Big Data Processing and Applications

Project document template

Trend analysis of Ireland tidal data

Matthew Utti	2305008
Rukshan Perera	2304805
Aleksi Patronen	Y5802907

Project description

Most of the important infrastructure and population in Ireland is concentrated around coastal areas and coastal storms which lead to coastal flooding - This poses a serious risk for them, and the sea water level is an important measurement that helps to identify coastal flooding. This project focuses on collecting key features like temperature, and tidal data around various coastal points in Ireland, analysing and identify significant trends in the sea water level changes over the years. Project goals are to describe sea level rise around the Isle of Ireland.

Related work

Ireland's digital ocean website already has several predictive models around water levels to provide water level data up to a day in advance. There has been a memorandum of up-to-date analysis of water level changes until 2018. Nejad et al. [6] found that the sea-level rise between 2003 to 2015 was some 10mm per annum at Dublin Port. Smaller figures were recorded over the course of longer periods, for instance, 1.67mm/year between 1938 and 2015 European Environmental Agency has suggested a figure of 1.7mm/year from 1900 to 2020 in global sea levels[13]- This would suggest an acceleration of the rise in sea level. Nejad et al. also found that the sea-level rise was cyclic, meaning that there are periods where the sea levels are decreasing rather than increasing, but overall, the sea level maintains a progressive increase. The study that Nejad et al. conducted was restricted to 60 km from the Dublin city port. P. L. Woodworth et al. [8], who studied similar sea level patterns around the British Isles, found that the acceleration of the sea-level rise in the 20th century is from 0.4 to 0.8 mm/year/century. This means that for the 20th century, the increase in sea level is from 0.4 to 0.8 more than in the previous year - This has a compounding effect, meaning that the sea-level rise is not linear. Reasons for sea-level rise are innumerable, but this is mainly due to thermal expansion and melting glaciers. Oceans have absorbed more than 90% percent of the excess energy caused by greenhouse gases. Sea-level rise can be used to make projections for the future according to the UK Environmental Agency in ‘Exploratory Sea Level projections. It was projected that, depending on carbon levels, the mean sea-level would rise between 0.6-2.2 meters in a low-carbon atmosphere and 1.7-4.5 meters in a high-carbon environment. World Economic Forum [12] have reported that the global sea temperatures are some 0.69 Celsius higher than century’s average. This would yield an average increase of temperature of 0.015 centigrade per year, when starting from the year 1977.

Data description

Data itself has been provided to the Marine Institute by various providers - It has been gathered from the Ireland Marine Institute website [1] where it is available for public consumption under the Creative Commons BY 4.0 license. This license lets others distribute, tweak, and build upon previous work, even commercially, as long as they credit Ireland’s Marine Institute for the original creation [1, 2].

The data used comprises of features such as timestamp, station ID, latitude, longitude, water level LAT, and water level OD Malin. The lowest Astronomical Tide (LAT) is defined as the lowest level that can occur due to the astronomical tide [11]. OD Malin is Head Vertical Datum which is the mean sea level of the tide gauge at Malin Head, County Donegal [11]. Hence, there's a fixed point to which the actual sea level is compared. There were in total 21 measurement stations in the data since 2007 and Some stations had recordings for only from 2021 until date. The data is downloaded in the form of comma-separated values and stored in the Rahtiapp server. The size of the data amounts to 1.6354 GB. The timestamp variables were modified into month and year readings based on some key values like minimum, maximum, standard deviation and mean values.

As we have many water level data collection stations below is an example data summary of the " Roonagh Pier" station and temperature from "Finnis Waver" station.

	Water Level (m)	Temperature
Min	0	8.83
Max	23	17.1
Std. Deviation	1.01875	2.4033
mean	2.4617	13.1127

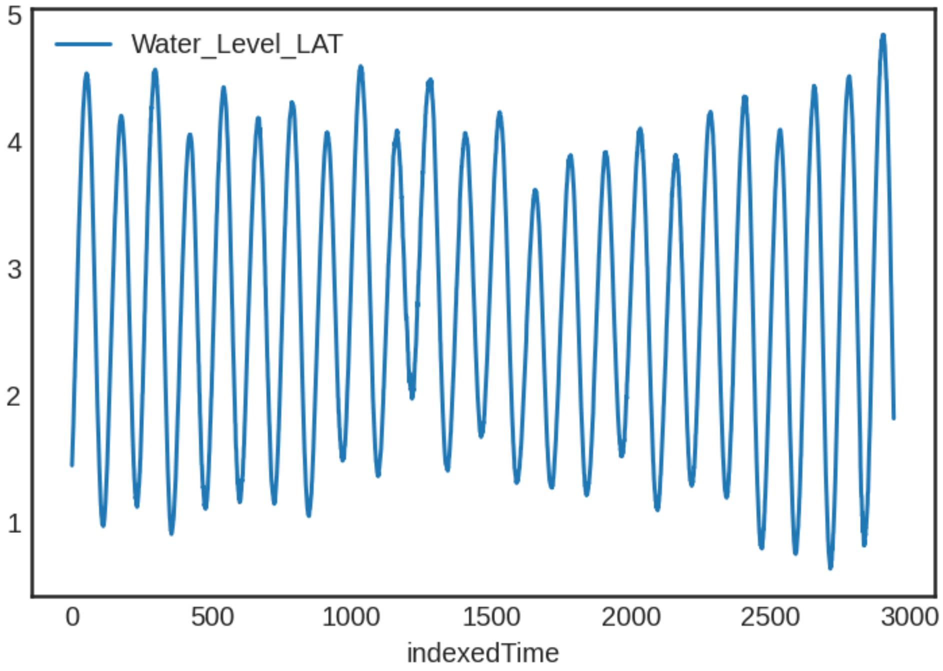

Figure 1 Water Level snapshot of some twelve days.

Figure 1 is a snapshot of one the Lowest Astronomical Tide of a station near Dublin. Time is indexed from zero to nearly 3000 representing a snapshot of a chosen 12-day period - It is demonstrated in the tidal action of the measurements. Overall, the water level rises and lowers, but some inherent movement exists where there's a decrease in tidal heights or an increase in tidal highs and lows.

Also, data about the temperature and wave patterns were downloaded from the Marine Institute website [7] - It consists of sea temperature measured in Celsius every five minutes, 9 stations, of which only one had a long enough operational history to be considered useful. The useful data from the station at Smart Bay started in September 2013. Other stations had data only for a couple of years or less which means that, other than Smart Bay data, they were not considered sufficient for trend analysis. The data's size totals into 0.71 GB.

Data selection for correlation analysis

Using latitude and longitude data by calculating the closest Buoy station which provides temperature data with sufficient row counts is used with each water level station. Since some Bouy points lack data points one station with maximum data points matched for these locations.

All the data and code for analysis can be found in the RahtiApp server. All the codes reside in notebooks and python functions we created are also written in the notebooks itself.

Methods and tools

Methods for the trend analysis are descriptive in nature. It includes statistics like mean, maximum, minimum, and standard deviation. To assess the changing tidal conditions on the Island of Ireland time of the year must be considered. That is, each statistic was calculated based on the month. **Approximated** data is used sometimes. Some approximated calculations were done based on figures visually. Meaning that, the approximate values were in some cases lifted out of the presented figures. Initial data cleaning is done in Spark. **Aggregations** like monthly and yearly max and average aggregations are done using both spark and spark sql. **Outlier** removal is done based on standard deviation to remove data too far from the mean value. For the purpose of visualization, data frames are converted to pandas data frames and seaborn and Matplotlib are used for visualization.

To find the correlation between the measured average temperatures and sea levels, there was a need to find the right stations to compare. Since the stations that measured the temperature and tides are not the same, we needed a way to correlate the measurements. A nearest station method was used by calculating the Euclidean distance between stations. A distance matrix was created, where the distance between different stations was calculated, and a shorter distance was chosen for pair creation. The correlation matrix was Pearson correlation. Pearson correlation ranks how well the observations correlate on a linear basis. As it was mentioned, oceans absorb energy from atmosphere, so the sea level swells up.

For trend analysis ordinary least squares linear regression was chosen. Monthly averages of water level LATs were calculated and the months themselves were indexed to an integer type. In linear regression, the outlier threshold was three times the standard deviation. Any measurement outside of that was considered an outlier. Linear regression was fitted on each station and the regression coefficients were averaged.

To see, which machine learning algorithm could be suitable for this kind of data. Linear regression and decision tree algorithms were pitted against each other.

Data analysis

This section covers, details, and describes the results. First, we describe the trends in water level, then we describe the trends in water temperatures and linear regression and the correlation results.

In the picture below is the average water level at one of the measurement stations. In this case, it is Galway Port. The records started in July 2005 and ended in December 2022. The water average yearly water level has increased by 25 centimetres in that timeframe. That makes an annual increase of 16.67mm/year.

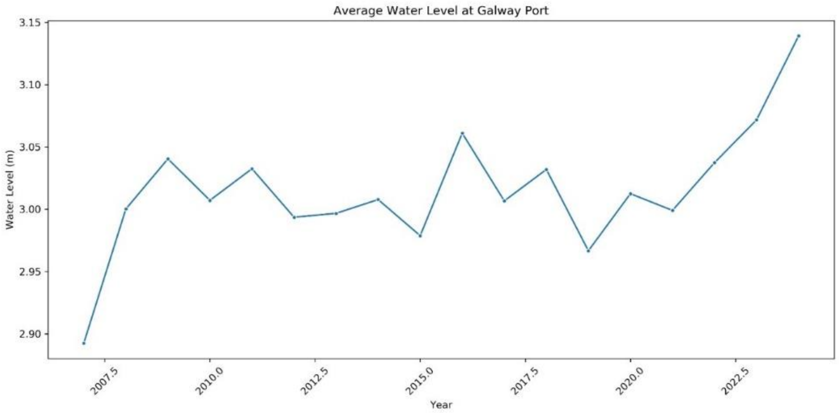

Figure 2 Average water level LAT at Galway Port yearly.

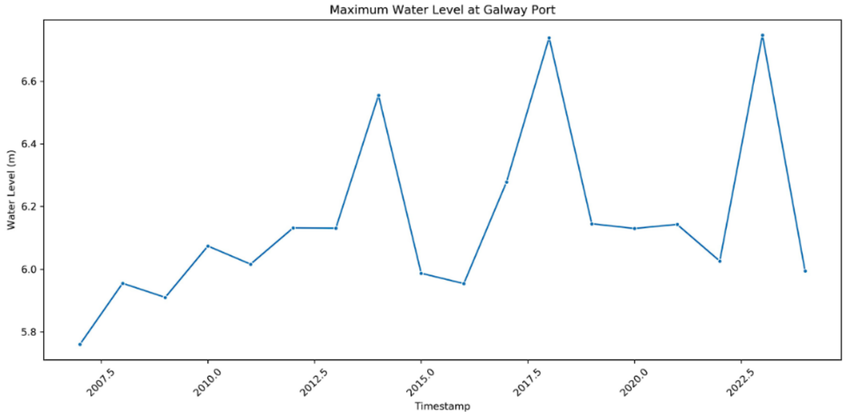

Figure 3 Maximum water level LAT at Galway Port yearly.

Figures 2 and 4 depict average water levels yearly at two different measurement stations. Figure 4 clearly shows that the time span is only some six years as the average water level rose. Also, Figure 3 displays the yearly maximum water level.

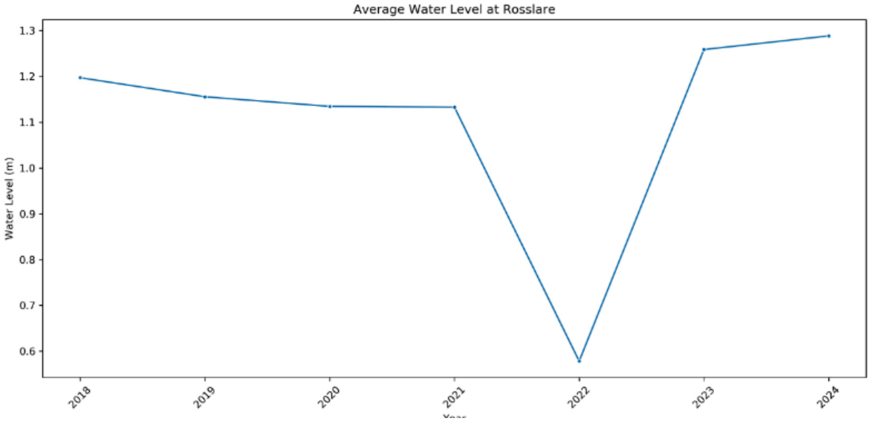

Figure 4 Average water level at Rosslare yearly.

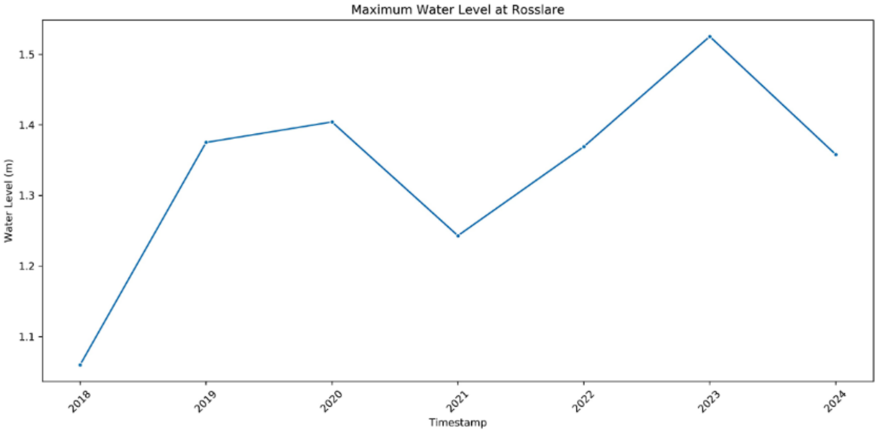

Figure 5 Maximum water level LAT at Rosslare yearly

From the graph below, figure 6, it’s obvious that the mid-point of the year reading has lower sea water level than rest of the year. This trend is seen throughout the year.

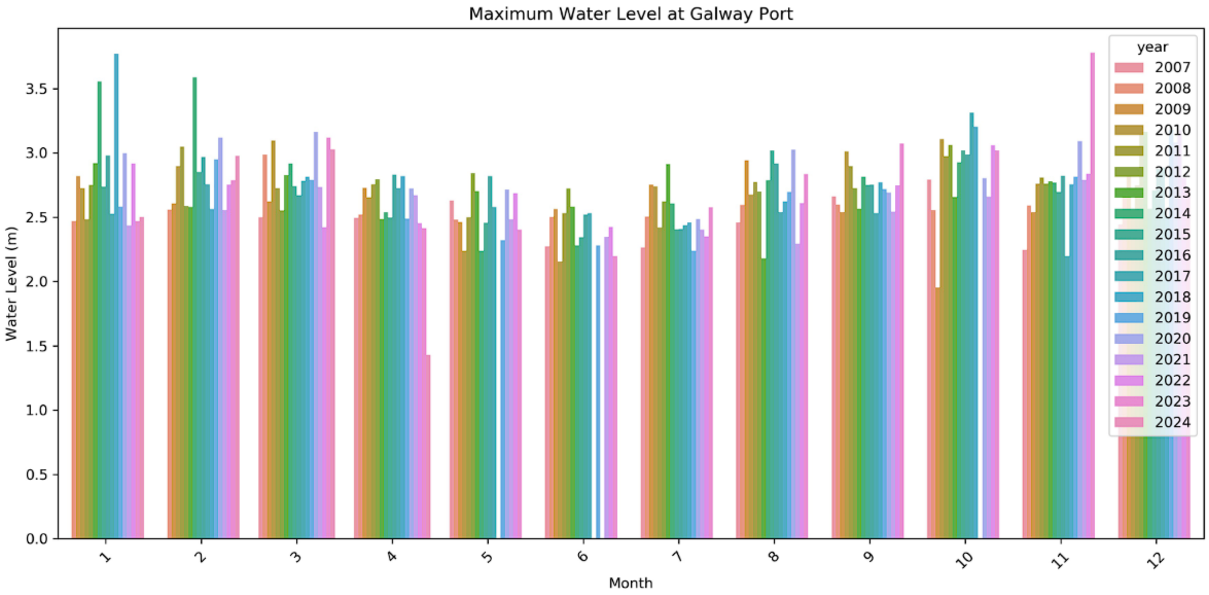

Figure 6 Maximum Water Levels at Galway Port by month from 2007 to 2024

Also, in figure 6 it is observed that there’s a characteristic wave pattern to water levels in years. In each month column, smaller columns represent the water level in years. In 2007, water levels were lower than in 2009 to 2012, when water levels started to decline again, then rejected lower levels again in years 2013 to 2017 - This, same wave pattern is also observed in figure 3.

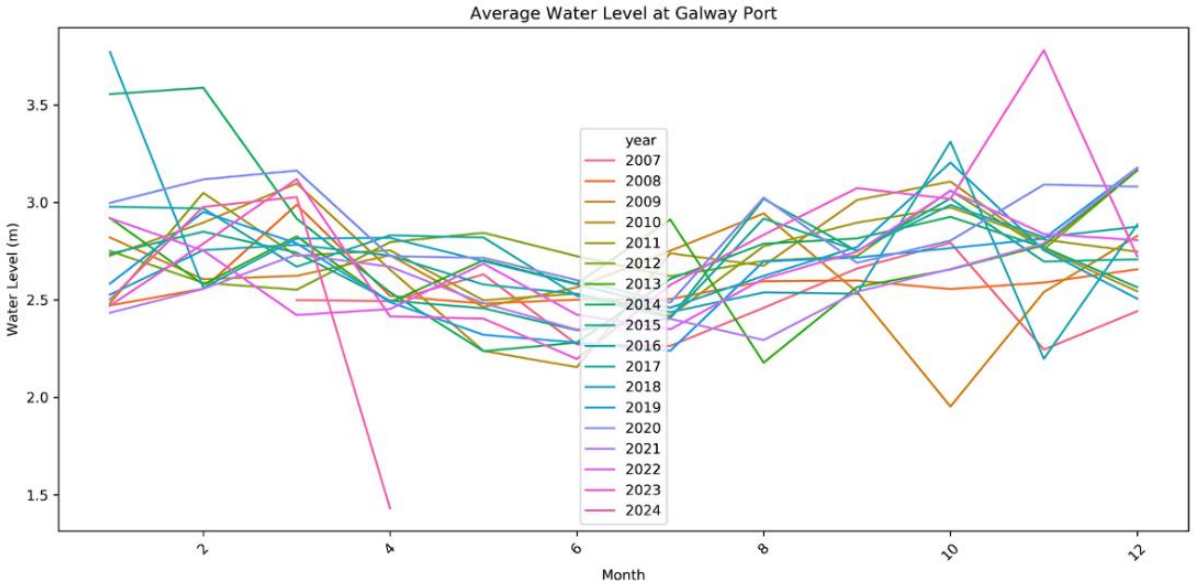

Figure 7 Average Water Level at Galway Port Line plot

According to histogram figures of Roonah Pier point we can see that water levels are highly distributed around 1-4 m range.

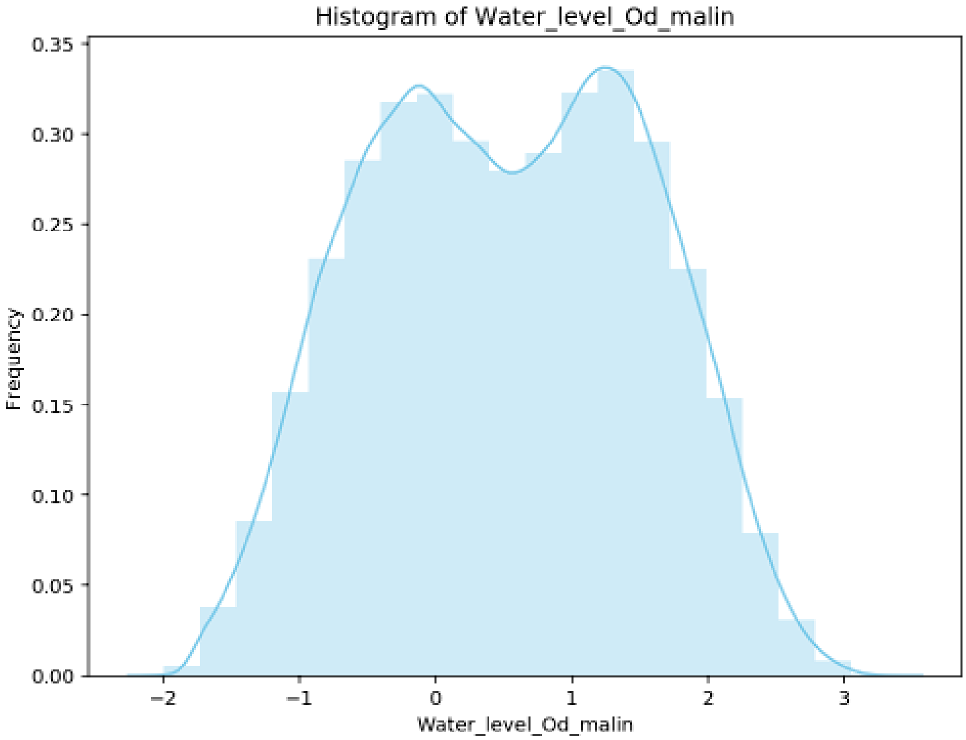

Figure 8 Distribution of OD Malin water level of Roonah Pier

The variations in plot heights for the 'year' distribution indicate fluctuations in the frequency of observations across the years 2020-2024.

The plot for 'water_level_od_main', figure 8, reveals a distribution that appears subtly uniform, albeit with a noticeable skew or asymmetry at the upper end, indicating a clustering of observations around the value of 0.32. The slight dip observed at the peak suggests a reduction in the frequency of observations. Around the peak value area, there is a clear correlation between the height of high tides and the depth of low tides, with a subsequent increase in frequency at higher or lower values. The presence of two distinct peaks corresponds to the high and low tide levels. Notably, there is an asymmetry between these peaks, with the high tide peak being taller and narrower compared to the low tide peak. This discrepancy likely stems from the fact that high tide occurs more consistently than low tide. Additionally, the shorter right tail of the distribution suggests a subtle upward trend.

Figure 9 presents a correlation heatmap for year and water levels. Excluding the obvious ones for variables correlating with themselves, there’s a slight positive correlation with water levels and year. This would suggest, as years go by water levels increase.

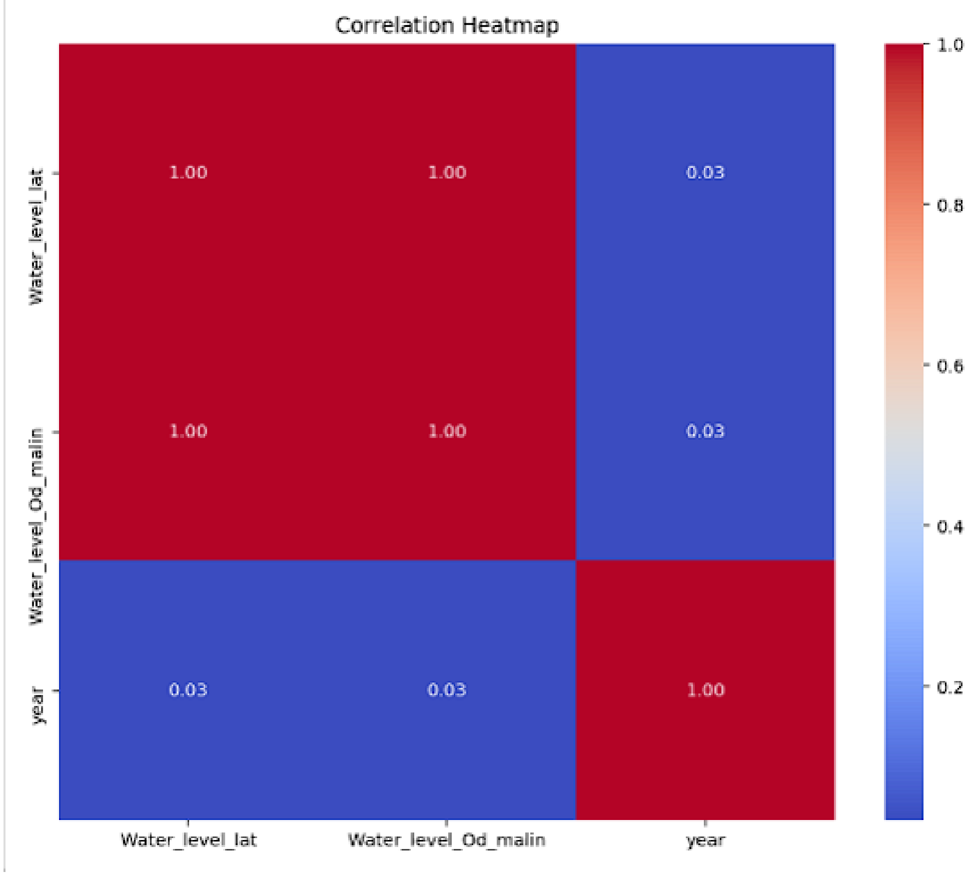

Figure 9 Correlation matrix on water levels and year

The heatmap plot above shows a positive inverse correlation between *water_level_lat* and *water_level_od_main*. The year is correlated slightly with the water levels.

From figure 10, below, one can see that despite the seasonal chances there's an upward trend. Data starts in September 2013 and reaches to April 8. 2024. The summertime peak has increasingly reached higher temperatures. In 2014 the higher average temperature in summer months was circa 16 degrees centigrade while in 2016, the summertime high average was increased to almost 16.5 degrees centigrade.

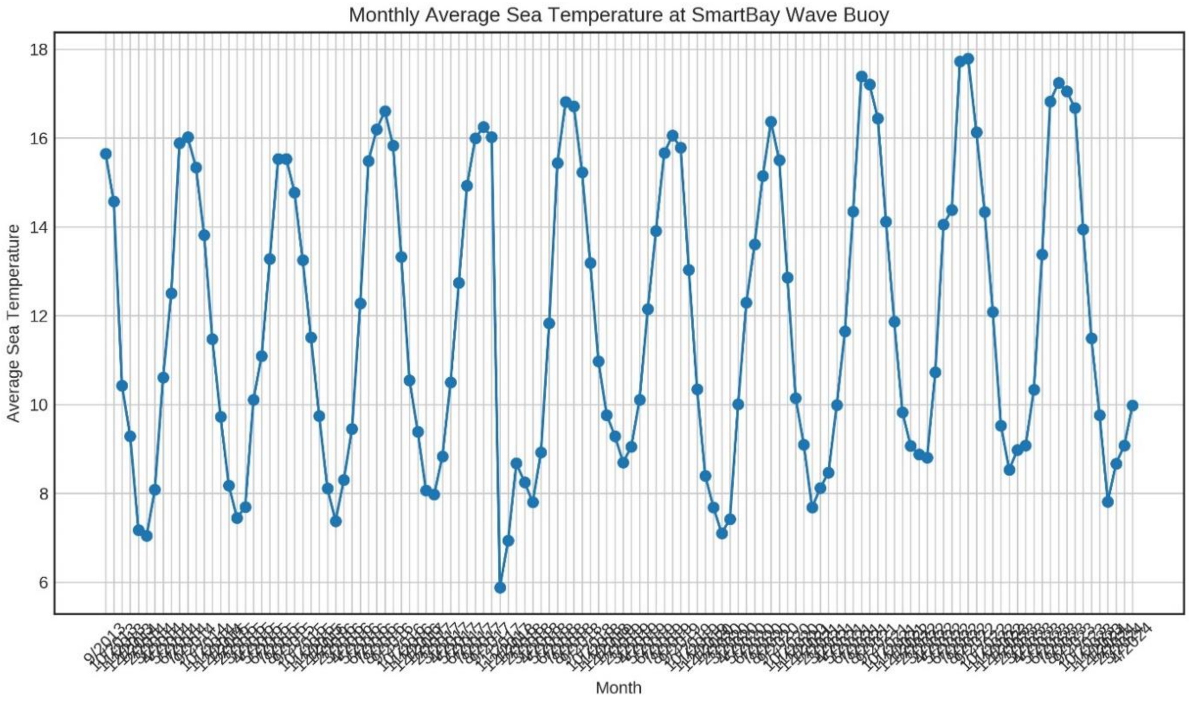

Figure 10 Monthly Average Sea Temp. At Smart Bay Wave Buoy

By 2022 the average, the temperature had increased to almost 18 degrees centigrade in the summer. This yielded an average of 0.25 Celsius centigrade increase in sea temperature per annum in average. The other stations had insufficient information for a similar review. As temperatures increases in the oceans, the water itself starts to expand due to thermal expansion.

Linear regression was conducted as described tools and methods section. On average, water levels had been rising by 10.899mm/year and the average water level at the beginning of the measurement period was 1.76m. In figure 12, one can see the obvious upward trend. Some measurements stations also show a negative trend. This is to be expected as there’s regional variation on the water levels.

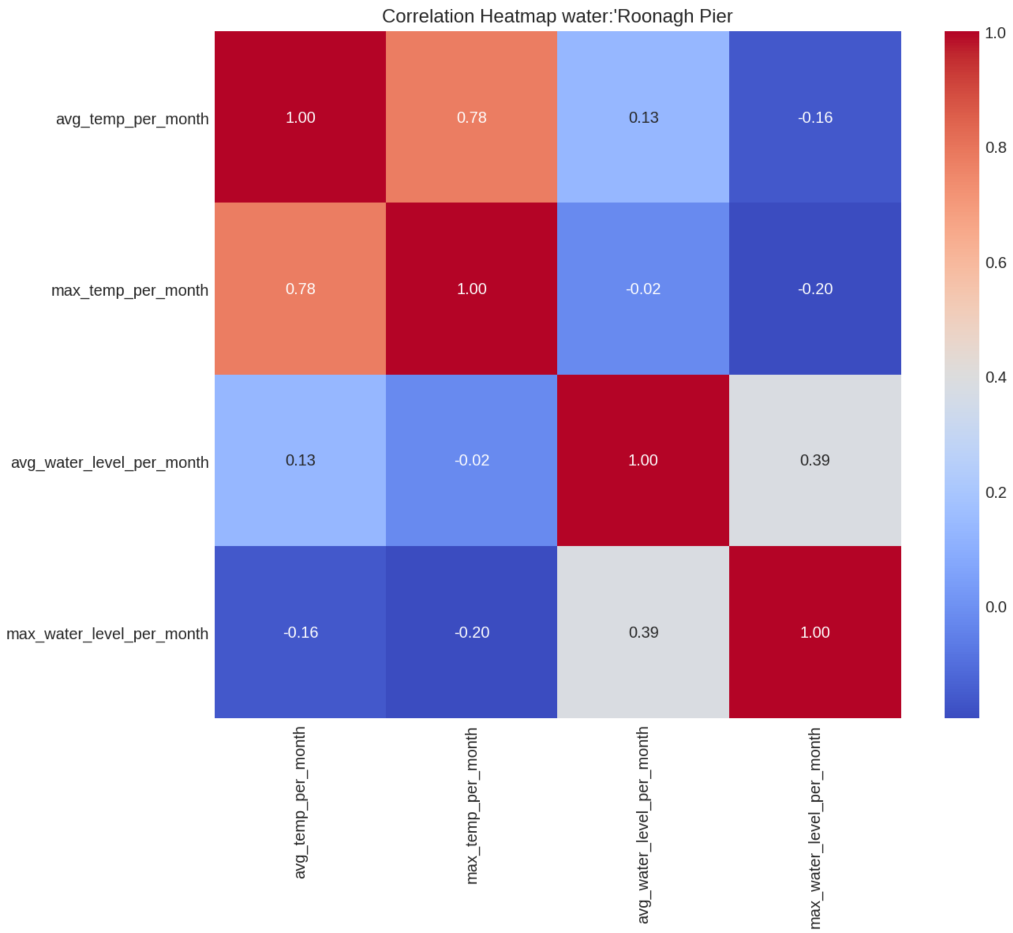

Figure 11 Correlation heatmap of avg./max sea temp. and avg./max sea levels

In figure 11 is a heatmap of the correlation measured at Roonah Pier and AMETS Berth C Wave Buoy at a distance of 52.89 km. This heatmaps is somewhat atypical of others were it shows a positive correlation average water levels and water temperatures. Most other measurements like this showed a negative correlation. We think this, somewhat surprising result, is only due to seasonal changes. As, it was seen on figure 7, ocean temperatures rise during the summer and the sea level lowers. This is due to evaporation.

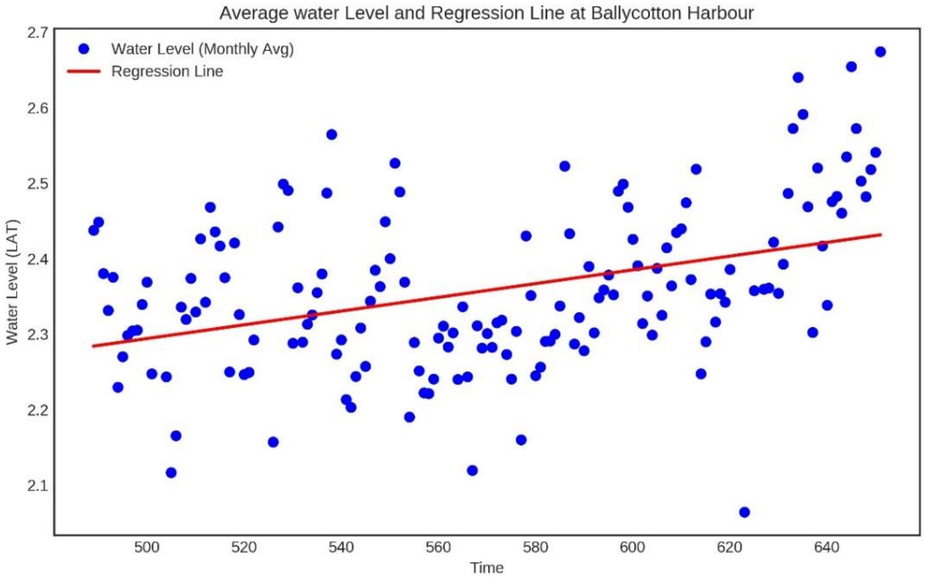

Figure 12 Linear Regression on Average Water Levels at Ballyvotton Harbour

Predictive Algorithms

In efforts to harnessing the robust capabilities of PySpark's machine learning algorithms, we probed into trends and relationships within the dataset through Linear regression and Decision Tree algorithms. With Linear Regression, Features are meticulously assembled using a VectorAssembler, and the dataset was put through a strategic partition into training and testing sets to predict water level measurements. The performance metric showed a Root Mean Squared Error (RMSE) of 1.0028, offering a quantitative glimpse into its predictive accuracy. Next, the Decision Tree Regression subsequently illuminates a superior predictive prowess as performance evaluation pitched the RMSE at 0.9447. Decision Tree Regression model's lower RMSE positions it as the preferred choice for applications requiring precise predictions and actionable insights in managing coastal flood risks, real-time flood monitoring, infrastructure management, and emergency response planning.

Conclusion

Our results on trend analysis are in line with the Nejad et al. results, who observed about 10mm increase in sea level per year. Ours was 10.9mm/year. Our conclusions about the sea temperature rise were 0.25 C/year, compared to the WEF’s 0,015C/year. Minding that both WEF and Nejad et al. come from a lot longer period and as the temperature increase accelerates. There was inherent challenge in the analysis, this was due to the tides. It was hard to find appropriate tools that could analyse tide data. Wavelet and Fourier transformations were considered but were left out due to excess complexity.

Also, to find a good predictive algorithm, decision tree stands out. This is due to the fact that sea level data exhibits nonlinear relationships with predictor variables such as time, temperature, and wind speed and Decision trees are inherently capable of capturing complex, nonlinear patterns in the data, making them suitable for modelling the intricate dynamics of sea level fluctuations.
Our station matching considered only the Euclidean distance but based on the geometry and using actual domain knowledge one must be able to match stations with good correlations. As an example, a large mountain range, coral leaf affects how wind and wave behaves. Our correlation data is not enough to say which station matching is best suited for model creation. In future, this matching can also be done using a model which analyses entropy for decision trees.

Also, for future exploration of this topic, spatio-temporal models should be considered. It would offer the benefit of linking each measurement station in time and space to each other. There are some data like wave direction, and wind direction like data which could contribute to such models.

References

[1] https://digitalocean.ie/
[2] https://www.climateireland.ie/impact-on-ireland/climate-hazards/coastal-flooding/
[3] https://creativecommons.org/licenses/by/4.0/legalcode
[4] https://www.digitalocean.ie/Data/DownloadTideData
[5] https://www.marine.ie/site-area/data-services/real-time-observations/tidal-observations-0
[6] https://os.copernicus.org/preprints/os-2020-81/
[7] https://erddap.marine.ie/erddap/tabledap/IWaveBNetwork.html
[8] P. L. Woodworth, M. N. Tsimplis, R. A. Flather, I. Shennan, A review of the trends observed in British Isles mean sea level data measured by tide gauges, *Geophysical Journal International*, Volume 136, Issue 3, March 1999, Pages 651–670, https://doi.org/10.1046/j.1365-246x.1999.00751.x
[9] https://oceanservice.noaa.gov/facts/sealevel.html
[10]https://assets.publishing.service.gov.uk/media/60378c448fa8f5048f78a5cf/Exploratory_sea_level_projections_for_the_UK_to_2300_-_report.pdf
[11] https://noc.ac.uk/files/documents/business/Datums-in-Ireland.pdf
[12] https://www.weforum.org/agenda/2023/01/ocean-weather-events-heat/
[13] https://www.eea.europa.eu/en/analysis/indicators/global-and-european-sea-level-rise