ARGO: Thermoclines and Typhoon Location

Axel Amzallag

Introduction

Every year, cyclones strike the eastern coast of Asia: the western Pacific equivalent of the hurricane. The damage that these cyclones, called typhoons, cause to infrastructure and human life is substantial. All of these typhoons begin as tropical storms in the Pacific ocean, which subsequently build up to stronger tropical storms and, sometimes, typhoons. Throughout the late 20th and early 21st century, scientists have improved their ability to track these cyclones and trace their paths. In order to support these scientists tracking these storms, it would be helpful to know where in the ocean the cyclones are most likely to begin. The earlier in their build-up process the storms can be discovered, the faster their paths can be mapped and predicted. In the end, this would help provide citizens of East Asian countries with more time to prepare for the arrival of typhoons and get to safety. At the corporate level, additional time to prepare for tropical storms can be used to protect assets. For regional and national governments, knowing the most likely locations of the beginnings of cyclones would help target tracking resources more efficiently and give more time to deploy emergency resources as the storms develop.

In order to provide more information on this phenomenon, we ask the question: Can we find a useful predictor of typhoon starting location during the summer months in the northwestern Pacific ocean? Since typhoons are formed from water near the surface of the ocean, in particular we were interested in the thermocline initial depth of the ocean at various latitudes and latitudes and typhoon starting location. Is there a statistically significant relationship between thermocline initial depth and typhoon starting location during the typhoon months?

Data Backgound, Collection, and Cleaning for ARGO

In order to begin tackling the issue of cyclone starting location, an oceanic dataset that contains variables and their connection to latitude/longitude was required. The ARGO dataset provided by the University of California, San Diego is ideal for this purpose. The project consists of around 4,000 buoys that dive into the ocean and collect measurements of various quantities. The Core Argo buoys – the original operational mission – are the main bulk of the fleet; they collect depth, salinity, and temperature, as well as location and time coordinates. The buoys dive into the ocean, on average to a depth of 1000-1500 meters, drift for 10 days, and then begin climbing back to the surface. As they climb, they collect measurements for the core variables. Once at the surface, they transmit the data from their dive to a satellite, which then transmits the data to a processing center. Finally, the scientists at the processing center clean the data, double-check the quality with high-quality boat measurements, and submit both the raw and cleaned data to an ftp server where the publically available data is kept. Importantly, it can take 1-2 years for all of the data to be quality-controlled appropriately and cross-checked by a high-quality boat reading. Therefore, using reliable data requires only using the data provided up until 2019.

On the server where the publically available data is stored, the initial breakdown is by ocean, and then the data is broken down further by year, month, and day. For the purposes of Pacific typhoons, the desired data was stored in

the Pacific Ocean directory. All years up to 2019, and all months/days can be used. The data is also provided in the .nc (netCDF) file format, a standard file format used by the National Oceanic and Atmospheric Administration (NOAA). So, there is a .nc file for each day that data has been collected since the beginning of the ARGO project. The first step of the process was to collect the data from the website; for this, a simple bash script was used that downloaded all desired days of data in the Pacific ocean. Next, the data was read into R and the package *ncdf4* was used to manipulate the data into a more workable format. All of the data was read into a data frame, the missing data values were removed, and then the newly cleaned data was outputted to a .csv file for ease of use in later calculations. The variables that were kept in the final .csv file to be used in the upcoming analysis were buoy ID number, longitude, latitude, Julian date (number of days since January 1, 1950), reference date (January 1 1950), temperature, pressure, and salinity. You can see an example of the top few rows of the data here below:

```
##   PLATFORM_NUMBER LONGITUDE LATITUDE    JULD REFERENCE_DATE_TIME TEMP_ADJUSTED
## 1         2900977   138.889   32.917 22127.9         1.95001e+13        28.734
## 2         2900977   138.889   32.917 22127.9         1.95001e+13        28.784
## 3         2900977   138.889   32.917 22127.9         1.95001e+13        28.639
## 4         2900977   138.889   32.917 22127.9         1.95001e+13        27.906
## 5         2900977   138.889   32.917 22127.9         1.95001e+13        26.462
## 6         2900977   138.889   32.917 22127.9         1.95001e+13        25.088
##   PRES_ADJUSTED PSAL_ADJUSTED
## 1           5.0        33.849
## 2          10.3        33.931
## 3          21.1        34.090
## 4          30.8        34.257
## 5          40.6        34.474
## 6          50.6        34.650
```

Initial Data Exploration

Data from individual buoys provided the first clue about which variables would be of interest and which would not. The main relationship that we were focused on was the behavior of the temperature as depth increases. Temperature is remarkably stable near the surface of the ocean, when the depth is small, but starts decreasing rapidly at a certain depth. This location is the beginning of what is called the *thermocline*: a layer of ocean water, immediately below the surface layer, where the temperature of the water drops precipitously. This pattern is remarkably consistent for the majority of the buoys near the equator, where most typhoons begin throughout the year. Buoys further north, much nearer the Arctic circle, have no noticeable thermocline at all.

In Figure 1, we can see a chart of temperature and depth for a group of buoys over the course of a year. In the figure, both plots are of the same buoys over the same year, but with different depths – up to 1500 meters on the left and 200 meters on the right – plotted. The green/orange lines, which start near a temperature of 30 degrees and dip down, have a clear thermocline where there temperature drops over a relatively small increase in depth. These measurements were made in tropical waters near the equator over the course of the year. The blue/purple lines, which start at very low temperatures and don't change much over time, have no thermocline at all. These measurements were made in sub-Arctic waters, and although they have a small thermocline in the summer/fall.

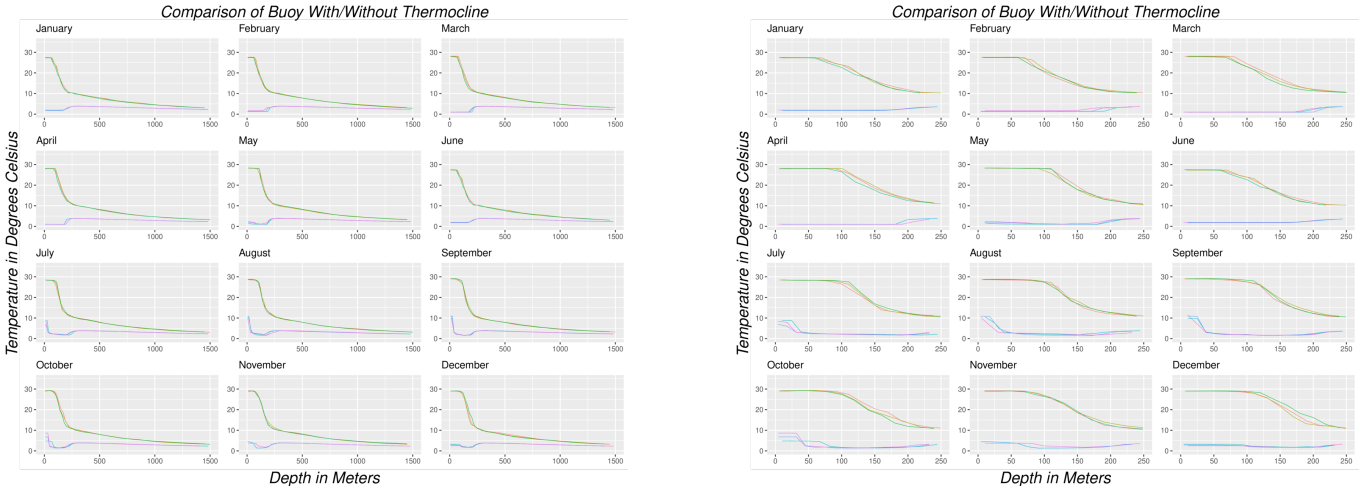

Figure 1: Depth versus Temperature for buoys in tropical and sub-arctic waters.

Once the initial look at individual buoys was completed, the next observation was the common locations where cyclones started. In an initial naive look at the paths of typhoons in the western Pacific Ocean, it appears that there are locations where tropical storms begin frequently and locations where they do not. Figure 2 shows the paths that were taken by typhoons in the first half of 2019. It can be seen from the figure that the paths do not seem to be starting in random locations. Almost all of the cyclones in the figure began in the southern area of the map, while the northeast portion has no tropical storms at all.

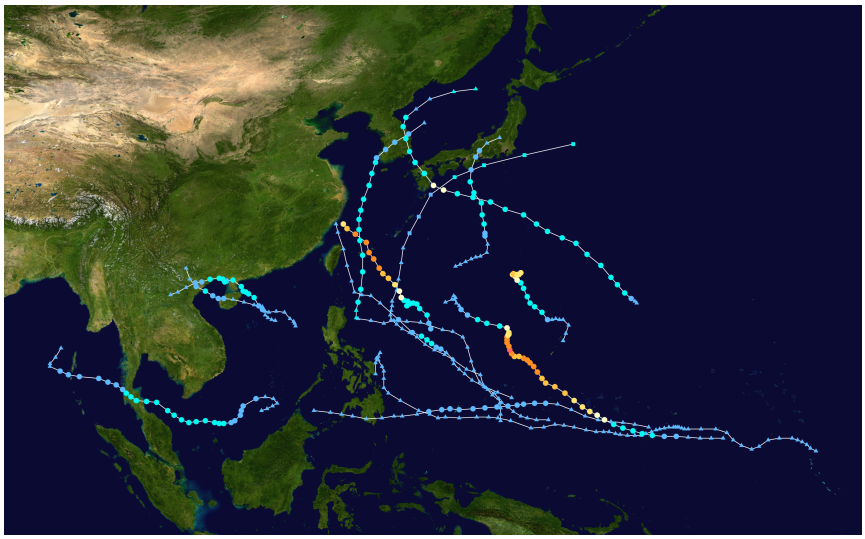

Figure 2: Map of Pacific cyclone season: first half of 2019. Image provided to the public domain by the Tropical Cyclones WikiProject (Link)

For typhoon data we used the US Navy's Joint Typhoon Tracking Center Data (Link). They have detailed records of every typhoon in the northwestern Pacific region dating back to 1945. However, the data is stored in a format that was not helpful for our purposes. As such, we made a small shell script that reformatted the files to be more easy to read into R. Out of the data that the Navy keeps, the parts that are of interest for our purposes are Latitude

and Longitude. Below, you can see an example of what the top of a typhoon file looks like:

```
##   X1 yr_cyc_num  date_time  lat  long vmax mslp typh_grade
## 1  1         01 2016052600 17.6 114.6   20 1005         DB
## 2  2         01 2016052603 18.2 114.7   20 1005         DB
## 3  3         01 2016052606 18.6 114.4   20 1005         DB
## 4  4         01 2016052612 19.1 113.6   25 1004         TD
## 5  5         01 2016052618 19.6 112.9   25 1004         TD
## 6  6         01 2016052700 20.4 112.7   25 1004         TD
```

Creation of the Initial Thermocline Depth Statistic

Typhoons are formed from water at the ocean's surface, not water from the depths, so creating variables that measured different aspects of the ocean's behavior near the surface of the ocean was the ideal way to follow up and note any statistical relationships between location and surface values. The primary statistic in our analysis was the depth at with the thermocline begins, or *thermocline initial depth* (TID). This statistic seems a logical one to use since tropical storms are formed from water of the top layer of the ocean. The depth of the top layer is therefore a variable that could be of interest. The best way to collect this statistic isn't fully fleshed out as of yet. Creating an objective description of thermocline depth is challenging, but researchers have come up with several methods to measure it (Fiedler, Comparison of objective descriptions of the thermocline, 2010). One option is to find the line segment with maximum slope with the restriction that the change in depth be greater than 20 meters. The median depth of that line segment can be chosen as thermocline depth. Another method is to choose the point where the temperature has dropped halfway between the SST and the temperature at 400 meters, called the *variable representative isotherm* (VRI) method.

For our analysis, the VRI method was used as our way of measuring thermocline depth. The primary reason we chose this method was because it had the most accurate results in Fiedler's paper. The VRI method had a root mean-squared error of around 10, while the other methods discussed had a root mean-squared error of around 20 (Fiedler 2010). The formula for the VRI method is a gradient-free estimation method with the following equation:

$$\text{TT} = 0.25\,[\text{T(MLD)-T(400m)}]$$

$$\text{TT: Thermocline temperature}$$

$$\text{T(MLD): Temperature of Mixed Layer Depth}$$

$$\text{T(400m): Temperature at depth of 400 meters}$$

Some weaknesses of the method are that it isn't tested on points with latitudes greater than 40 degrees and doesn't work well on data with surface temperatures less than 10 °C. However, since typhoons begin near the equator and that is neither of these categories, the VRI method of measuring thermocline depth can be used pretty safely in our analysis. After creating a function that takes as input our data and outputs the thermocline depth as determined by the VRI method, we get a thermocline depth measurement at every point and with a mean (dependent on the month) of around 70 meters and a standard deviation of 45 meters. Below, in Figure 3, you can see an example of

our Thermocline depth data from August 2010, and an example of the VRI method on a single buoy's thermocline profile.

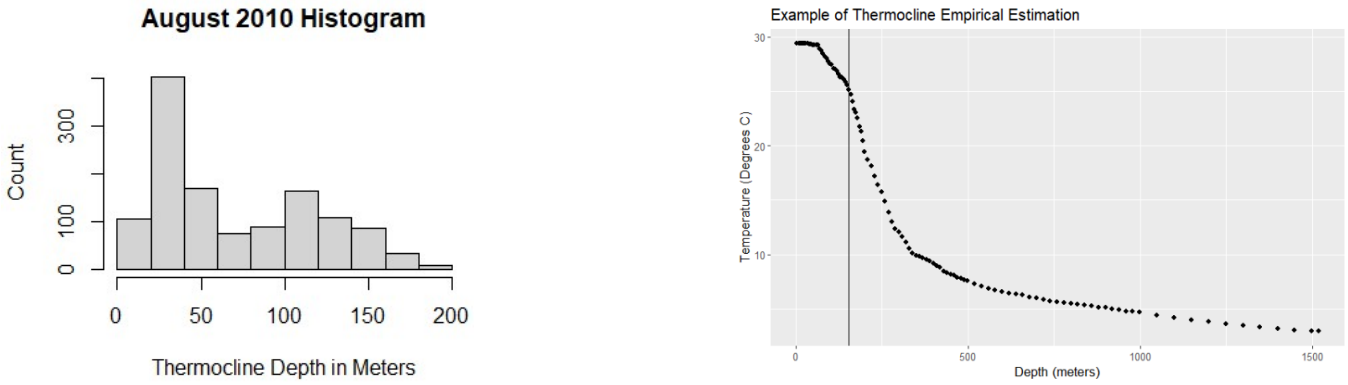

Figure 3: (Left) Histogram of the outputs from the VRI for August 2010. (Right) Example of the VRI method for a single buoy, marked by a vertical line.

Model for Thermocline Depth at any Point

Since the data is quite sparse, there is no way to use empircal measurements for the typhoon analysis, and so a continuous model of the thermocline is necessary. So, our intermediate goal is to estimate the thermocline depth at every single possible point, instead of only the sparse points where we have buoys. Once we have a model that estimates the thermocline depth at every point, we would be able to analyze its relationship to typhoons. In order to do this, we can use a general additive model (GAM). Generalized additive models are often used with temperature and climate data because that data is continuous and often follows a very nonlinear distribution similar to a Gaussian curve. This generalization of linear mixed models is made up of a sum of penalized splines, where we can choose the type of spline, the basis dimension, and the degree of smoothing for each term. The general formula for a GAM is

$$g(\mathbb{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + ... + f_m(x_m).$$

In our particular circumstance, due to our limited amount of data and the sparsity of the data points, we chose to train our model on all of the typhoon data from 2016-2018. The reason for this is because, as was seen in Figure 1, the thermocline is actually quite consistent over the months of the year in the tropical regions of the ocean. Therefore, we decided that adding the training data from times outside of typhoon season (July-October), and adding month both as a stand-alone variable and as an interaction variable with latitude would be worthwhile. The form of our model is as follows:

$$\mathbb{E}\left(\sqrt{\text{Thermocline Depth}}\right) = f_1\left(\text{Latitude,Longitude}\right) + f_2\left(\text{Latitude}\right) + f_3\left(\text{Latitude,Month}\right) + f_4\left(\text{Month}\right).$$

For the splines, we chose sphere splines for the longitude/latitude interaction, and we settled on penalized thin plate splines for latitude. We let the month and latitude/month terms have random spline effects. Initially, we ran a model assuming a Gaussian response variable, but changed to a scaled t-distribution after a residual analysis. The distribution was quite clearly heavy-tailed and the heavy tails of the t-distribution were a much better fit to our

given data, as can be seen in Figure 4.

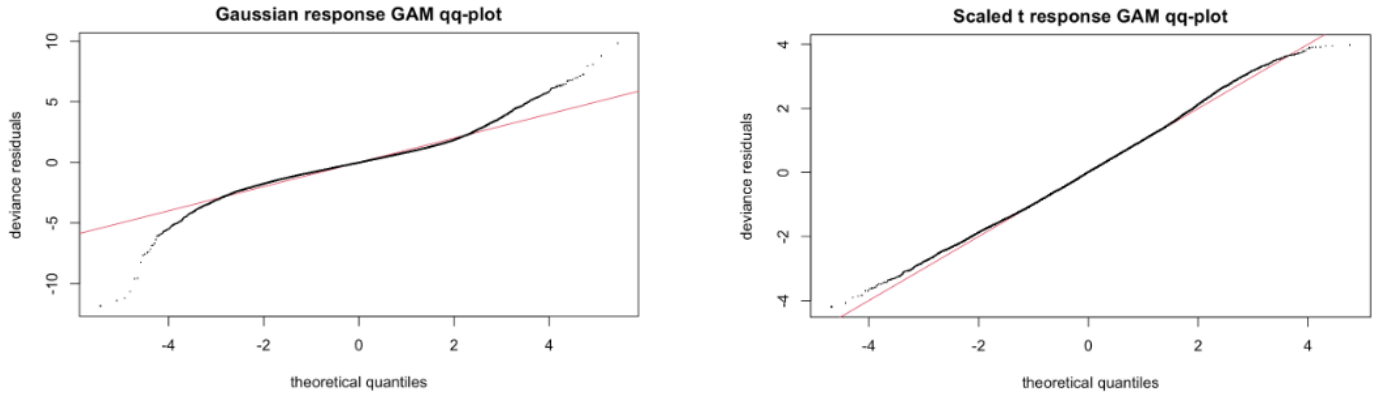

Figure 4: QQ Plot for the Gaussian response and the scaled t response variables.

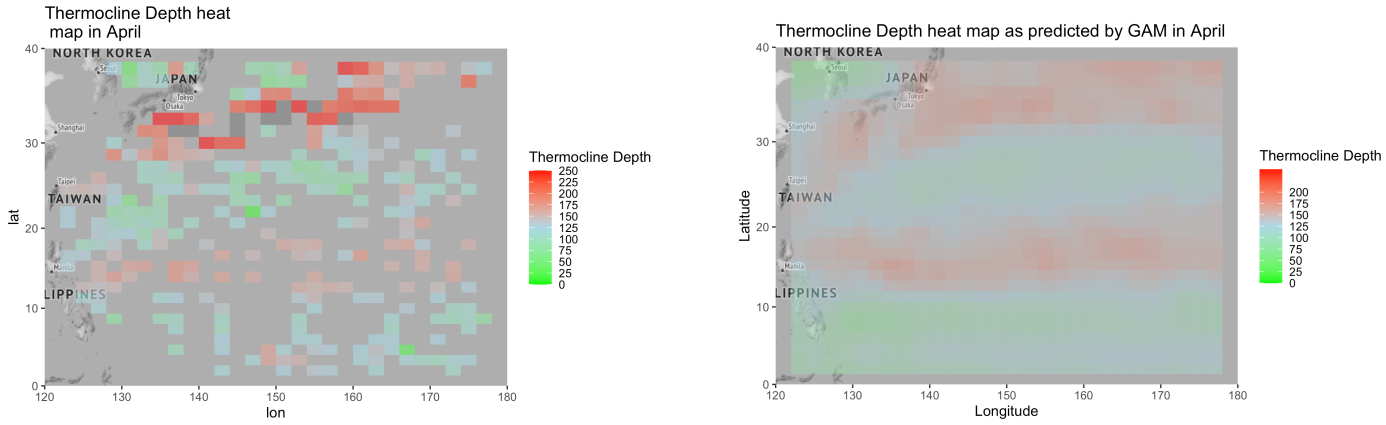

Figure 5: Heat map comparing the empirical thermocline and the predicted thermocline outputted by the GAM

Our model, once we switched to the scaled t-distribution rather than the Gaussian distribution, passed all of the usual criteria for checking models, such as qq plots and residual analysis. In the tropical zones where typhoons begin, the model was quite accurate. Testing our model on the data from typhoon season in 2019, we got a root mean-squared error of around 20 meters compared to the actual values from the 2019 buoys. While this root mean-squared error is higher than desired, it is small enough that this model can still be used confidently for the upcoming typhoon analysis. A visual example of the model for the month of April can be seen in the heat maps of Figure 5, with the empirical values and sparsity on the left and the continuous model built from that empirical data on the right.

Final Typhoon Analysis

Now that we have our model of thermocline depth, we are ready to perform our analysis of thermocline depth with the typhoon data. One challenge here is that there are only about 20 typhoons per year, and so there is very little data to work with. To get around this challenge and simplify the problem, we chose to create bins by latitude and longitude. What we did in this case was create small bins of 1° latitude by 1° longitude and we counted the number

of typhoons that started in each of the bins. For thermocline depth, we calculated the thermocline depth at the centroid of each bin and used that value as the thermocline depth for that bin. So, the problem becomes simplified to an analysis of 240 values of the independent variable (Thermocline Depth) and 240 values of the response variable (Number of Typhoons). Our first analysis was a simple univariate linear model,

$$\mathbb{E}\left(\text{Number of Typhoons}\right) = \beta_0 + \beta_1 \cdot \text{Thermocline Depth}.$$

We found that this was a statistically significant relationship, but looking at the data we clearly have a nonlinear relationship. Below, in Figure 6, there is a plot of the regression and the line of best fit, which doesn't appear to match the distribution of the data.

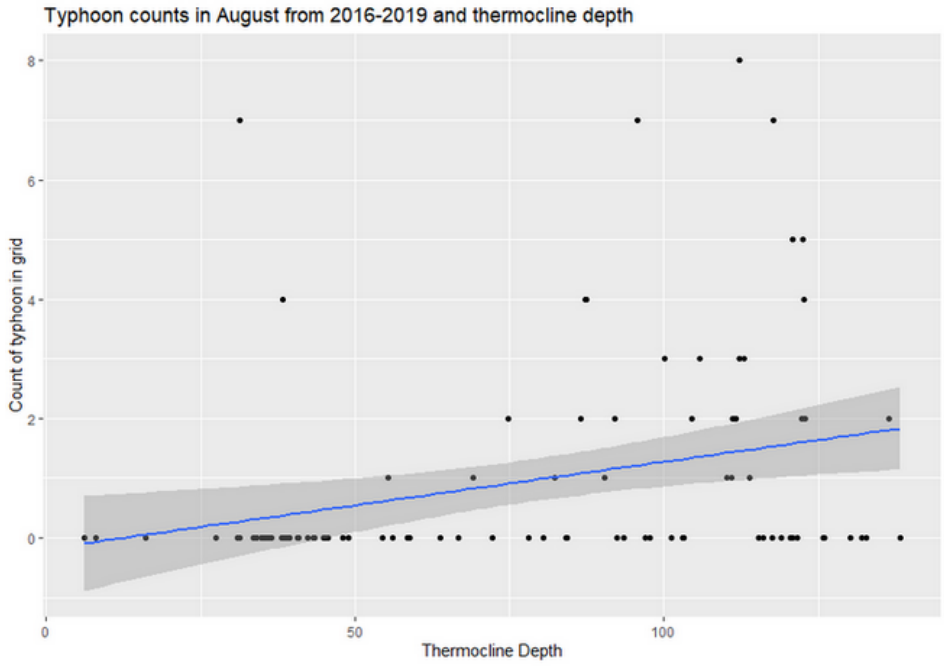

Figure 6: Linear regression of Typhoon Starting Counts on Thermocline Depth

Due to the non-linearity of the data and the response variable being in terms of counts, a Poisson or negative binomial regression would be more logical than a linear regression. Since the data is both zero-inflated and overdispersed, we model using a negative binomial regression. The negative binomial regression is a generalized linear model defined by

$$\text{Negative Binomial Regression: } \begin{cases} y_i | x_i & \sim \text{NB}(\mu_i, \theta) \\ \mu_i & = e^{x_i^\top \beta} \end{cases}.$$

So, instead of our previous naive linear model, we have the model

$$\mathbb{E}\left(\text{Number of Typhoons}\right) = \exp\left\{\beta_0 + \beta_1 \cdot \text{Thermocline Depth}\right\}.$$

When we run this model with our data, we get that the result is still statistically significant ($p < 0.05$), and we get the coefficient $\exp\left(\hat{\beta}_1\right) = 1.0232$. This we can interpret as follows: every 1 meter increase in thermocline

depth increases the number of typhoons in that bin by a multiplicative factor of 1.02. So, there is an exponential relationship between the variables that is statistically significant. This analysis shows evidence of a correlation between thermocline depth and the likelihood of a typhoon starting. If countries in the East Asian region measure the temperature at different depths prior to typhoon season, there will be a correlation between those values and the locations where typhoons will start during the typhoon season. While not causal or predictive, this can definitely be actionable for deployment of typhoon tracking assets.

Conclusion

In this analysis, we showed a correlation between thermocline depth and typhoon likelihood. However, there are confounding variables to analyze further, such as surface temperature. Surface temperature could not be modeled in this analysis due to time constraints. In order to analyze the effect of surface temperature on typhoon likelihood, it would be necessary to once again create a statistic for surface temperature and create a continuous model from the sparse data. As an example of how this could help our analysis, thermocline depth near 20° latitude in April was deeper than 5° latitude in August, yet the latter consistently has typhoons forming while the former does not. So, thermocline depth is only part of the picture, and there must be other elements at work. This is why we chose only to analyze thermocline depth and typhoons during the summer months, since the spring months had behavior that was clearly confounded in some way. Further analysis could look at this further, and try to get a more robust method for predicting typhoon likelihood across all months rather than only a pre-selected subset of months. On this front, the ARGO dataset is not sufficient to fully complete the analysis. Surface wind data, as well as more granular temperature data, would probably be required to complete a more thorough and deep analysis.

In conclusion, we successfully were able to get an empirical estimate of the initial depth of the thermocline in tropical and temperate regions of the ocean. Using the variable representative isotherm empirical estimation method and then modeling using a general additive model, we constructed a continuous measurement of thermocline depth across the typhoon region. Finally, with a binning simplification and a negative binomial model, we showed a statistically significant relationship between thermocline depth in the summer months and the likelihood of a typhoon starting in that location. Our result finds that a 1 meter increase in thermocline depth in the summer is correlated with a 1.02 multiplicative increase in number of typhoons during typhoon season.