# STAT 4620/5620 Final Project

# Predicting Wave Heights Using Random Forest Methods

Alexander Boutilier

B00801543

&

Seoyeon (Cali) Park

B00768397

Winter 2021/2022

# 1 Abstract

**Keywords**: *Oceanography, Wave Height, Data Manipulation, Random Forest, Prediction, RMSE*

Canada has put extensive research into collecting valuable oceanographic data by placing multiple buoys along the coastlines. For this report, raw time series data from 4 different buoys selected along the coast of Nova Scotia were combined and sorted, in an attempt to accurately predict wave heights based on a random forest model, and will attempt to analyze which meteorological variables are most significant in prediction. Other scientific questions will be answered during the data exploration stages. Model performance will be based on R-squared and mean root mean squared error values.

# 2 Introduction

Forecasting of wave heights are one of the main problems scientists are trying to solve. Scientifically, this problem involving fluid dynamics is quite complicated, as there are numerous different components that may affect the height of the waves. Some of these variables may include meteorological components such as temperature and pressure of the atmosphere and wind speed, placement of the buoy, and geographical location (longitude/latitude). In the North Atlantic, wave heights are extremely difficult to predict from the variability of meteorological states that arises between seasons. Due to large scale storms of varying degrees, the time series values of the recorded wave heights may be highly inconsistent during the fall and winter months.

Canada has put extensive efforts into collecting marine data. The buoys placed along the coast have provided valuable data made available for mariners, fishermen, surfers, and other recreational ocean users. The weather forecasts however, can be inaccurate, and most of these people simply rely on various websites which forecast the wave heights in Nova Scotia without disclosing their prediction model source code. Thus, individuals are not given the specific forecasting methods that interpret the raw data collected from weather stations, such as the Halifax harbour buoy. Therefore, if we can come up with a model to improve the accuracy of these forecasts, and predict when such dangerous storm conditions will occur, we can help fishermen avoid going out to sea in hazardous conditions that were not accurately forecasted, as well improve the safety of beach goers in general. In addition, having better prediction for large wave events can help people prepare for the infrastructure impacts such as erosion to roads and power lines.

For the analysis, we aim to answer the following questions:

1. **Based on previous data collected, are we able to determine whether or not the waves are getting larger or more consistent over the past 20 years?**

2. **Are the ocean temperatures in Atlantic Canada increasing due to changes in the Gulf Stream?**

3. **What month of the year produces the best quality surf for the Halifax area (waves, wind, temp)**

4. **Can we accurately predict wave heights using random forest methods on the collected buoy data?**

Using the time series wave data from *MEDS - Fisheries and Oceans Canada* , we have extracted 4 buoy datasets: Banquereau Bank, East Scotian Slope, La Have Bank, and the Halifax Harbour. (Fisheries and Oceans Canada & Government of Canada, 2020) A large portion of this report will be dedicated to data manipulation/data cleaning, since there are large inconsistencies with the data recordings. Modelling with random forests will be a useful tool for model building, as well as model predictions, due to the robustness of the method.

# 3 Data

## 3.1 Definition of Variables

The following definitions were extracted directly from the *MEDS - Fisheries and Oceans Canada* website. Since the data extracted were raw, a lot of data cleaning was required. This process as well as the compilation of the different data were done using the `dplyr` package. All of the data have the same variables, therefore when compiled, had an indicator index of the location to ensure that the data corresponded to its respective buoys. (._b for Banquereau, ._l for La Have, ._e for East Scotia, and ._h for Halifax)

Initially, Banquereau data had 184,756 observations, La Have data had 127,104 observations, East Scotia had 211,064 observations, and Halifax had 133,286 observations.

Table 1: Variable definitions

| Variable | Definition |
|---|---|
| Quality Control Flags | Based on IGOSS quality codes |
| Date Time Format | mm/dd/yyyy hh:mm:UTC |
| Latitude/Longitude | Location of buoy. West and North of the equator are positive values |
| Depth | Depth of water below the buoy (m) |
| VWH | Vertical Wave Height (reported by the buoy) (m) |
| VCMX | Maximum Zero Crossing Wave Height (reported by the buoy) (m) |
| VTP | Wave Spectrum Peak Period (reported by the buoy) (m) |
| WDIR | Wind Direction (° True) |
| WSPD | Horizontal Wind Speed (m/s) |
| GSPD | Gust Wind Speed (m/s) |
| ATMS | Atmospheric Pressure at Sea Level (mbar) |
| DRYT | Dry Bulb Temperature (° celcius) |
| SSTP | Sea Surface Temperature (° celcius) |

Each definition of the IGOSS quality codes can be seen in the appendix. (Table 3)

## 3.2 Initial Data Processing

As mentioned above, raw data for the 4 buoys placed along the coast of Nova Scotia were extracted from the *MEDS - Fisheries and Oceans Canada* website. The buoys chosen for this analysis were the Banquereau, East Scotian Slope, La Have, and the Halifax buoys. These buoys were specifically chosen based on their location along the coast, which can be seen in the map below:

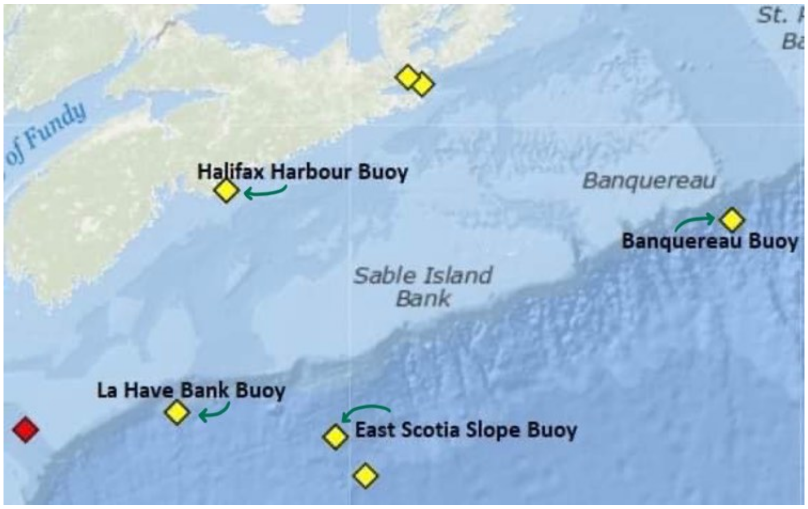


Figure 1: Map of buoys along the coast of Nova Scotia.

Each dataset went through the same data cleaning process. Each column was checked for NA values, and if the column consisted entirely of NAs, it was removed. Then, specific columns were analysed. When looking at WDIR, WSPD, and GSPD, there were two columns measuring the same covariate, which were concluded to be from the buoy having two different devices to measure. For example, these were indicated by WDIR and WDIR.1. If there was no measurement for WDIR but one for WDIR.1, it replaced the non measured WDIR value, and vice versa. After both columns were filled, the average of the two were calculated to be used in the analysis. Then, WDIR and WDIR.1 were removed from the data and replaced. The ATMS variable also had a secondary reading, ATMS.1. However, there were no NA values in either column so the two columns were averaged like the other three variables, and the average column was kept.

After going through the 4 separate initial data cleaning processes, the full data was combined using the `inner_join` function, using the common DATE column. From the combined full data, the DATE column was separated into DAY, MONTH, and YEAR.

## 3.3 Data Quality Check

After the full data was created, a data quality check was performed using data visualization plots.

Initial data quality check began with the sea surface temperature measurements. For other data quality checks, see `https://github.com/calipark1213/Stat4620---Project`. As we see in figure 2 below, for the

Halifax harbour buoy, there were massive recording errors in 2007 and 2008, as well as 2013-2016.

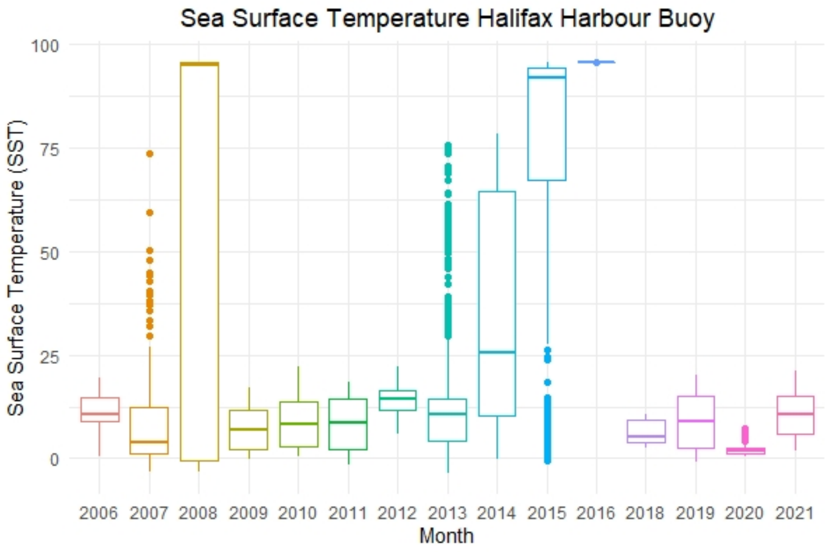


Figure 2: Sea Surface Temperatures (2006 - 2021) measured in degrees celcius.

In addition, there seems to be no recorded data in 2017. Sea surface temperatures in Nova Scotia are impossible to reach as high as 100 degrees celcius, therefore these values were identified as recording malfunctions by the buoy. A reason for these high recordings may be due to the temperature sensor on the buoy switching from measuring in celcius to fahrenheit. To fix these temperatures, the outliers were partitioned into a separate data frame. After, the other outliers after removal were removed based on data from `seatemperature.org` (Global Sea Temperatures - A-Connect Ltd, 2021).

Similar temperature recording issues were shown in the East Slope buoy, so the same procedure of inspecting the temperature anomalies were conducted. As well, the sea surface temperatures for the other two buoys were inspected, as the variable was misrecorded frequently.

The next variable that was inspected for outliers was the vertical wave height variable. Again, the steps below were conducted on the Halifax buoy, and the other buoys were inspected in a similar manner. In 2011 and 2013, it seems that the wave height sensor in the buoy did not record the wave heights correctly. The values selected here were turned into NAs, and were imputed later using the `mice` package, with the predictive mean matching method.

The dataset being potentially biased was not an issue, as it was sourced from the Canadian government, and as the raw data reported from a buoy is incredibly unlikely to be biased.

## 3.4 Final Data

Across all 4 datasets, there were a total of 23 variables, which were combined and collected into one dataset called ”fullimp.RData”, resulting in 64,933 observations. The variables included in our model after data cleaning are defined in table 1.

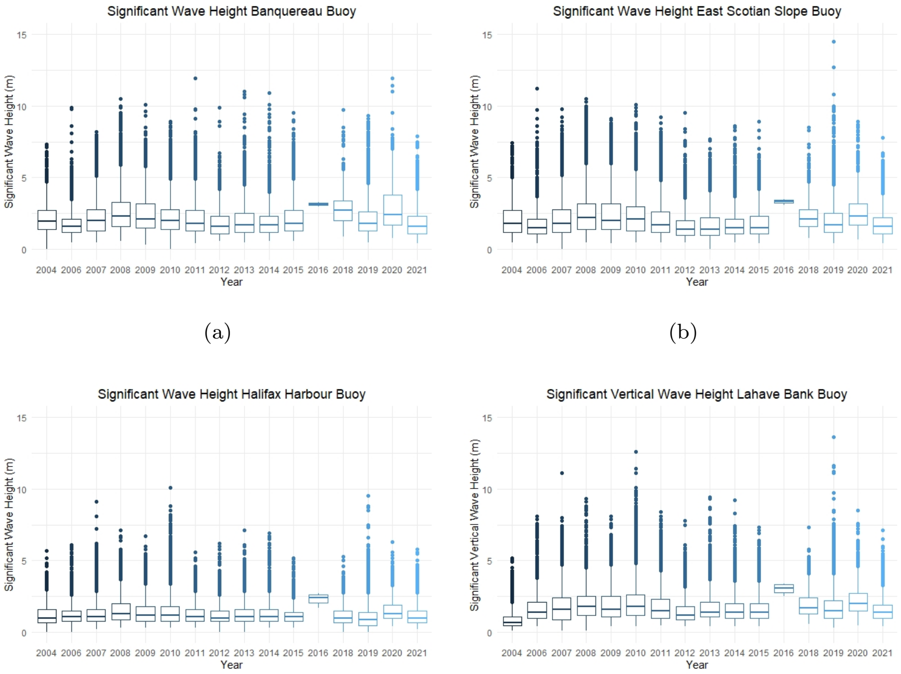
# 4 Methods

## 4.1 Scientific Questions

The first 4 scientific questions can be answered using visualizations.

### 4.1.1 Question 1

First, let’s explore whether the waves are getting larger over the past 20 years of historical data.



(a)

(b)

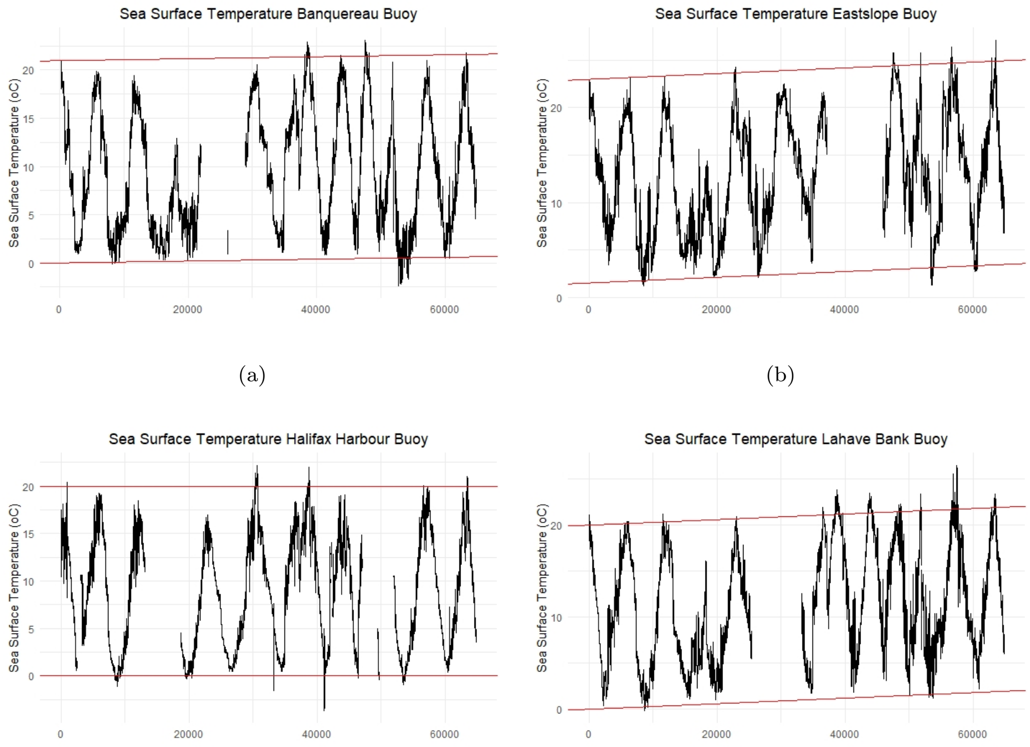
(c)

(d)

Figure 3: Boxplots of VWH measurements from 2004 - 2021.

### 4.1.2 Question 2

To assess if the Gulf stream is changing, the sea surface temperatures for the buoys over the entire time series data were analysed.



(a)

(b)

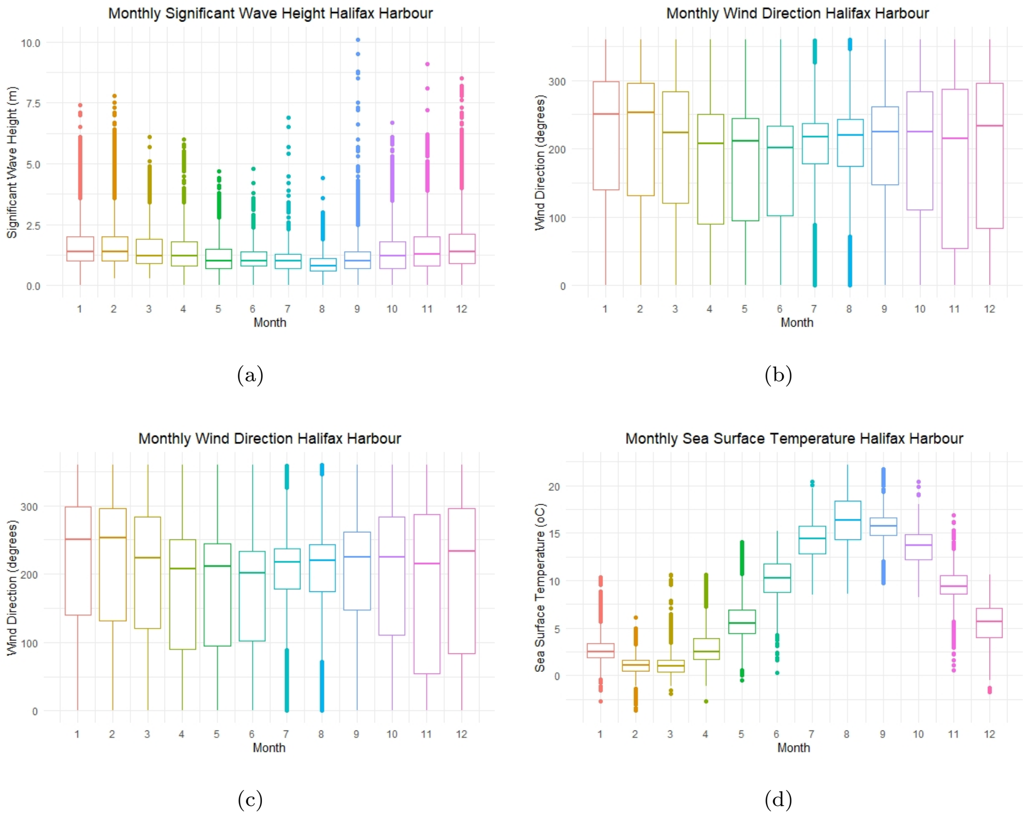
(c)

(d)

Figure 4: Time series plots of the Sea Surface Temperatures.

### 4.1.3 Question 3

To determine which months of the year produces the best quality surf for the Halifax area, the wind, wave height, and temperature variables were analysed.



(a)

(b)

(c)

(d)

Figure 5: Monthly variable plots for Halifax Harbour buoy.

### 4.1.4 Question 4

To assess whether the wave heights can be accurately analysed, a random forest model was fitted to the data for this regression problem. To account for the auto-correlation that arises from the time series data, the training, validation, and testing data were created by splitting the final fully imputed data into chunks. This idea is similar to the block bootstrapping method commonly used for time series data.

The random forest algorithm is a very useful tool for this research, as it has the ability to model both categorical and continuous variables. The random forest model was fitted using the `Rborist` package. This package allows for a high computational performance compared to other packages. (Seligman, 2019) Then, the model was tuned by performing a grid search along the hyperparameters. The performance of the model fitted to the training set were measured on the validation and the test set via mean squared error.

# 5 Results

## 5.1 Scientific Questions

### 5.1.1 Question 1

Without taking the obvious outliers in 2016 into consideration, the waves do not seem to be getting larger over the past 20 years. However, 2020 was certain a significantly an above average year compared to recent years. As mentioned before, there are missing data from 2016 and 2017 when the Halifax Buoy went offline.

### 5.1.2 Question 2

The 4 time series we have plotted begin in 2004 ending in January 2022. All 4 buoys contained recording errors during the span and had to be removed leading to gaps in the sea surface temperature readings. Therefore, accurately assessing a trend in the sea surface temperatures for the 4 buoys was not feasible. However, through visual inspection, there is a clear trend in the SST as a result of the Gulf stream. For the Halifax harbour buoy, there has not been a significant increase of temperature over the previous years. However, for the East Slope and La Have buoy, there is a noticeable upward trend in the minimum and maximum sea surface temperatures (in red) towards the end of the recorded data. We suspect that this fluctuation in temperature may be due to potential impacts from climate change patterns. To validate this inference, cross referencing the sea surface temperature trend for the East Slope buoy can be conducted, since it is not too far away. While the sea surface temperatures observed at the East Slope buoy are warmer than the La Have Bank Buoy due to being further from the North American continent the noticeable upward trend is also present with the same slope. This offers significant proof of the Gulf stream increasing sea surface temperatures off of the coast of Atlantic Canada.

### 5.1.3 Question 3

A good surfing criterion satisfies the following conditions: VWH larger than 1 meter, wave period above 8 seconds (which means that the waves are organized and have more power.

Looking at the vertical wave heights plot of the data, the months December through February produce the largest waves on average for Halifax, because they all had significant mean vertical wave heights over 1.25

metres with similar number of large wave event outliers reaching heights of up to around 8 metres. However, in terms of the largest possible waves, September has produced the largest wave event outliers namely due to hurricane season peaking during the month. In terms of the worst month of year for waves, August has a mean vertical wave height of below 1 metre with very few large outliers. This indicates that Halifax does not get substantial swell from hurricanes until September and the peak of hurricane season where the sea surface temperatures are at their max leading to more powerful storms that move up further into Atlantic Canada. In summary solely based on wave height data over a month, the three winter months of December, January, and February really end up tying backing up the claim of ”the Winter is the best” by experienced Nova Scotian surfers.

Next, the plot of the wind directions by month was analysed. While the wind direction plot by month may seem confusing at first glance, upon further inspection, the months of January and February have means of over 250 which corresponds to roughly the WSW wind direction and are closest to the ideal wind conditions of North West to North offshore wind. This is a result of the polar jet coming down from the arctic as it sweeps across Canada causing strong North West winds and very cold air temperatures. However, as soon as spring begins we can see that the wind direction moves to just above 200 degrees or approximately South-South West on a compass creating unfavourable wind conditions for surfing. Then when we reach the summer, in July and August the variance in wind directions substantially decrease as the coastal storm activity reaches a minimum thus not causing rapid changes in wind direction as it moves through. This is also indicative of poor surfing conditions as there is no South-West, South, or South-East winds creating waves in the first place. Overall though, the months of January and February are the best months for ideal wind direction conditions followed by the fall months.

Similar to the vertical wave height, the wave spectrum peak period reaches a max mean of just under 10 seconds during the winter month of February. While the months December through March are quite similar in wave period indicating that the winter is the best for longer period waves. September has significantly larger quantiles indicating that while September has lower average wave period, during a hurricane swell there can be much longer wave period. It is not uncommon to experience a wave period of over 17 seconds during a good hurricane swell in September. Therefore, the best month for quality waves in terms of period is really either February or September depending on if Atlantic Basin is in an active hurricane year or if Nova Scotia is being impacted by a large number of Nor'Easters during the winter.

Based on the data from the Halifax Buoy, it is clear that the sea surface temperature peaks during the months August or September around 16-18 Celsius, the first month of which coincides with the smallest wave heights experienced in the entire year. Therefore, while the warm sea surface temperatures are ideal for 4/3 millimeter wetsuits and no gloves and boots, August is certainly not good. However, September benefits from the influx of hurricane swells which have significant wave height and period while still enjoying the warm sea surface temperatures courtesy of the Gulf Stream. On the other hand, from December through April, the sea surface temps are extremely cold even reaching negative values for short periods during January and February. This makes surfing in the winter in Nova Scotia incredibly challenging as 6/5 mm hooded suits, 8mm gloves, and boots are required to survive the freezing temperatures not to even mention the much larger surf experienced along the coast due to the Nor'Easters.

In summary, February produces the best conditions for surfing on average for the Halifax region based on mean wave height of over 1.3 metres, mean wave period of almost 10 seconds, and average WSW wind direction. However, the main drawback of February for surfing is unsurprisingly the average sea surface temperature of 1 Celsius making hypothermia a real risk and surf sessions longer than 2 hours almost impossible. However, depending on a surfer's experience level, the quality of wetsuit they have, and their cold threshold, February can either be the best or worst month. For less hardcore surfers September produces decent surf, while it may be less consistent and smaller than the winter months, it is also much warmer allowing for long 4-5 hour surf sessions with only a 4/3 mm wetsuit. While we may have not objectively answered the question of which month is the best for surfing in the Halifax area for all skill levels, the question is subjective by nature due to the extreme weather we experience and the different thresholds individuals have for the cold. If a person is a experienced surfer and does does not mind the extremely cold temperatures of winter, February is the best month. However, if a person is less experienced or does not like the cold, September is the best month.

### 5.1.4 Question 4

To assess whether the wave heights can be accurately analysed, a random forest model was fitted to the data. To account for the auto-correlation that arises from the time series data, the training set included the first 45000 observations from the fully imputed data frame, the validation set included observations from 45000 - 55000, and the testing set included the 55000 - 64933 observations.

The first model computed was to fit a random forest model for the vertical wave height as a function of the rest of the covariates in the model. Using this, the model hyperparameter selection across the grid search was conducted to look through, and a new model was fitted using the best parameters from the hyper grid.

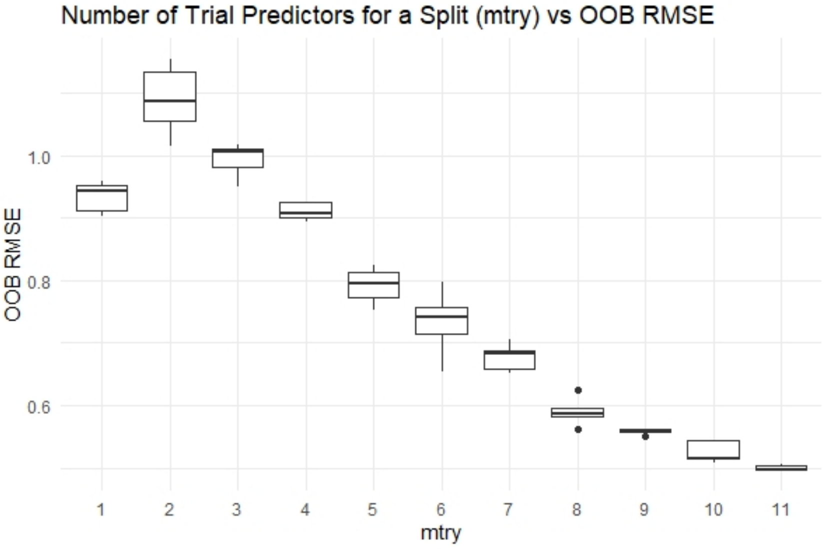


Figure 6: Number of trial predictors for a split VS OOB RMSE value.

From this plot comparing the number of trial predictors (mtry) for a split VS the out-of-bag root mean squared error values (OOB RMSE), there is a clear relationship as the number of mtry increases,the OOB RMSE value decreases substantially. Looking at the mtry equal to 9, 10, and 11, there is very little variability in the OOB RMSE rate, although the sizes are different, therefore, mtry = 11 was chosen. The best minimum node size was selected to be 6, achieving a OOB RMSE value of 0.4962. Then, the final (tuned) selected model was then evaluated on the testing set, to observe the predictive performance.

Furthermore, the performance of our tuned random forest model on the training, validation, and test set were compared using the mean squared error values, and the variable importance plot of the chosen RF model was plotted. The predicted vs observed values plot of the tuned random forest plot seems to follow a linear trend, deviating slightly towards the high tails.

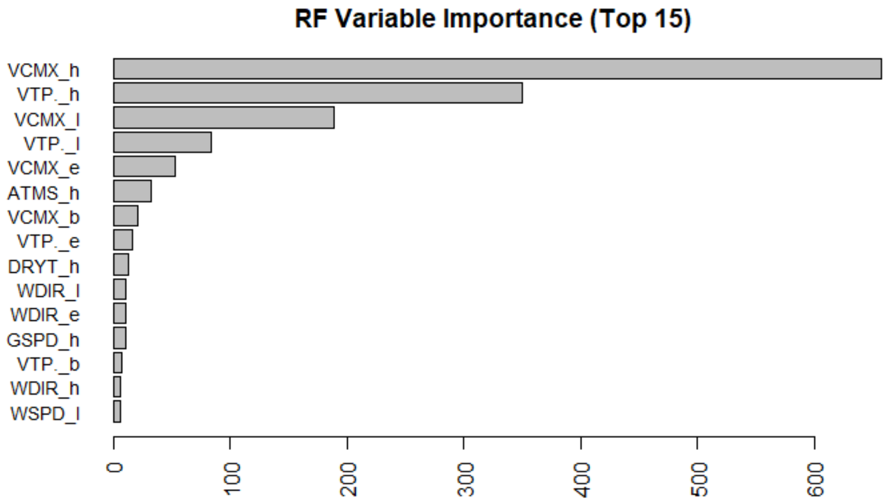
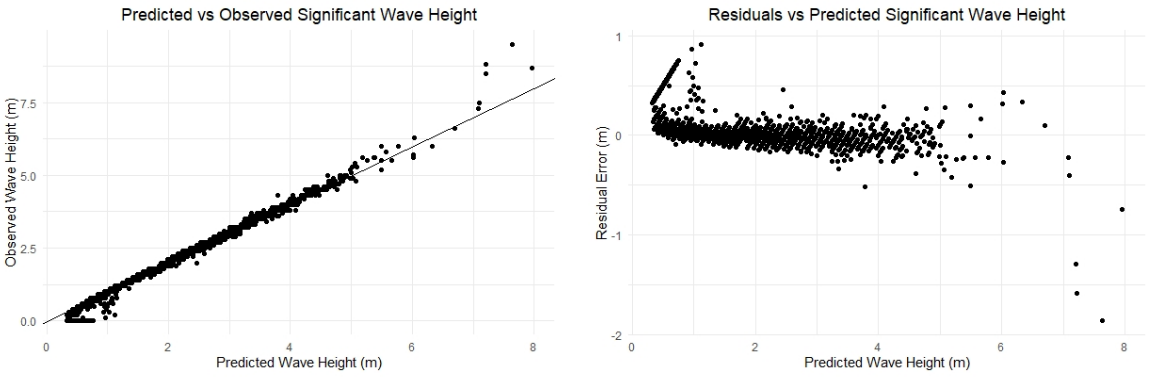


Figure 7: Variable Importance plot of the RF model.

Table 2: Mean squared error values for each dataset

| Data | MSE |
|---|---|
| Training | 0.0023 |
| Validation | 0.0035 |
| Testing | 0.0310 |

Additionally, the residual plots of the tuned random forest model was plotted:



(a)

(b)

Figure 8: Left: Plot of observed vs predicted from tuned RF model. Right: Plot of residuals vs predicted from the tuned RF model.

From these plots, notice that the residuals of the predictions are densely located around 0 and has a small number of outliers as VWH becomes larger than 6-7.5 metres. The residuals also seem to have a tail like pattern as significant wave heights approach 0.

# 6 Conclusion

In conclusion from the visualizations of the data, the waves do not seem to be increasing in height over the past 20 years. By comparing the time series plots of sea surface temperatures recorded from the 4 different plots, the significance of the Gulf stream influencing the increase of the sea surface temperatures off of the coast of Atlantic Canada was confirmed. From plot 5, February was deemed to produce the best conditions for surfing on average for the Halifax region. However, for a less experienced individual, September was deemed to be the best month. Finally, the fine tuned random forest model fitted by tuning the hyperparameters via a grid search performed tremendously well. The MSE value of the training, validation, and testing set were evaluated to be 0.0023, 0.0035, and 0.0310, respectively.

# 7 Acknowledgements

This report aims to satisfy the STAT4620 Data Analysis Final Project component. For this report, the work was evenly divided. Throughout the semester, both Alex and Cali met together to work on the analyses. Most of the time spent initially were dedicated to cleaning the data, and since the number of observations were large, Alex's PC was used. As well, the random forest model for the analyses was conducted on the PC. The report created by Cali on `www.overleaf.com`. Cali also created the project repository on `www.github.com`

# 8 Appendix

The full analyses can be found on `https://github.com/calipark1213/Stat4620---Project`. (Park, 2022)

Table 3: Quality Control Flags

| Flag number | Definition |
|---|---|
| 0 | Blank: No quality control has been performed |
| 1 | Good: QC has been performed: record appears correct |
| 3 | Doubtful: QC has been performed: record appears doubtful |
| 4 | Erroneous: QC has been performed: record appears erroneous |
| 5 | Changes: The record has been changed as a result of QC |
| 6 | Acceptable: QC has been performed: record seems inconsistent with other records. |
| 7 | Off Position: There is a problem with the buoy position or mooring. Data may still be useful |
| 8 | Reserved |
| 9 | Reserved: Indicates missing elements |

# References

Fisheries and Oceans Canada, & Government of Canada. (2020). *MEDS format descriptions.* Retrieved from https://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/waves-vagues/formats-eng.html

Global Sea Temperatures - A-Connect Ltd. (2021). *Halifax water temperature: Canada sea temperatures.* Retrieved from https://www.seatemperature.org/north-america/canada/halifax.htm

Park, S. (2022). *Stat4620 project repository.* Github. Retrieved from https://github.com/calipark1213/Stat4620---Project

Seligman, M. (2019). *Package rborist.* Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/web/packages/Rborist/index.html