Part 1: Open questions

1. **What questions would you ask the WSL on a kick-off call before beginning on the project?**

   For technical requirements, I would inquire about the following things:

   1. How long does it take WSL to organise the event? This would influence the minimum forecasting time span of the model.

   2. The duration of the event. This would set the time interval during which "good" surfing days are to be determined, for example, across the duration of a week or across 5 days. A good surfing day is one where the wave height is at least 3m.

   3. The minimum number of "good" surfing days required during the course of the event. This would enable the model to recommend suitable time intervals for the event to be held.

   4. If there is flexibility on the starting day of the event. This would affect the data aggregation method of the model. For example, if the event can start on any day of the week, then statistics of the wave across a time-window that begins on any day of the week (i.e. a "rolling-window") can be used. However, if the event has to start on a fixed day of the week, for example a Monday, then statistics of the wave have to be computed across a time-window that strictly begins from Monday.

   I would also inquire about circumstances that could lead to the event being cancelled (for example, the occurrences of rip currents, heavy rains, high UV radiation exposures, threat of shark attacks during shark migration seasons, etc.). This could lead to an estimator for such circumstances to be developed.

2. **Before diving into the analysis and forecasting, which new data sources would you consider to complement the buoy data? How would you consider using these data on an ongoing basis?**

   I would consider the following new data sources:

   1. The temperature of the water.
   2. The atmospheric air pressure.
   3. The wind speed, wind duration, and the area over which the wind is blowing (i.e. the fetch).

There are companies that provide these data on an ongoing basis. These usually provide APIs that can be integrated with to procure the latest data at required time intervals.

3. **How would you handle missing data in this project?**

Due to the strong periodicity in the data, I would handle missing data by computing the average of the metric across different years of the same date. The method would consider data from years before as well as after the year of the missing data point. The results for missing data on wave height at Waimea bay are shown below in Figure 1; results of all the other metrics at the other locations are in the document titled "strong_missingData.pdf". There was insufficient data available to fill missing data at Stations 21418t and 46402t.

Figure 1. Plots of missing data imputed using the mean of values on the same date across different years.

4. **Which approaches would you consider for this project? Explain your recommendations as if you were communicating directly to the CEO of WSL.**

Given that the wave heights at Waimea bay exhibit strong periodicity across time, I would consider building a model that would learn this periodic behaviour. The Fourier series based method lends itself to model such trends. This technique combines multiple sinusoidal functions (or sine waves) of different amplitudes and phases to produce the trends observed in the data. The amplitudes and phases of the sinusoidal functions can be computed using the observed data. The resulting model is a sum of these sinusoidal

functions. It can extrapolate the trends into future time intervals, thereby forecasting the wave heights into the future.

5. **How would you recommend that these results be communicated to the WSL on an ongoing basis?**

   As more data becomes available to the model the precision and accuracy of the forecasts improve. Therefore, updates on the prior predictions can be provided on a regular basis. The results can be communicated at regular intervals via a shared platform or an API can be provided to the client which they can integrate into.

Part 2: Proof of concept

As a starting point, we propose a method that attempts to model the periodicity exhibited in the time series of wave heights in Waimea bay. To do this we first aggregated the data into weeks and computed the weekly mean of wave heights. We also explored a weighted mean technique, where the mean of the wave height in a week is weighted by the number of "good" surfing days. The weight mean can be expressed as:

weighted-mean = mean_wave_height * (no. "good" surfing days in a week + 1)/8

As shown later, this weighting technique produced better results for forecasting weeks with at least 2 and 3 "good" surfing days. A non-rolling widow based weekly aggregation was performed, i.e. the mean was computed across 7 days from every Monday. We then decomposed the whole time series into trend, season and residual using the LOESS (locally estimated scatterplot smoothing) technique. The results of the decomposition are plotted in Figure 2.

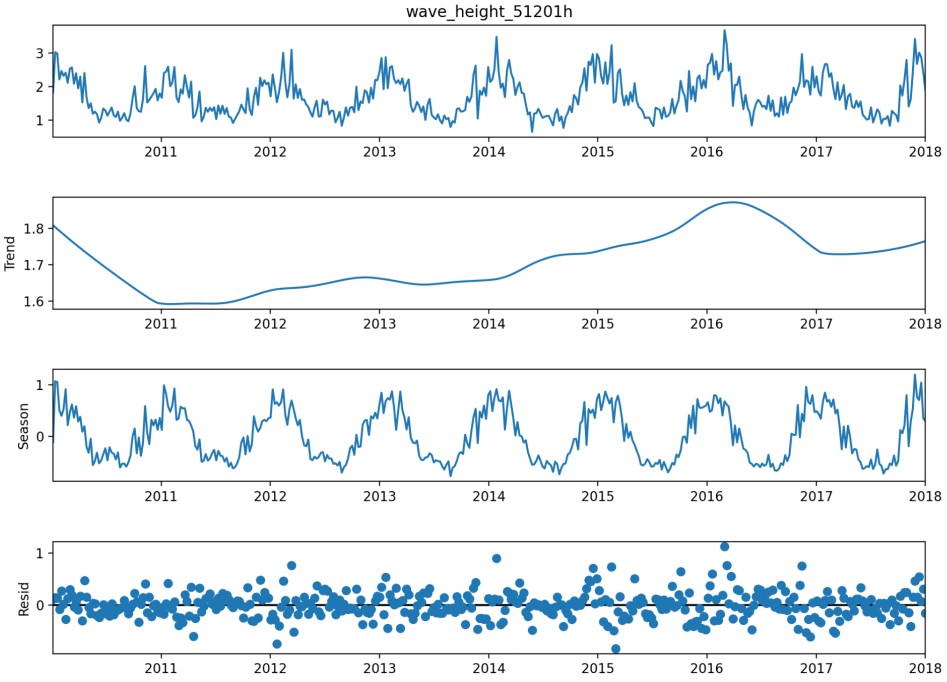


Figure 2. Plots of trends and seasonal decomposition, as well as the residual, of wave heights in Waimea bay.

The residuals show a pattern that reflects peaks in wave heights. We attempted to capture this pattern as well. Therefore, not only trend and season but also residuals were further transformed into their respective Fourier components using the Fast Fourier Transforms (FFT) method. Figure 3 shows the FFT of the trend, season and residual of weekly wave heights.

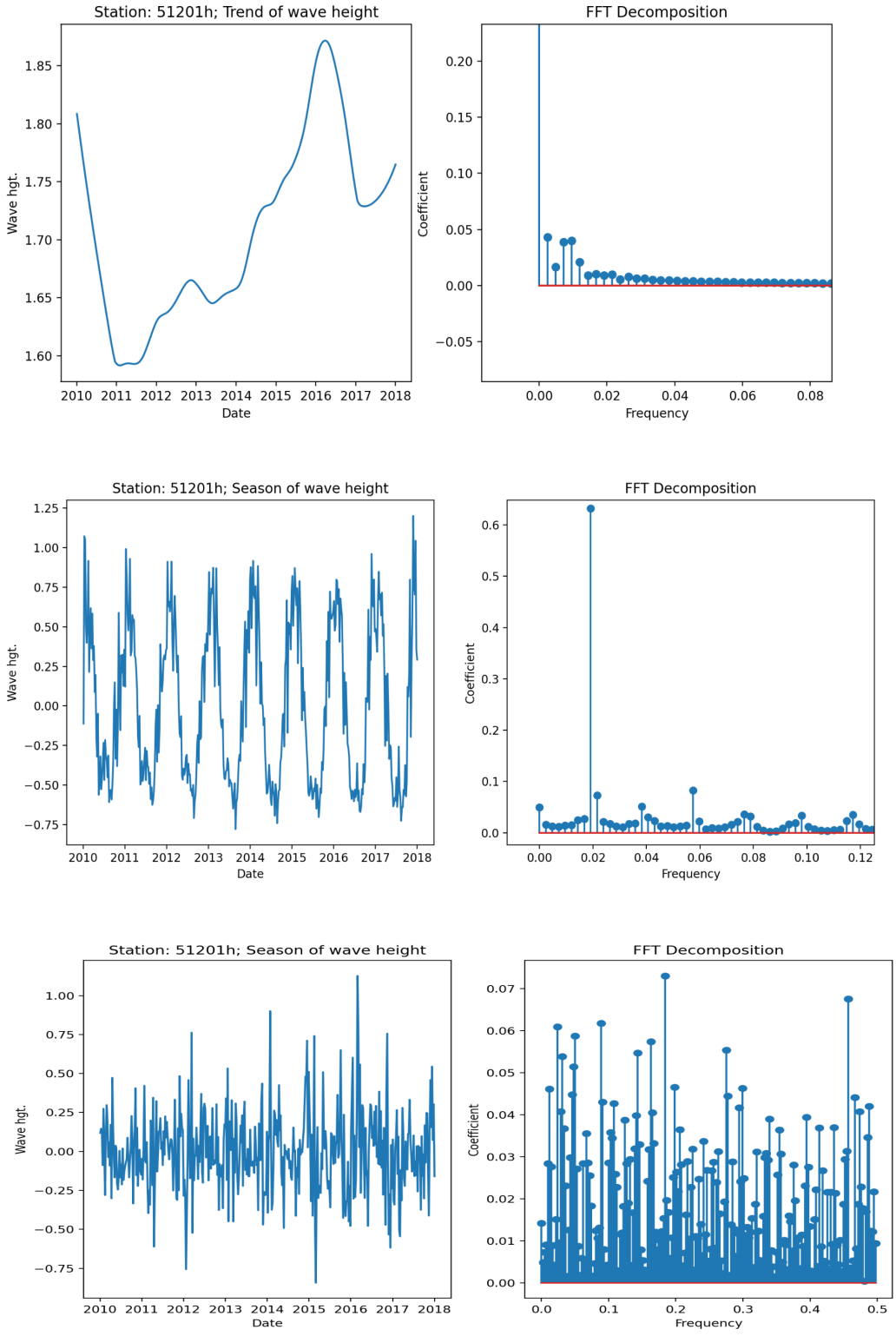


Figure 3. Plots of FFT of trends, season and residual of wave heights in Waimea bay.

The FFT components of trend, season and residual are then used to reconstruct the signal. The reconstructed signal can be extrapolated into the future, providing estimates for the forecast of wave height.

Training and testing the model

The model was trained and tested on different time intervals of the series. We set training and testing time intervals to be between 1Jan2010 - 31Dec2015 and 1Jan2016 - 31Dec2017 respectively. In order to train the model, the components trend, season and high peaks of the series, derived by decomposing the series using LOESS, were transformed into its Fourier components. Care was taken to prevent forward-looking bias in the testing interval by ensuring that only the training interval segment of the components were transformed. Testing involved reconstructing the series and extrapolating it into the testing time interval, after which out-of-sample metrics were computed to analyse the performance of the model.

In order to prevent the model from overfitting the data, the number of coefficients that were generated from the FFT of the training data (i.e. harmonics of the training data), were used as a regularisation parameter. Over-fitting the training data was prevented by carefully selecting the number of harmonics used in reconstructing the signal and extrapolating it into the test time interval. Using all the harmonics would imply fitting the noise whereas using a single harmonic would imply using the mean value of wave height as an estimate for forecasting. Therefore, we performed an out-of-sample test for various percentages of harmonics used in reconstructing the signal. This implies that trend, season and high peak components were reconstructed at various combinations of percentages of total harmonics to reconstruct and extrapolate the final series.

To examine the performance of the model, we first computed the number of “good” surfing days for every week beginning from Monday. A “good” surfing week was then determined as one that has at least N “good” surfing days. The performance of the model was judged on its ability to forecast “good” surfing weeks. The area under the Receiver Operating Characteristic (ROC) and the Precision-Recall (PRC) curves were used as metrics for out-of-sample performance. The table below shows the performance of the model for various values of N, and for varying percentages of harmonics used in reconstructing and extrapolating the signal.

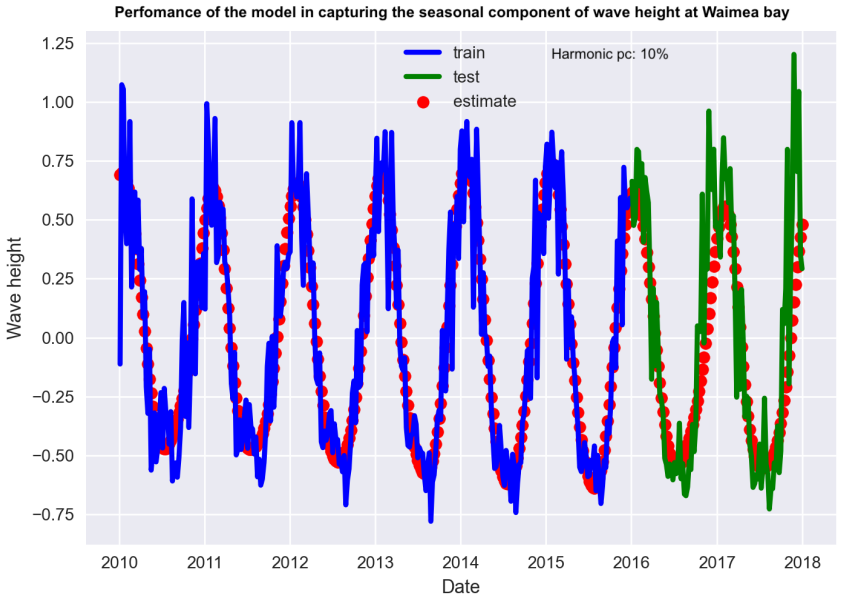
Performace of estimator in forecasting weeks starting Monday with at least 1-, 2- and 3- good surf days, between 1Jan 2016 - 31Dec 2017

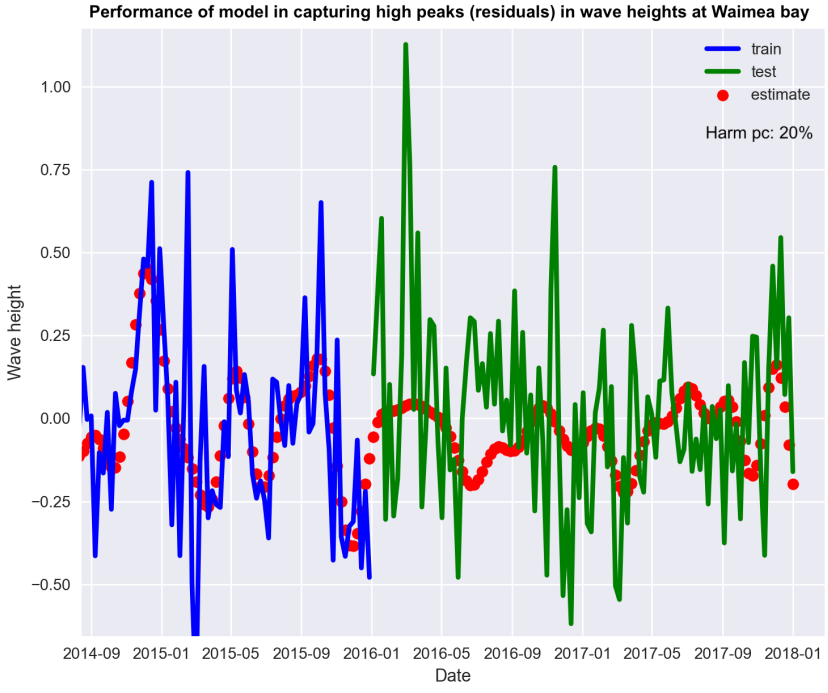
Fraction of harmonics of Trend/Season/Residual	Count of "good" surf days in a week					
	At least 1 good surf day		At least 2 good surf days [3]		At least 3 good surf days [3]	
	ROC AUC	PRC AUC	ROC AUC	PRC AUC	ROC AUC	PRC AUC
0.1% / 0.1% / 0.1%	0.392	0.334	0.328	0.176	0.372	0.151
10% / 20% / 10%	0.935	0.847	0.891	0.653	0.861	0.614
20% / 20% / 20%	0.956	0.902	0.904	0.597	0.884	0.339
40% / 10% / 20%	0.958	0.914	0.905	0.642	0.895	0.358
40% / 10% / 100%	0.922	0.866	0.882	0.704	0.902	0.417
80% / 80% / 80%	0.882	0.726	0.828	0.492	0.856	0.293
100% / 100% / 100%	0.870	0.697	0.839	0.529	0.847	0.294

[1] A "good" surf day is one where the wave is at least 3m high.
[2] Percentage of weeks with at least N good surf day(s) between 1Jan2016 - 31Dec2017: N=1, p.w.=28%; N=2, p.w.=17%; N=3, p.w.=10%. For N>3, p.w.<3%.
[3] A weighted-mean is used instead of a weekly mean.

We observe that the model performs best in predicting weeks with at least 1 “good” surfing day using 40.0% of trend harmonics, 10% of season and 20% of high peak harmonics from the training data. Figure 4 shows the performance of the model in capturing the seasonal, trend and high peak components of the wave heights at the aforementioned combination of harmonic percentages. Figure 5 shows a comparison of the true and forecasted wave height, together with the error. The error indicates that the model was not able to fully capture all the high peaks at 20% harmonics in the testing interval.
The ROC and PRC curves in Figure 6 show that at an 82% true positive rate the model generates a 8% false positive rate and a 80% level of precision in forecasting weeks with at least 1 “good” surfing day.
We note that there are 29 weeks with at least 1 “good” surfing day (making 28% of the data), whereas less than 20 weeks (or 17%) of with at least 2 or more “good” surfing days in the testing time span. This low number of positives could contribute to the limiting performance of the model in forecasting weeks with 2 and 3 “good” surfing days. Furthermore, 27% of weeks with at least 1 “good” surfacing day have 3 or more “good” surfing days and 46% of them have exactly 1 “good” surfing day. As such forecasting weeks with at least 1 “good” surfing day may not be good enough to hold the event.

These observations call other techniques to be explored. State space based techniques, such as Kalman filters, can be used to implement the Seasonal ARIMA (SARIMA) model for this particular problem.
Models that establish relationships across different data sets may tend to be useful. For example, cointegration can be tested between wave heights at waimea bay and the data provided from different locations. A long-run relationship can then be explored between them for forecasting purposes.
Deep learning models such as Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) models could also be experimented with; however, these techniques would require a longer history of data.


Perfomance of the model in capturing the seasonal component of wave height at Waimea bay
train
test
estimate
Harmonic pc: 10%
Wave height
Date
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
2010
2011
2012
2013
2014
2015
2016
2017
2018


Performance of model in capturing high peaks (residuals) in wave heights at Waimea bay
train
test
estimate
Harm pc: 20%
Wave height
Date
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01

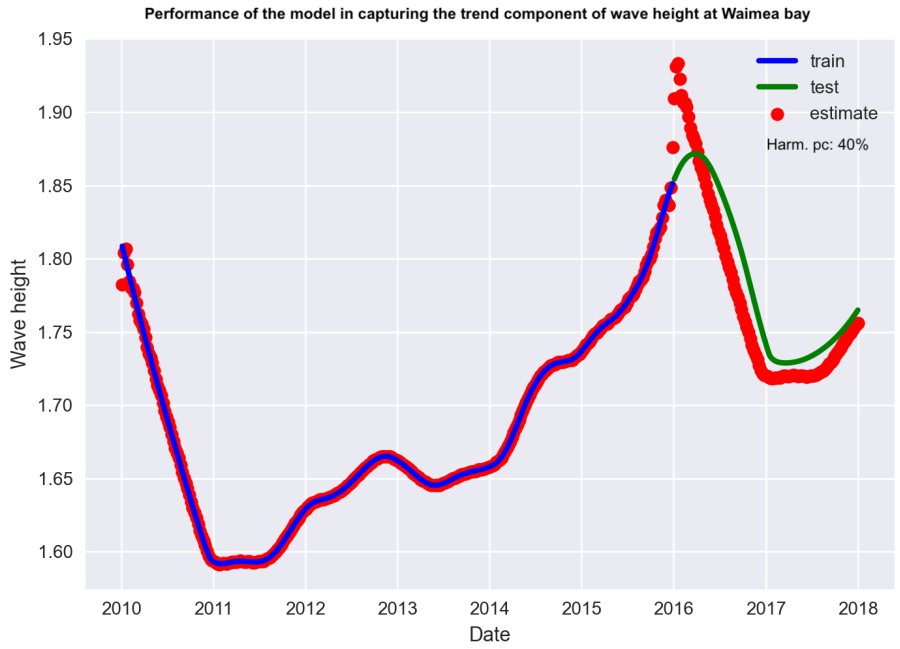


Figure 4. Performance of the model in capturing the trend, seasonal and peak components of wave heights at Waimea bay.

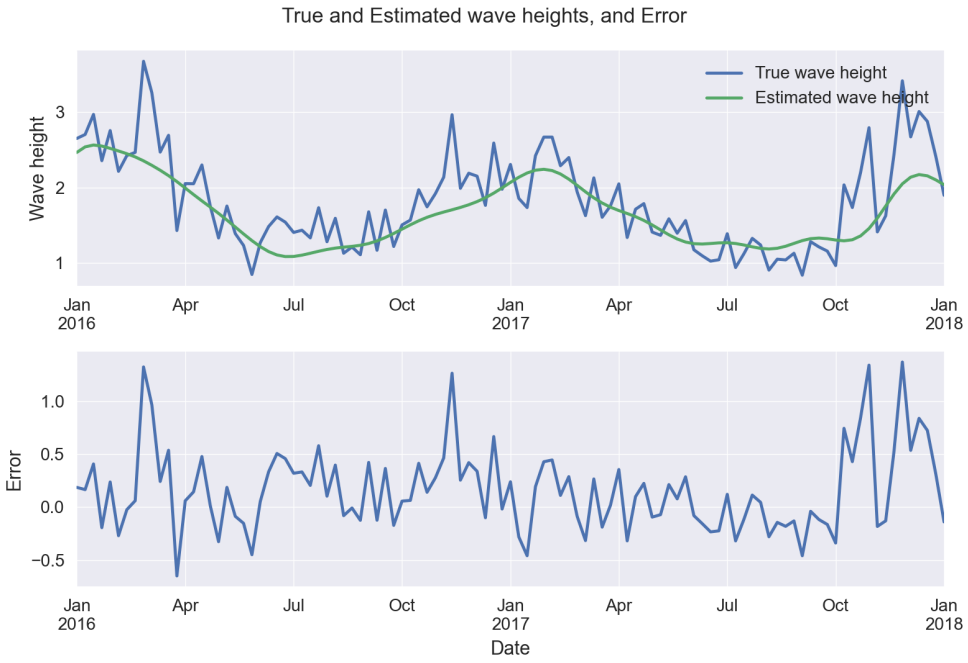


Figure 5. Plots comparing the true wave height to the forecasted wave heights, and the error generated between them.

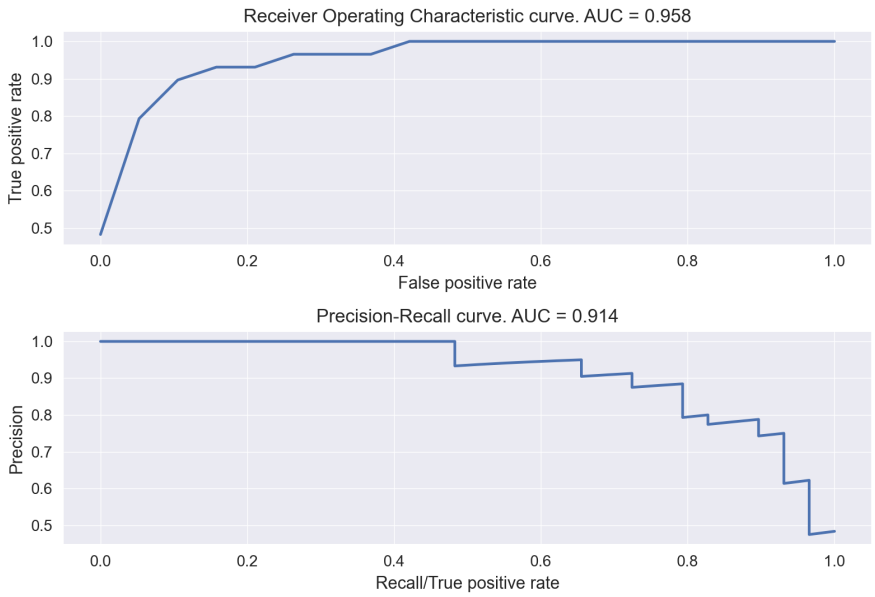


Figure 6. ROC and PRC curves of the model at harmonics in forecasting weeks with at least 1 “good” surfing day