A Machine Learning Method to Correct Wave Model Gridded Output

323845

Ashley Ellenson, School of Civil and Construction Engineering, Oregon State University
Yuanli Pei, School of Electrical Engineering and Computer Sciences, Oregon State University
Greg Wilson, College of Earth, Ocean and Atmospheric Sciences, Oregon State University
H. Tuba Ozkan-Haller, College of Earth, Ocean and Atmospheric Sciences, Oregon State University

OVERVIEW

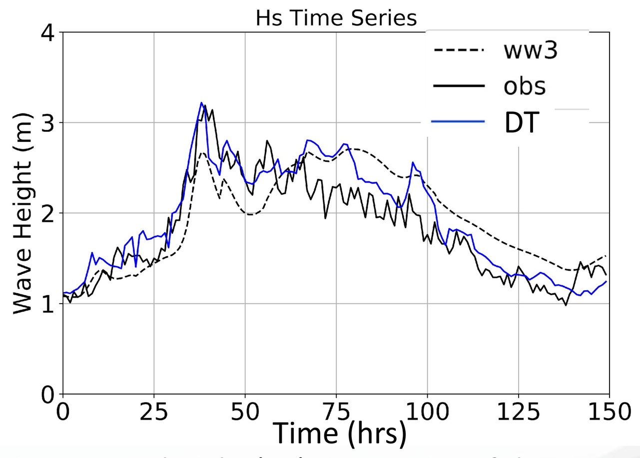


Fig 1. Wave height (Hs) time series of the original wave forecast, the observations, and the corrected time series.

The goal of this study is to correct wave height forecasts through the use of a machine learning technique. The technique is a bagged Decision Tree (DT).

Machine Learning Techniques

- Generally, machine learning algorithms make predictions by learning patterns between input data, called features, and the output desired, called the target.
- The learning occurs during a "training" phase, where the algorithm is given both the input features and the associated target.
- The predictions are made during the "testing" phase, where the algorithm is only given input features and predicts the associated target.

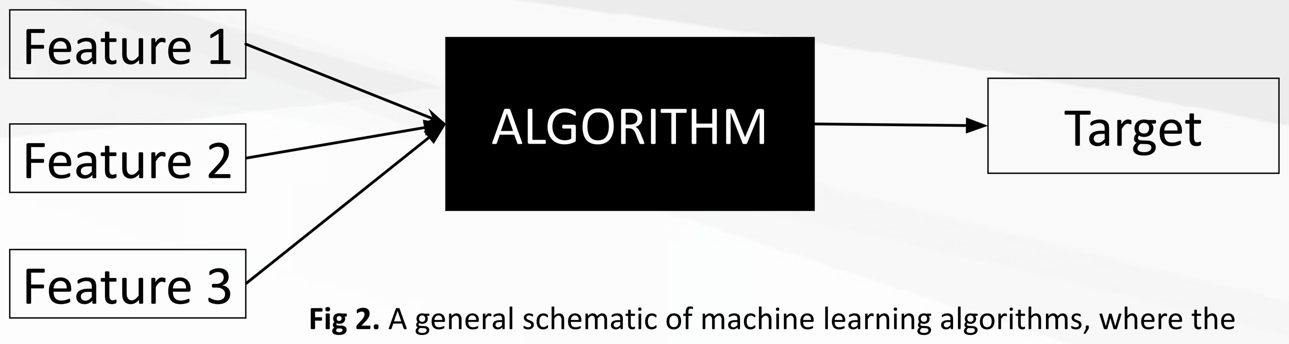


Fig 2. A general schematic of machine learning algorithms, where the algorithm is given input features and predicts a target.

WHAT IS DECISION TREE?

Decision Tree (DT) creates sub-spaces of the target values with respect to associated features. The mean target value of each subspace is its prediction. It creates sub-spaces by following a criterion.

The criterion is to minimize the mean squared error (MSE) between the mean target value ($\widehat{Y}_i$) of the subspace and the rest of the target values in that subspace (Y_i).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Where n is the total number of members in that subspace.

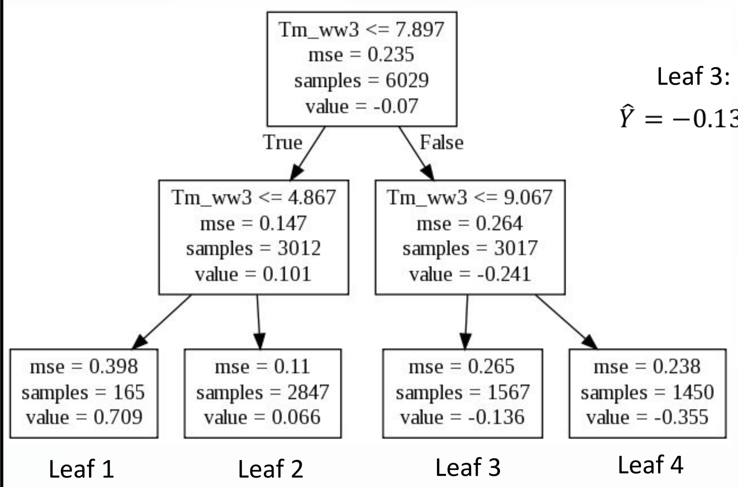


Fig 3. An example of a Decision Tree. *Samples* are the number of samples in that sub-space. *MSE* is the mean squared error between the member values and the average value, and *value* is the mean target value associated with that subspace.

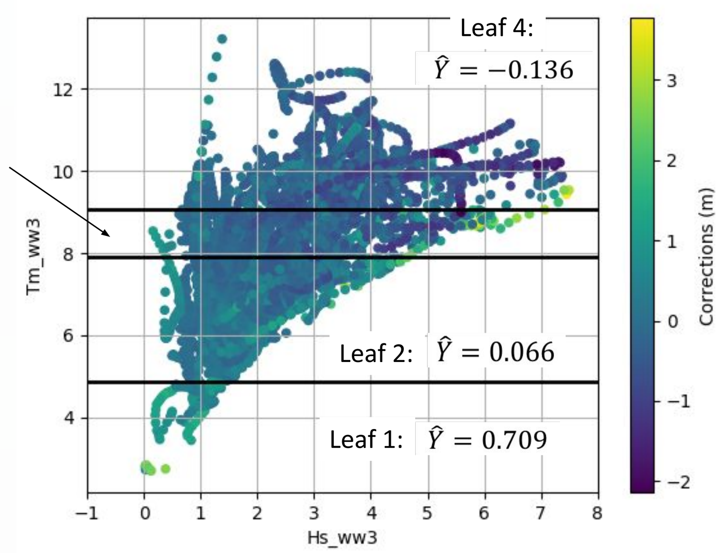


Fig 4. The feature space and sub-spaces associated with the DT in **Figure 3.** The feature space is sub-divided into the final sub-spaces, or leaves according to the boolean decisions of the tree (Tm >= ##). The mean value, or DT prediction, of that sub-space is indicated.

To make a prediction, the DT assigns a data point that falls into the sub-space the mean value of that sub-space. The final subspace is called a terminal node or "leaf".

This study employs an ensemble method, where more than one tree is considered in a Bagged Decision Tree. A random subset of the data is considered for each tree in the bagging method.

This work is funded by the National Science Foundation's Research Traineeship: Risk and Uncertainty Quantification in Marine Science

METHODS

Input features include hourly time series of environmental parameters and the output target consists of an hourly time series of errors.

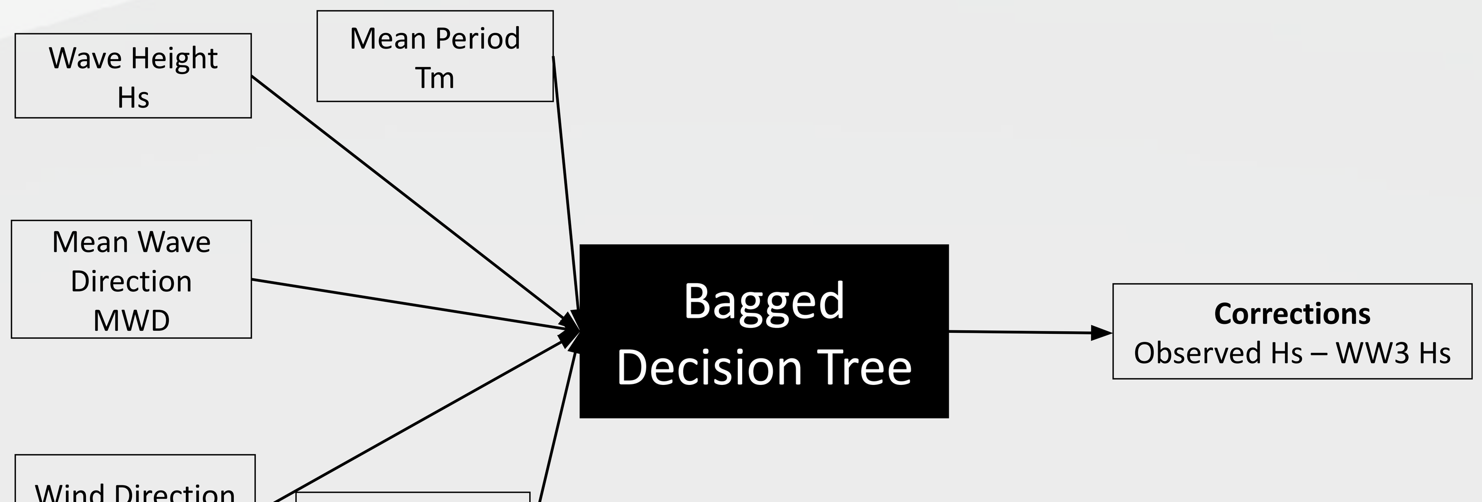


Fig 5. A schematic of the application of the DT ensemble algorithm to wave forecasts. The feature inputs and target output is indicated.

Data

- Wave model output is from WaveWatch III (WW3) with ST2 physics 24-hour forecast horizon.
- Wind data is from Global Forecast System (GFS) input to WW3.
- Training data consists of summer months (April 1 – September 30) from years 2012-2014
- Testing data consists of summer months from 2015

Input Features

Input features include wave information (Hs, Tm, and MWD) and wind information (wind direction, wind speed).

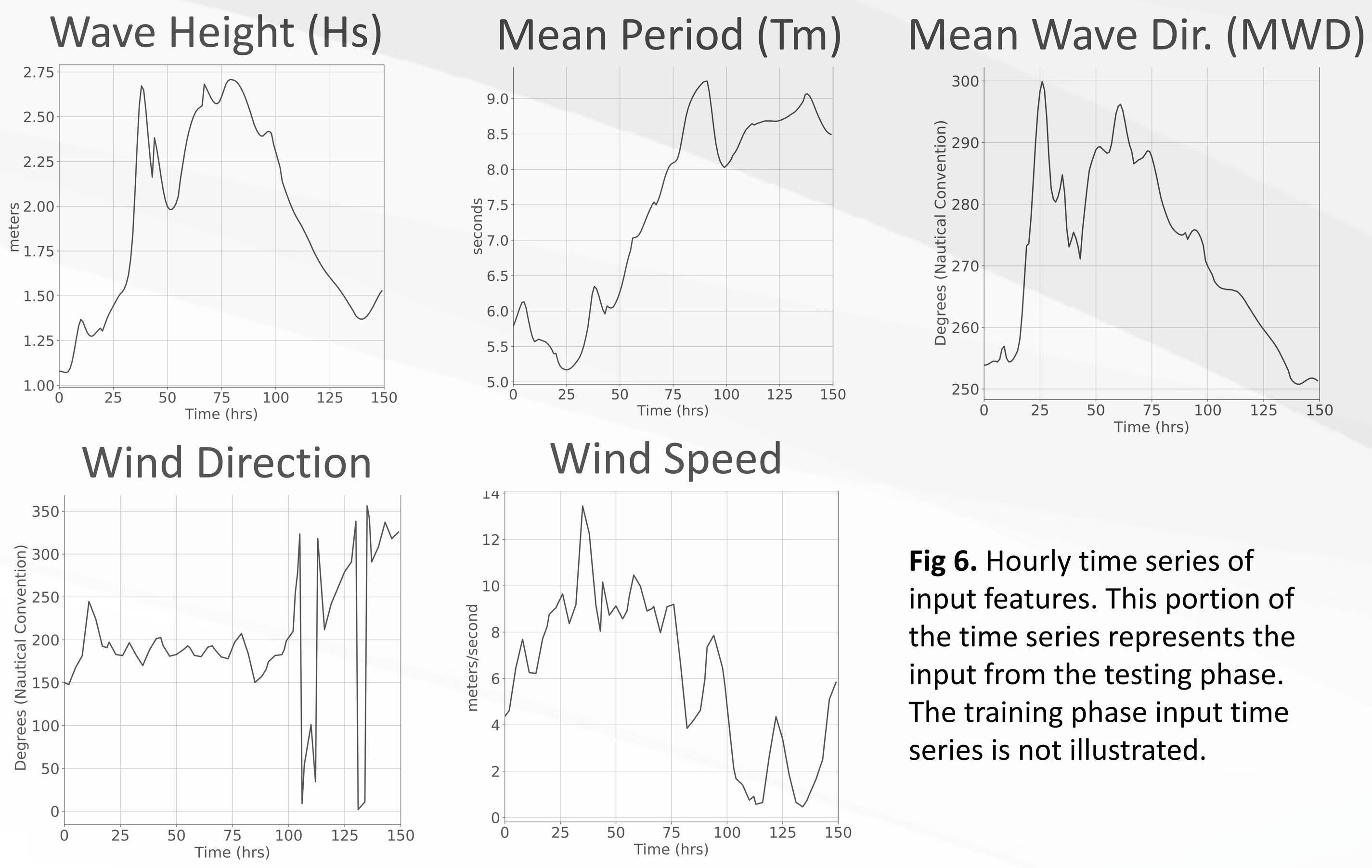


Fig 6. Hourly time series of input features. This portion of the time series represents the input from the testing phase. The training phase input time series is not illustrated.

Target

The target is the difference between the observed wave height and the modelled wave height (the "corrections").

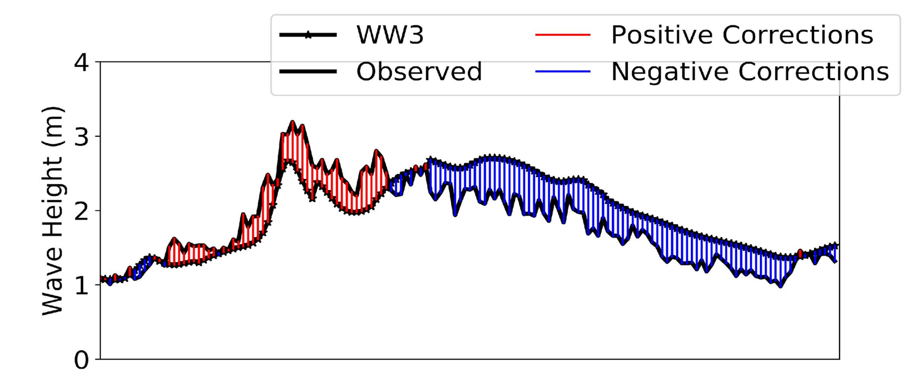
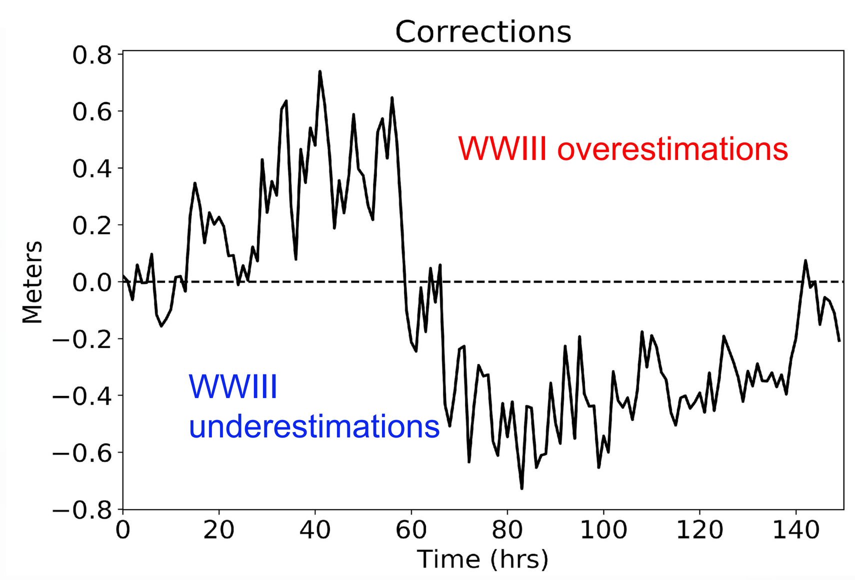


Fig 7. Hourly time series of wave forecast and observations. The difference between the two time series or the corrections, is colored. These corrections are the target for the DT.

Fig 8. The colored corrections from **Figure 7** as a time series. This time series is the target associated with the input features from **Figure 6.**



RESULTS

DT experiments, or "data denial" tests, consisted of removing one feature at a time and leaving the remainder as input. The goal is to understand the effect of removing information from a certain environmental parameter on DT performance.

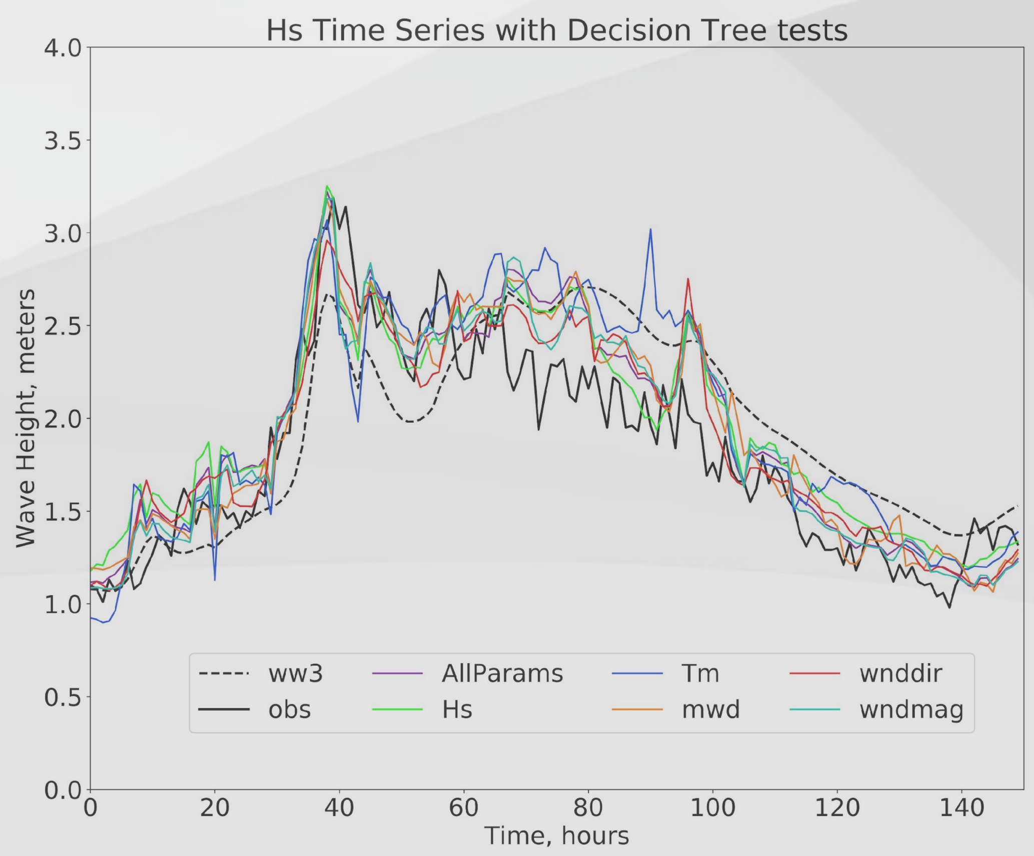


Fig 9. DT final time series from the testing phase. Each data denial experiment is colored. The feature removed from the input data set is indicated on the label.

Error Metrics

- All trials improve RMSE and correlation coefficient of original model.
- DT performs worst when mean period is removed.
- DT performs best when wind direction is removed

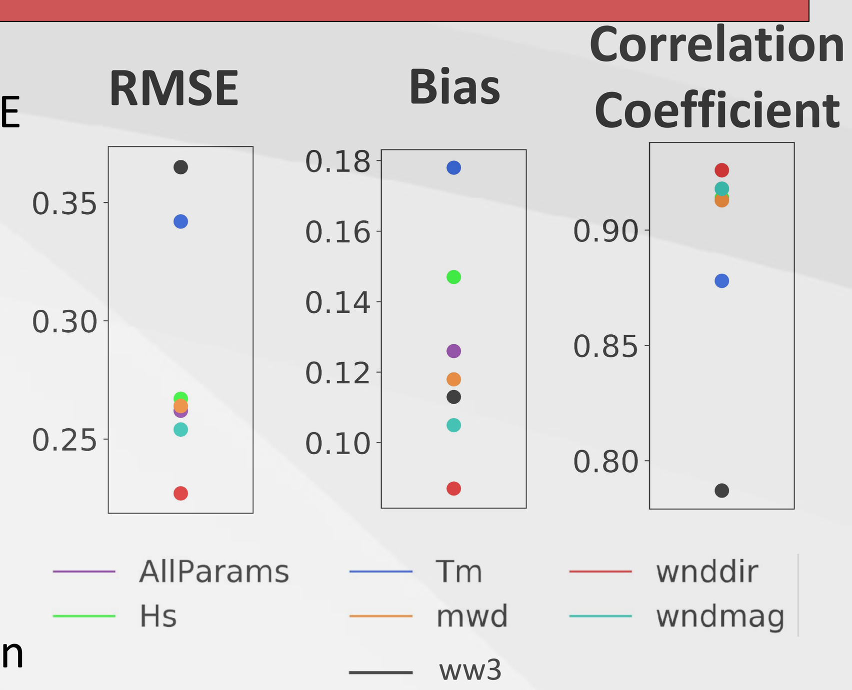


Fig 9. Error metrics Root-Mean-Square-Error (RMSE), bias and correlation coefficent associated with each data denial test.

Feature Importances

The three features which are most often used by decision tree are:

1) Tm
2) MWD
3) Hs

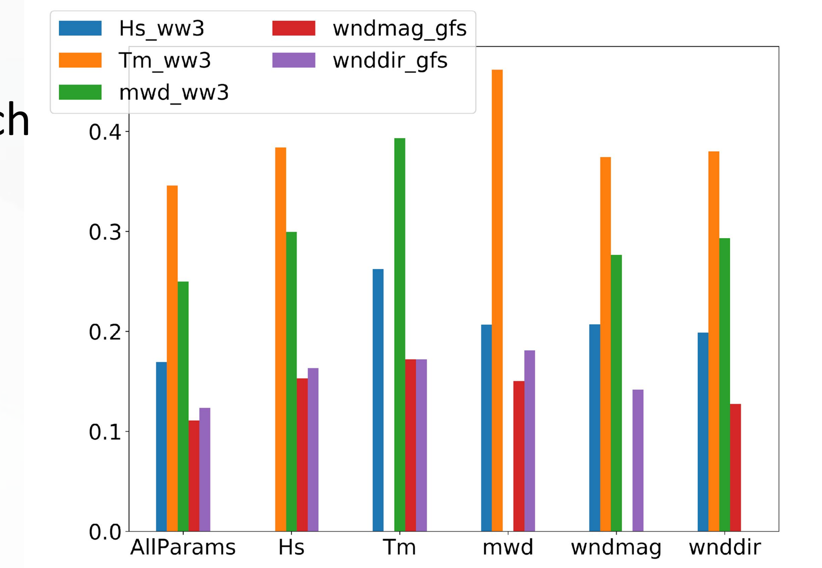


Fig 10. Feature importances indicate how frequently DT uses that feature to split the target space.

CONCLUSIONS

Decision tree was applied on correcting forecasts of wave height and successfully improved the performance for all data denial runs with respect to error metrics RMSE and correlation coefficient.

DT performs best when wind direction is removed.

Mean period is the feature considered most often by DT to make its final predictions in all runs which include this feature.