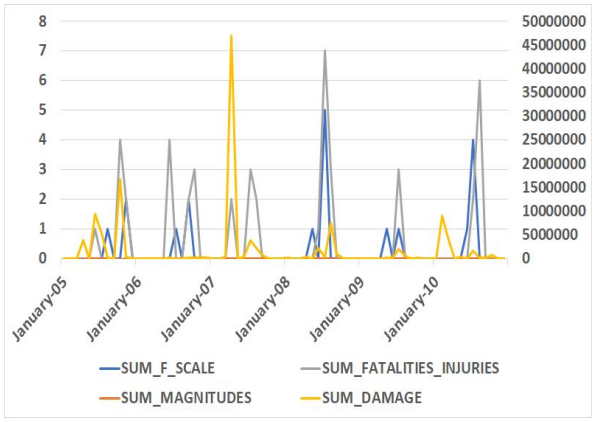

Figure 7: Storm/Flood Severity Metries over Time

3 DATA CLEANING, PREPROCESSING, AND FEATURE ENGINEERING

For this report, the data is examined on a monthly level. This requires calculating the amount of precipitation per month, the number of precipitation-related storms per month, the sum of the floods' F-scale and magnitude scores per month, the number of injuries and fatalities caused by floods per month, the total amount of damage caused by flooding per month, and the number of dangerous floods per month for all 72 months between 01-2005 and 12-2010. The temperature and water level measurements were already reported as monthly averages.

3.1 Data Cleaning

3.1.1. Removing Extraneous Features

Storm Events Database contains many features for each storm event. The only fields pertinent to this analysis are BGN_DATE, EVTYPE, F, MAG, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP, LATITUDE, LONGITUDE so all other columns were removed.

3.1.2. Truncating Data to 2005-2010

The years 2005-2010 is the longest span of time that all the data sets have in common. Since we require every record to have entries for every feature, each data set was constrained to records from 2005-2010 only.

3.2 Data Preprocessing and Feature Engineering

3.2.1 Filtering for Lake Ontario Data

By filtering LATITUDE and LONGITUDE, records for only the Lake Ontario shoreline region can be isolated. This was achieved using the following parameters: `filter(LATITUDE >= 4300, LATITUDE <= 4410, LONGITUDE <= 8000, LONGITUDE <= 7600)`

3.2.2. Floods and Precipitation-related

The EVTYPE column from *Storm Events Database* labels the type of storm each record is. The events relevant to our analysis are the flooding events (FLOOD, FLASH FLOOD) and events that are associated with precipitation (TSTM WIND, HAIL, WATERSPOUT, THUNDERSTORM WIND, LIGHTNING, MARINE THUNDERSTORM WIND, FUNNEL CLOUD, HEAVY RAIN, MARINE HIGH WIND).

3.2.3 Unit Conversions

All environmental measurements were time stamped using the UNIX date format, which was converted to the conventional MM/DD/YYYY using the Excel formula `=A1/(24*60*60) + DATE(1970,1,1)`

The temperature measurements were converted from celsius to fahrenheit using the Excel formula `=CONVERT(B1,"C","F")`.

The OVER_BASIN_PRECIP measurements were rounded to the nearest cubic meter per second.

The PROPDMG and CROPDMG calculations were recorded alongside corresponding PROPDMGEXP and CROPDMGEXP columns, which denote the magnitude of the corresponding PROPDMG/CROPDMG. For example PROPDMG=2 with PROPDMGEXP=M means $2,000,000 of property damage was attributed to that storm event. For this analysis it is important to have PROPDMG and CROPDMG to be represented by their true values as shown in the above example. So PROPDMG and CROPDMG were combined with the corresponding PROPDMGEXP and CROPDMGEXP to convert to their true values. Lastly, for all records, the PROPDMG and CROPDMG columns were added together to create the SUM_DAMAGE column, which more accurately reflects the total damage caused by each storm (in dollars).

The FATALITIES and INJURIES measurements were added together to calculate SUM_FATALITIES_INJURIES since it is redundant to keep these features separate.

3.2.4 Filtering Dangerous Floods

Many of the flooding events reported in the Lake Ontario region did not cause any property or crop damage, and did not result in any injuries or deaths. Such floods are not interesting since this report is only concerned with hazardous floods. All non-hazardous floods were removed so the only flood events that remained either caused some amount of damage or resulted in some number of injuries or deaths.

3.2.5 As Monthly Totals and Averages

For all data sets the MM/DD/YYYY format was parsed to MM-YYYY and then separated into their own columns called MONTH and YEAR. For each month 01-2005 through 12-2010 the OVER_BASIN_PRECIP, OVERLAKE_PRECIP, F, MAG, SUM_FATALITIES_INJURIES,

SUM_DAMAGE columns were added together to reflect monthly sum totals.

A count of the hazardous floods for each month was calculated to get a monthly NUMBER_DANGEROUS_FLOODS, and a count for the number of precipitation-related storms per month was calculated and recorded as NUMBER_PRECIP_STORMS.

The temperature and water level measurements are already reported as monthly averages.

3.2.6 Merging the Data Sets

Since all of the data sets now contain only 72 rows corresponding to a month from 01-2005 to 12-2010, the MONTH, YEAR columns can be used as a primary key on which the 7 data sets can be merged. The resulting data set contains 72 records (1 for each month) with across the 14 combined attributes MONTH, YEAR, OVER_BASIN_PRECIP, OVERLAKE_PRECIP, OVERLAND_TEMP, OVERLAKE_TEMP, WATER_SURFACE_TEMP, WATER_LEVEL, NUMBER_PRECIP_STORMS, SUM_F_SCALE, SUM_MAGNITUDES, SUM_FATALITIES_INJURIES, SUM_DAMAGE, NUMBER_DANGEROUS_FLOODS.

4 ANALYSIS AND MODELING

4.1 Pairwise Correlation of Features

To get an understanding of how the 14 features correlate with one another, a pairwise correlation plot was constructed using the R library corrplot:

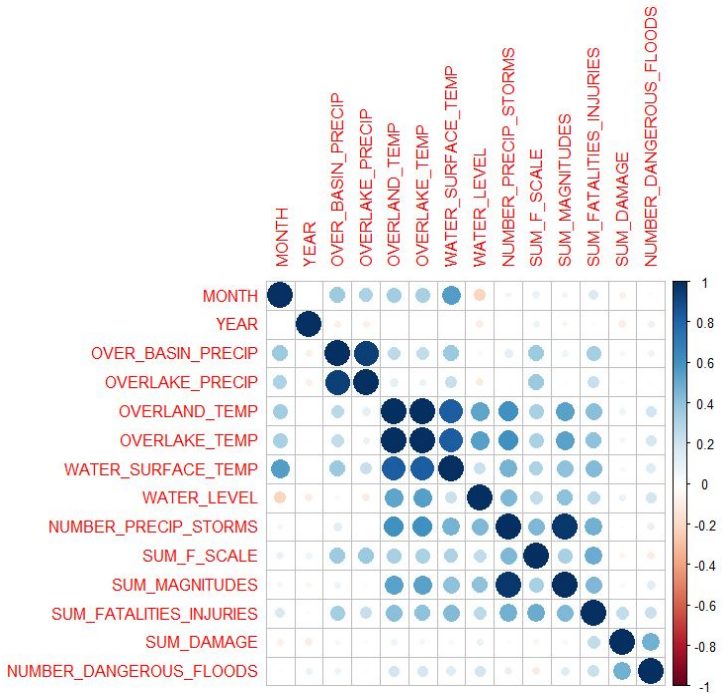

Figure 8: Pairwise Correlation Plot (`method="circle"`)

In addition, a scatterplot matrix of all the features was created using ggplot2's function ggpairs to visualize a pairwise matrix of scatterplots:

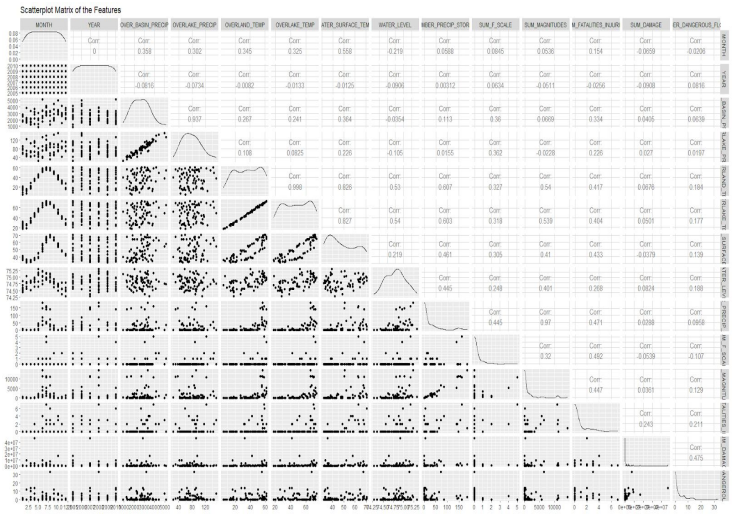

Figure 9: Scatterplot Matrix of the Features

4.2 Feedforward Neural Network

The R library `neuralnet` was used to create an artificial neural network with 12 input variables `MONTH`, `OVER_BASIN_PRECIP`, `OVERLAKE_PRECIP`, `OVERLAND_TEMP`, `OVERLAKE_TEMP`, `WATER_SURFACE_TEMP`, `WATER_LEVEL`, `NUMBER_PRECIP_STORMS`, `SUM_F_SCALE`, `SUM_MAGNITUDES`, `SUM_FATALITIES_INJURIES`, `SUM_DAMAGE`, 1 output variable `NUMBER_DANGEROUS_FLOODS`, 1 hidden layer, and 9 nodes in the hidden layer.

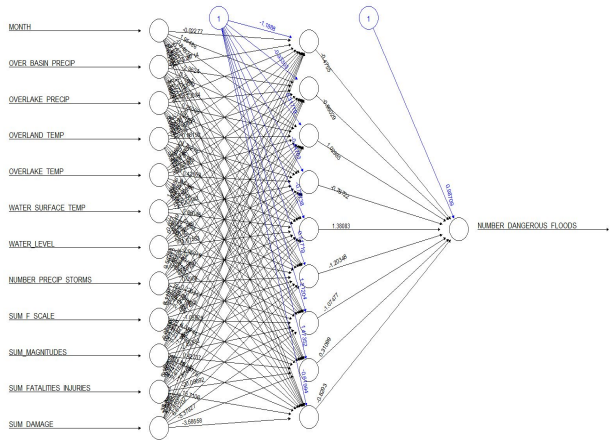

Figure 10: Neural Network with Labeled Weights

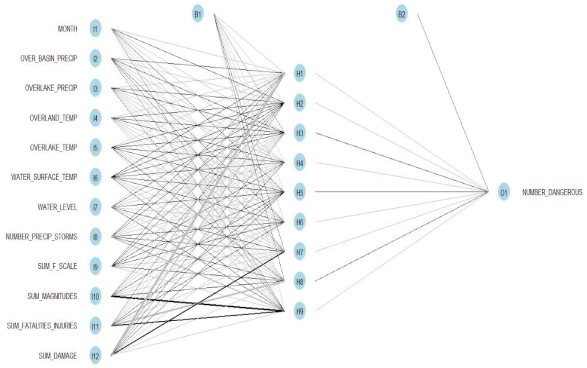

Figure 11: Neural Network with Visualized Weights

A manual k-fold cross validation was performed to determine the appropriate train-test split of the dataset across 90:10, 80:20, 70:30, 60:40, 50:50, and 40:60 train-test ratios. Using `neuralnet`'s `compute` function on each of the ratio splits, each split's Root Mean Square Error (RMSE) was calculated and it was determined that using 80% of the data to train the neural network resulted in the lowest RMSE (5.929935).

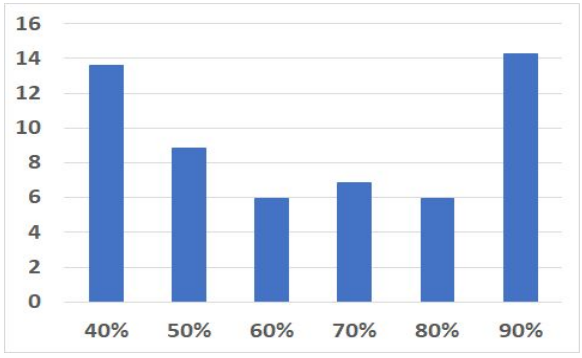

Figure 12: Percentage Used to Train versus RMSE

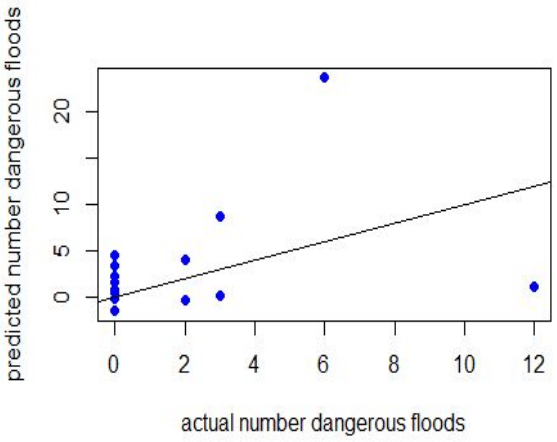

Figure 13: Neural Network Predicted versus Actual Number of Dangerous Floods

5 CONCLUSIONS

Root Mean Square Error (RMSE) is a standard measurement for the error of a predictive model and is formally defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

n is the number of observations

One can notice that the RMSE equation closely resembles the formula for Euclidean distance. This means that heuristically RMSE can be thought of as a normalized distance between the vector of predicted values and the vector of observed values. In other words, RMSE answers *How far off should we expect the model to be on its next prediction?*

Training on 80% of the months, the neural network was able to predict the number of floods for each of the remaining 20% of months with a RMSE of 5.929935. This means that when provided a month's environmental measurements only, this model can predict the number of hazardous floods which occured that month with an *average* error of about 6 floods.

Between 01-2005 and 12-2010, the average number of floods per month was 2.7, the median 0, and a range of 0-33 floods. This indicates that *most months did not experience floods* but some months had a very high number of floods (for example 04-2007=14, 08-2008=33, 03-2010=14).

It would appear that these outlier months threw off the neural network's *average* error, which is a sign of model overfitting. This is highlighted by Figure 13 where the model predicted 0 floods when there were actually 12, and over 20 floods when there were actually 6.

However Figure 13 also indicates that the model recognizes *non-eventful months with a much higher degree of accuracy*—denoted by the cluster of correctly predicted months with fewer than 5 hazardous floods. *This suggests that environmental metrics can be used to predict non-eventful months (the majority of months), but they fail to capture months with an outstanding number of hazardous floods.* This explains the rather high RMSE.

For future work, predicting a month with an unusually high number of floods remains a challenge. The model's performance can be improved by either considering a longer time period so that the neural network sees more instances, or by introducing more features that better correlate with super high frequencies of hazardous floods. While this shows that hazardous flooding is predictable to a degree, a more sophisticated model which considers more features is required to correctly identify the extreme months.

6 ACKNOWLEDGMENTS

I thank Professor Pawlicki for offering me a wonderful opportunity to apply what I learned throughout my data science journey with the University of Rochester. This project was a nice denemount to my time at Rochester—from CSC 171 and 172 through CSC 240, 242, 265 and others. While completing this assignment I realized that I would like to specialize in predictive models going forward. Stay tuned.■

REFERENCES

[1] National Weather Service, "Storm Events Database." N.O.A.A., April 2020. repdata_data_StormData.csv, https://www.ncdc.noaa.gov/stormevents/

[2] National Weather Service, "Lake Ontario OverBasin Precipitation (cubic meters per second)" N.O.A.A., April 2020. https://catalog.data.gov/dataset/lake-ontario-overbasin-precipitation-cubic-meters-per-second#sec-dates

[3] National Weather Service, "Lake Ontario Overlake Precipitation (millimeters)" N.O.A.A., April 2020. https://catalog.data.gov/dataset/lake-ontario-overlake-precipitation-millimeters19101#sec-dates

[4] National Weather Service, "Monthly minimum overland air temperature for Lake Ontario (Celsius)" N.O.A.A., April 2020. https://catalog.data.gov/dataset/monthly-minimum-overland-air-temperature-for-lake-ontario-celsius#sec-dates

[5] National Weather Service, "Lake Ontario Mean Monthly Overlake Daily Air Temperature [(max+min)/2] (Celsius)" N.O.A.A., April 2020. https://catalog.data.gov/dataset/lake-ontario-mean-monthly-overlake-daily-air-temperature-maxmin-2-celsius#sec-dates

[6] National Weather Service, "Lake Ontario modeled water surface temperatures (Celsius)" N.O.A.A., April 2020.

https://catalog.data.gov/dataset/lake-ontario-modeled-water-surface-temperatures-celsius#sec-dates

[7] National Weather Service, "Aggregated Lake Ontario Water Levels" N.O.A.A., April 2020. https://catalog.data.gov/dataset/aggregated-lake-ontario-water-levels#sec-dates