The Application of Bagged Decision Tree to Corrections of Wave Height Forecasts

Ashley Ellenson, Yuanli Pei, Greg Wilson, Tuba Ozkan-Haller, Xiaoli Fern
OSU Coastal and Ocean Engineering Brown Bag
April 19, 2018

◄ Hide

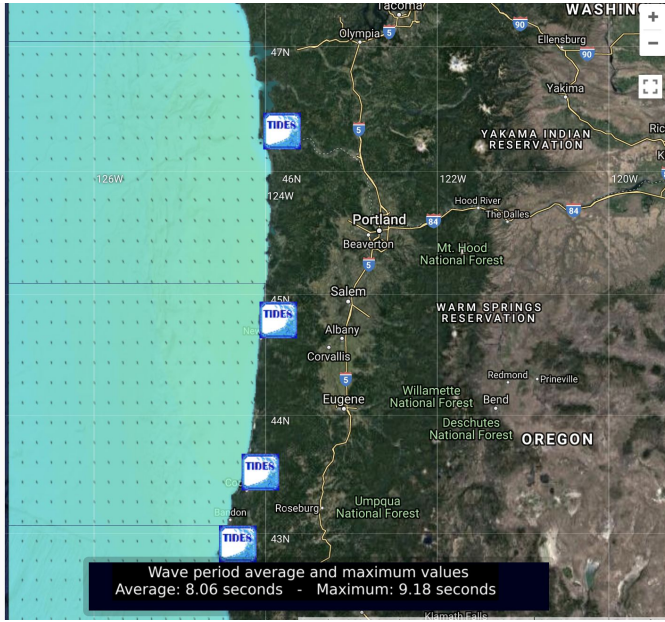
The keys for Currents and Wave Direction are estimates. See About page for details.

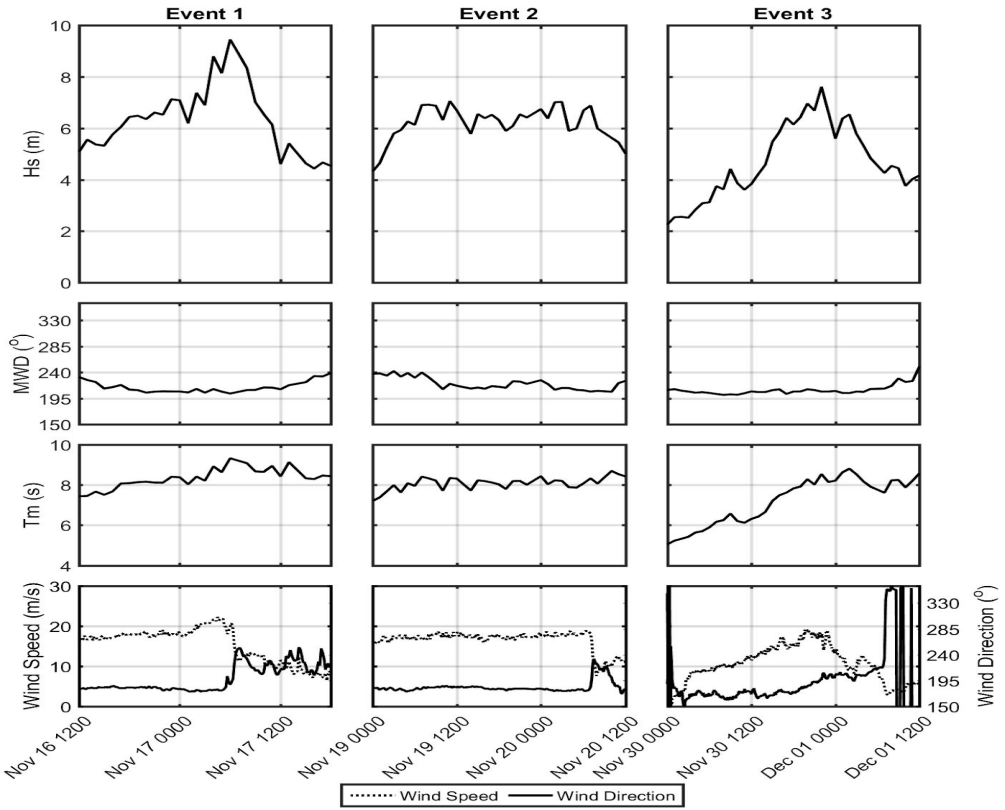
Forecast Date/Time

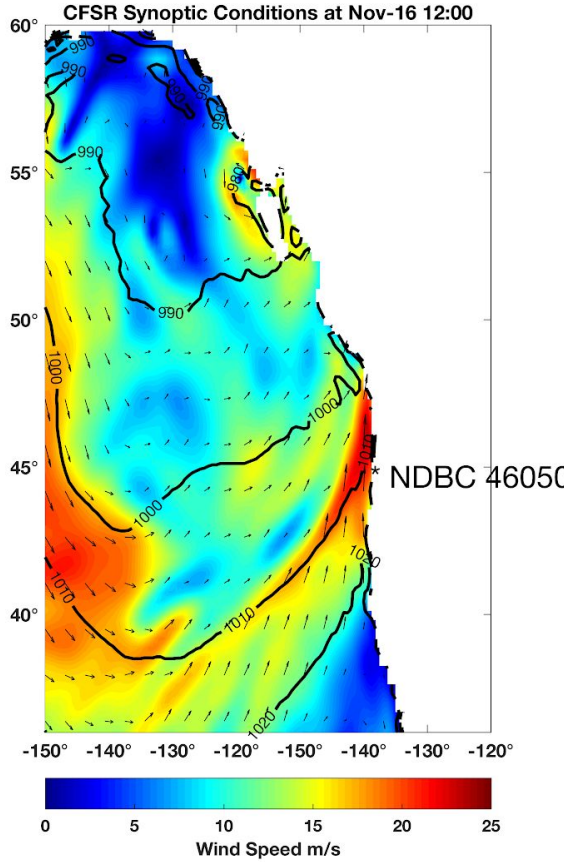
Prev | April 17, 2018, 5 p.m. | Next

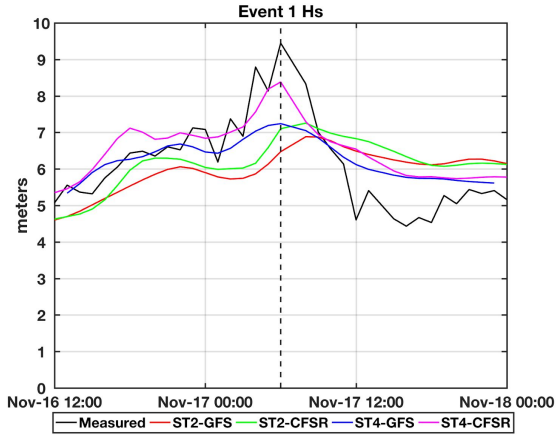
Map Data Layers

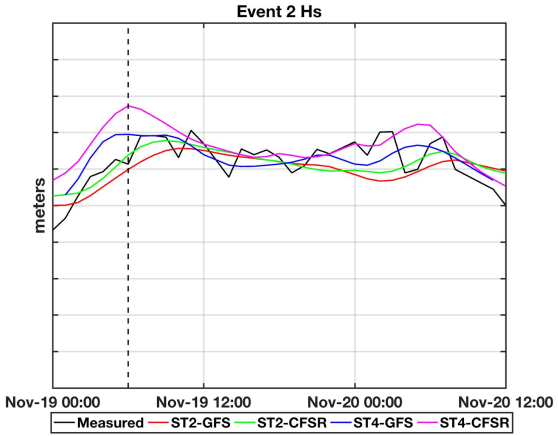
Surface Temperature	OFF
Surface Currents	OFF
Wave Height	ON
Wave Direction/Period	ON
Wind	OFF
Bottom Temperature	OFF
Surface Salinity	OFF
Bottom Salinity	OFF
Sea Surface Height	OFF
Thermocline	OFF

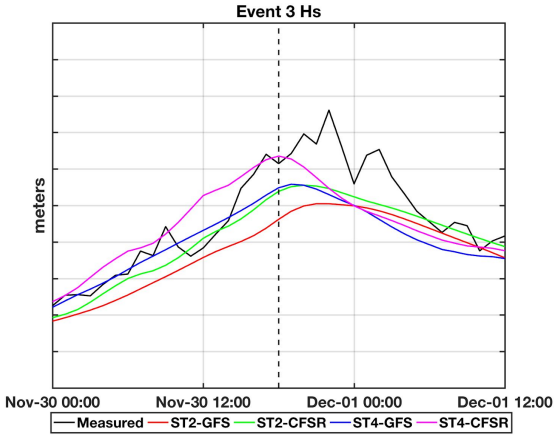


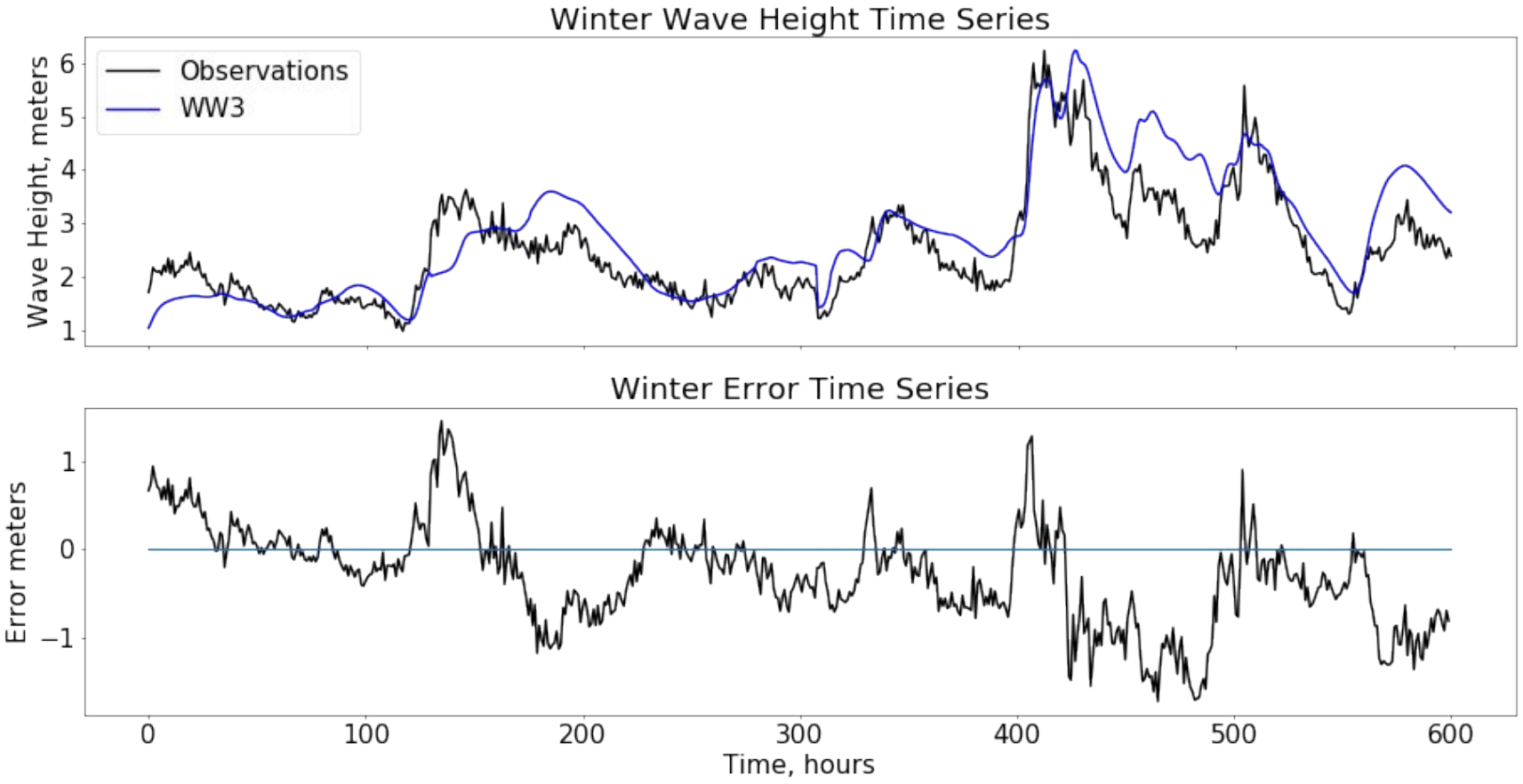

Event 1
Event 2
Event 3
Hs (m)
MWD (°)
Tm (s)
Wind Speed (m/s)
Wind Direction (°)
Nov 16 1200
Nov 17 0000
Nov 17 1200
Nov 19 0000
Nov 19 1200
Nov 20 0000
Nov 20 1200
Nov 30 0000
Nov 30 1200
Dec 01 0000
Dec 01 1200
Wind Speed
Wind Direction


CFSR Synoptic Conditions at Nov-16 12:00
NDBC 46050
Wind Speed m/s

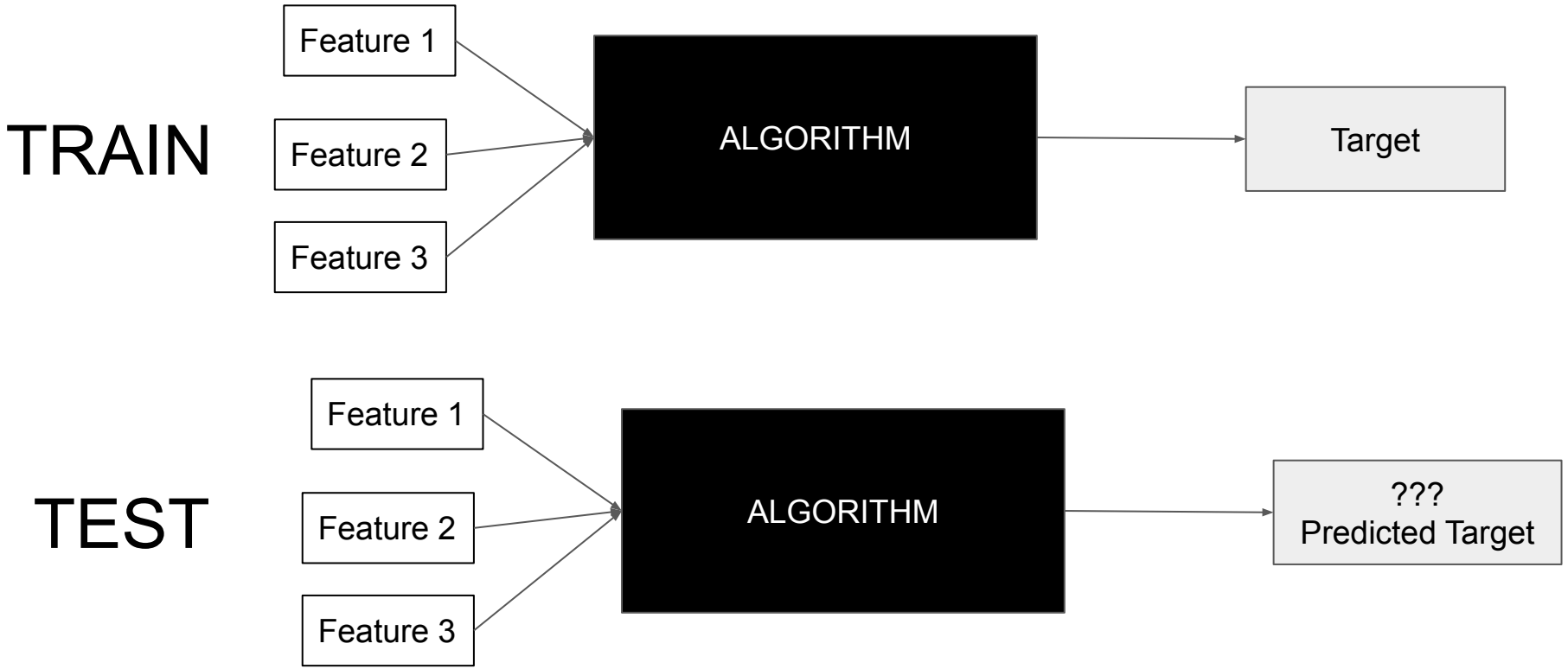

Event 1 Hs
meters
10
9
8
7
6
5
4
3
2
1
0
Nov-16 12:00
Nov-17 00:00
Nov-17 12:00
Nov-18 00:00
Measured
ST2-GFS
ST2-CFSR
ST4-GFS
ST4-CFSR


Event 2 Hs
meters
Nov-19 00:00
Nov-19 12:00
Nov-20 00:00
Nov-20 12:00
Measured
ST2-GFS
ST2-CFSR
ST4-GFS
ST4-CFSR


Event 3 Hs
meters
Nov-30 00:00
Nov-30 12:00
Dec-01 00:00
Dec-01 12:00
Measured
ST2-GFS
ST2-CFSR
ST4-GFS
ST4-CFSR


Winter Wave Height Time Series
Observations
WW3
Wave Height, meters
1
2
3
4
5
6
Winter Error Time Series
Error meters
1
0
−1
0
100
200
300
400
500
600
Time, hours

Machine Learning Algorithms: A General Overview



Decision Tree On Wave Forecasts: A Simple Case

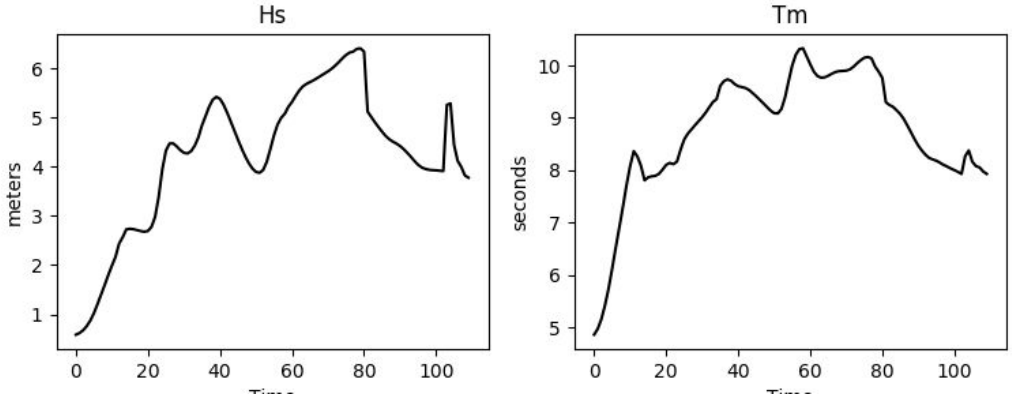
Decision Tree On Wave Forecasts

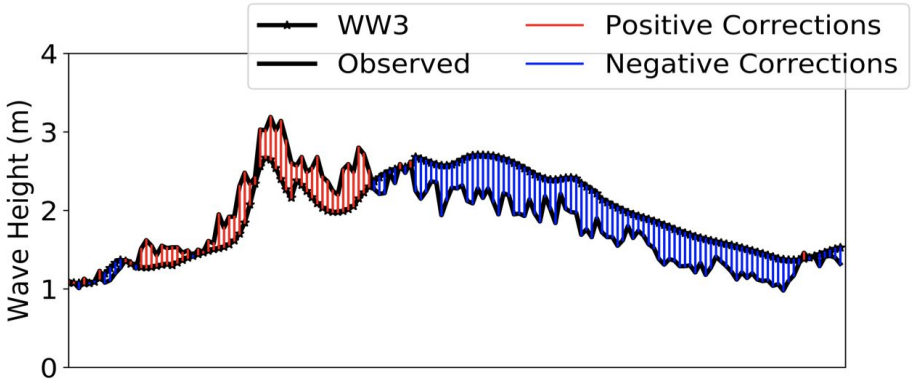
Input Features

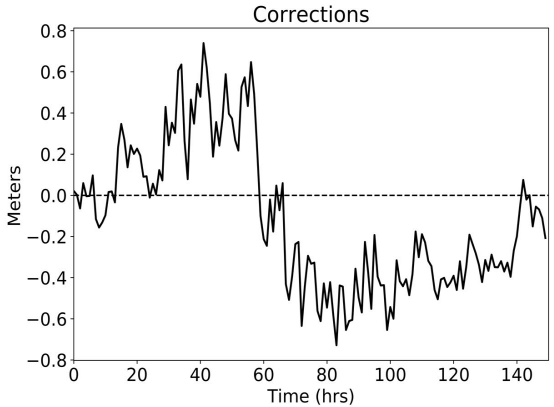
x(t) = [Hs(t), Tm(t)]

Target

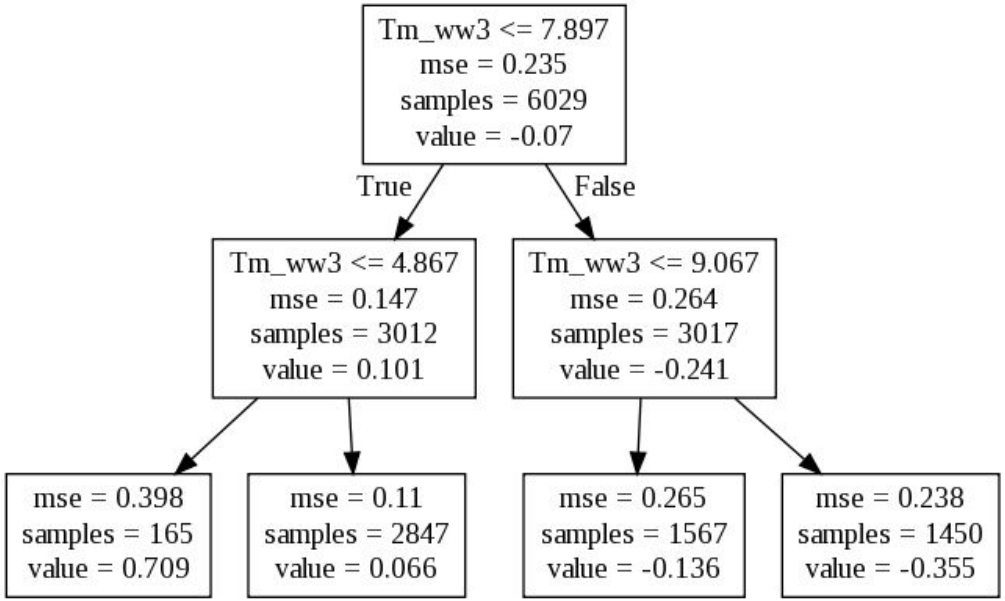
y(t) = Hobs(t) - Hww3(t)

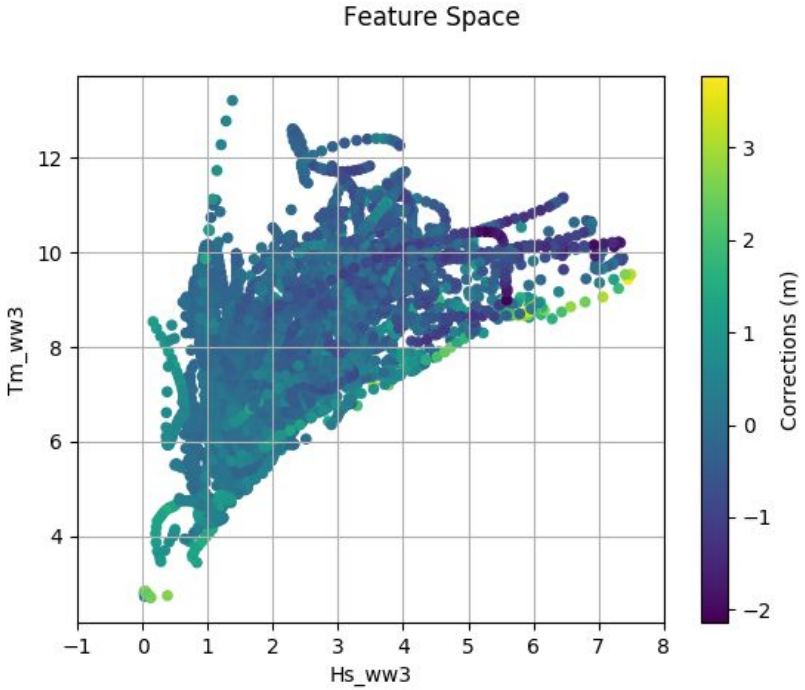




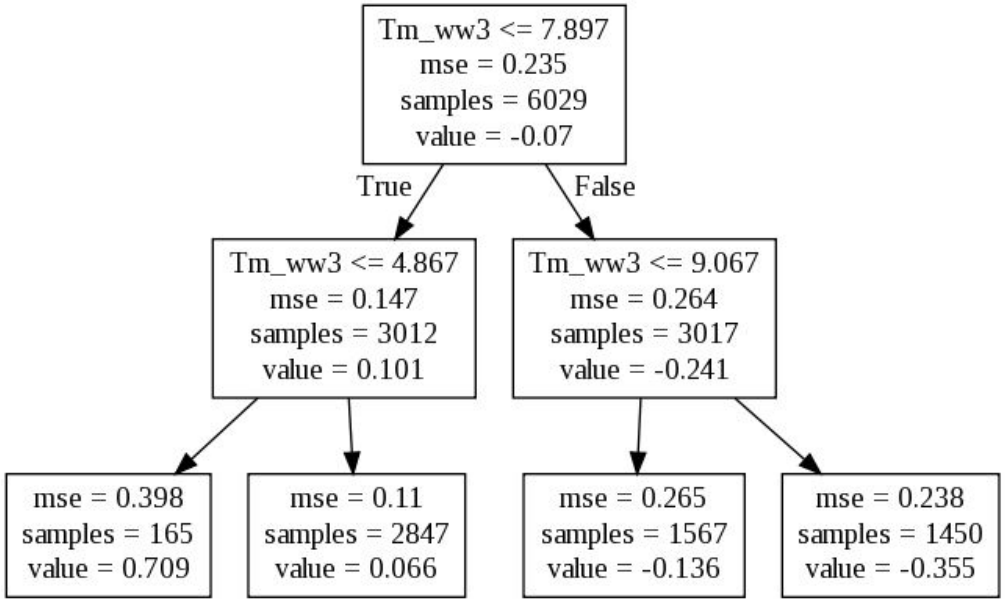


Decision Tree Logic

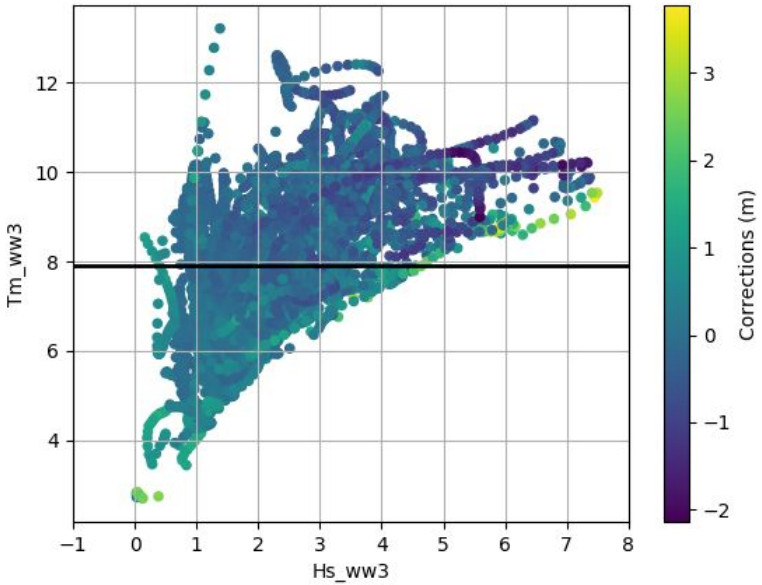




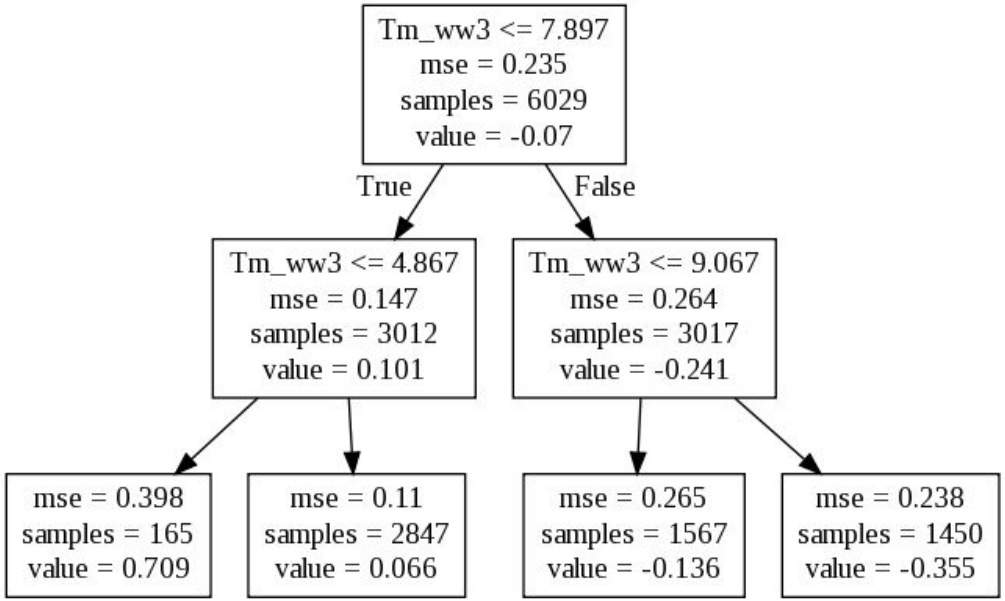
Decision Tree Logic



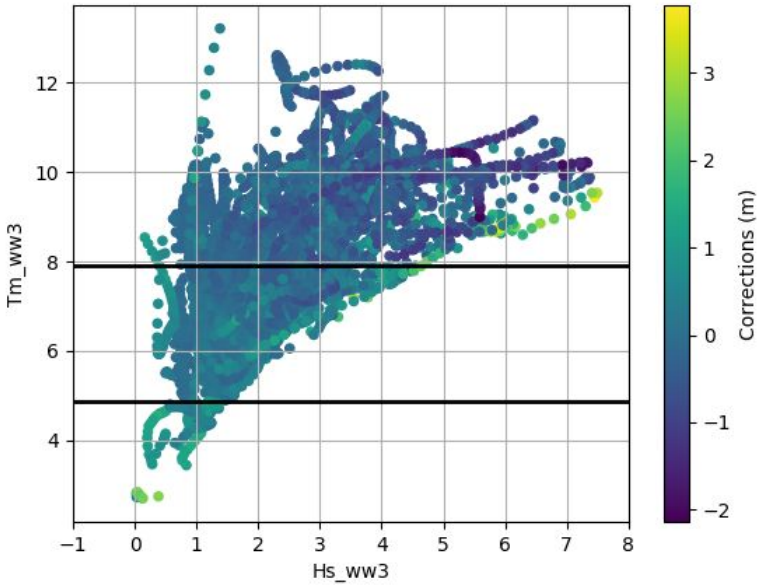
Feature Space



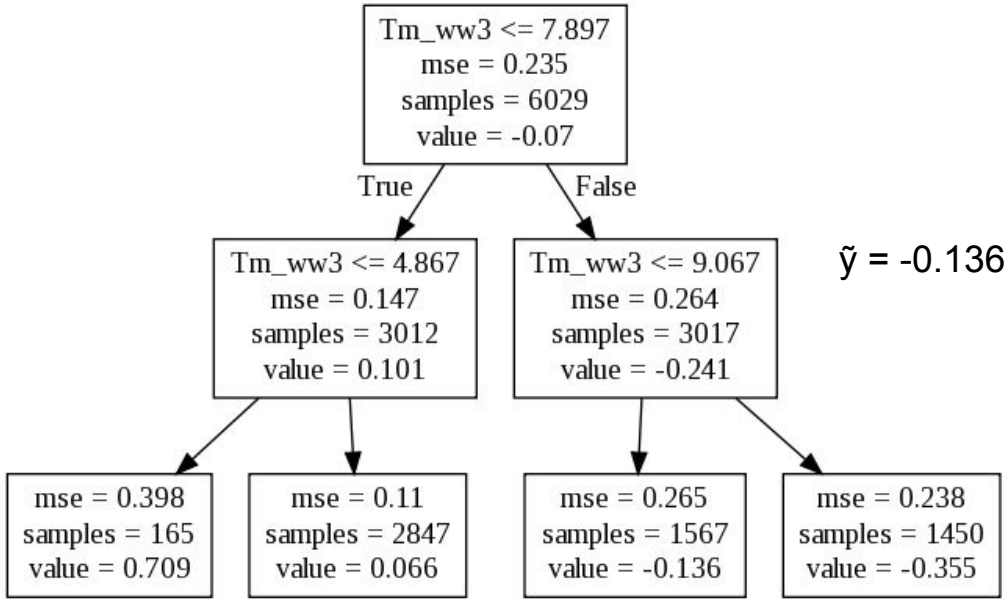
Decision Tree Logic



Feature Space

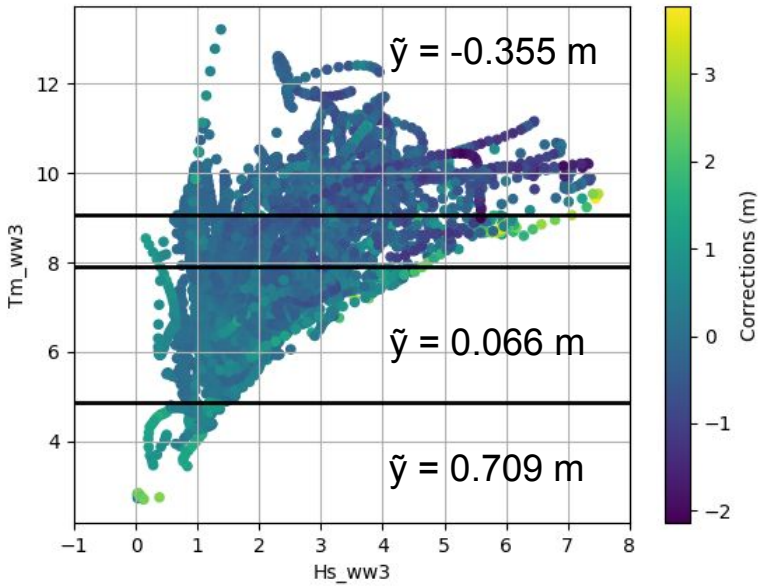


Decision Tree Logic

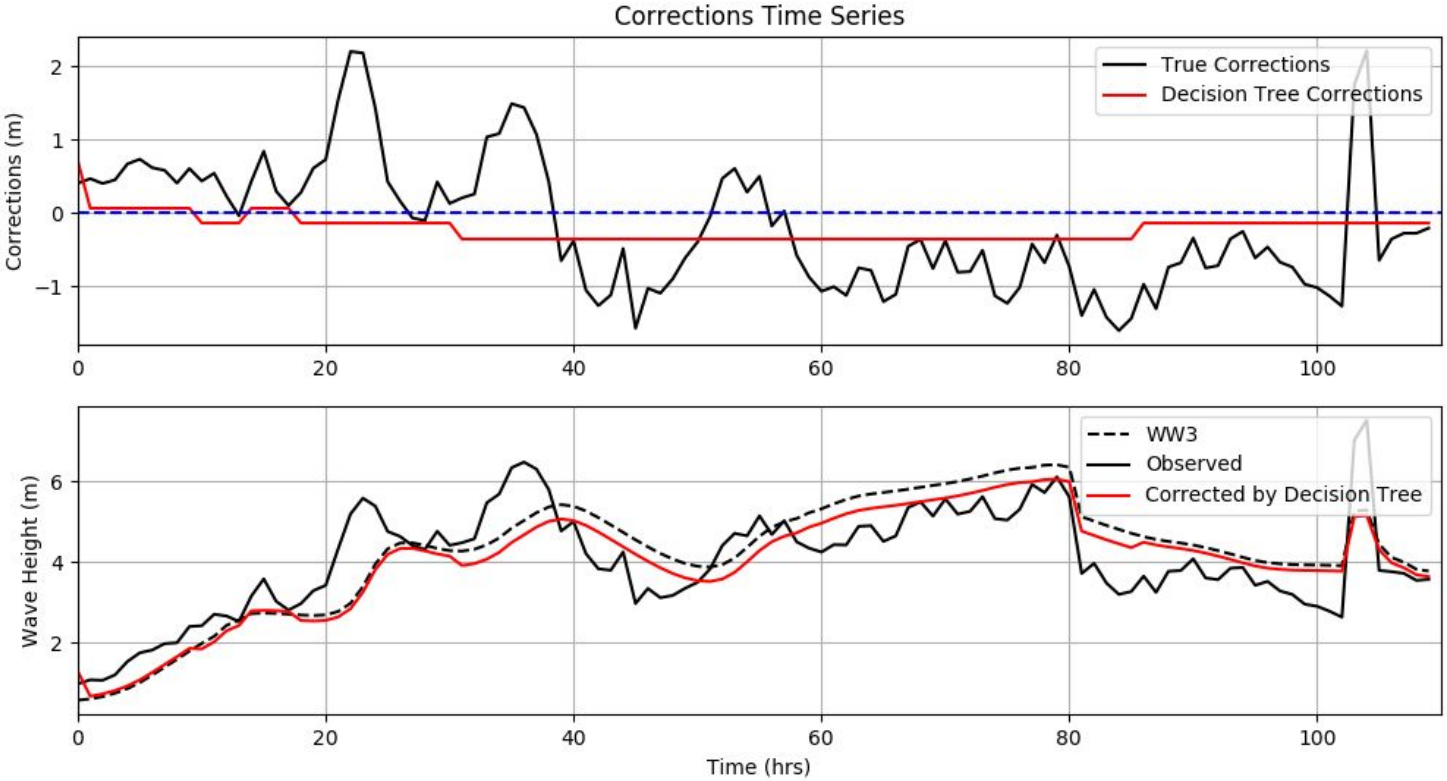


ỹ = -0.136 m

Feature Space

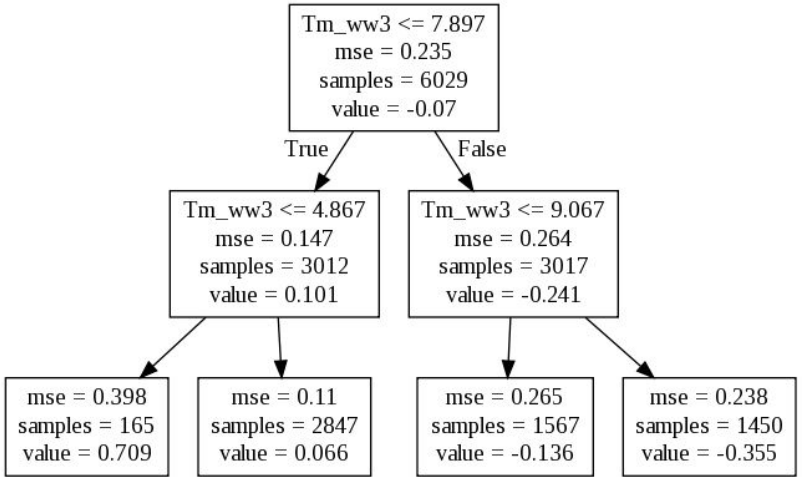


Decision Tree Output



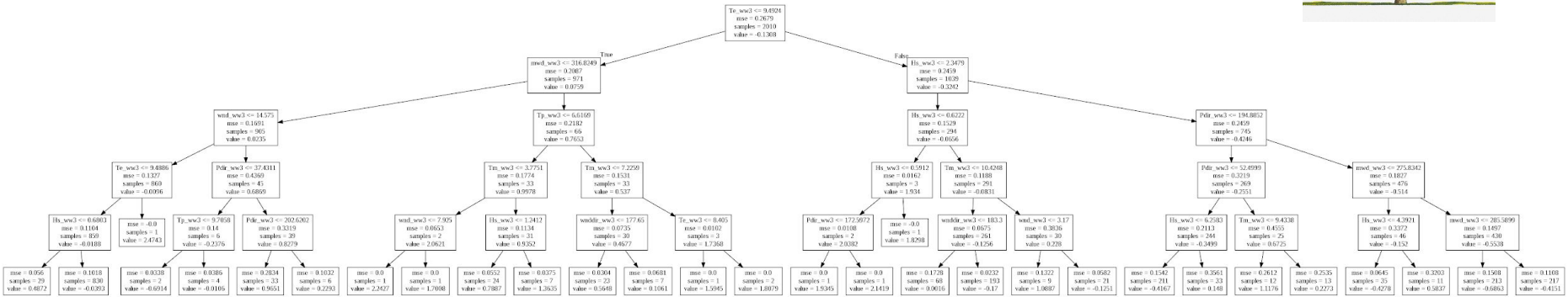
Decision Tree On Wave Forecasts: A Complicated Case

Random Forest Parameters: Tree Depth

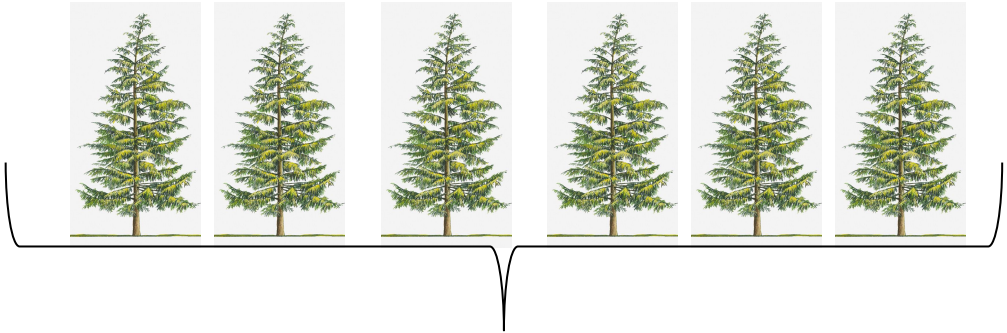


Tree depth: 2

Tree depth: 5



Bagged Decision Tree Parameters: Number of Trees



Final Prediction is average of each tree's prediction

Bagging Method

Each tree is trained on a subset of the data.

Data points are picked uniformly and with replacement.

Determining Forest Parameters

vs

Decision Tree Variations for Different Problems

Ensemble Methods

- Bagging
- Boosting
- Random Forest

Prediction Methods

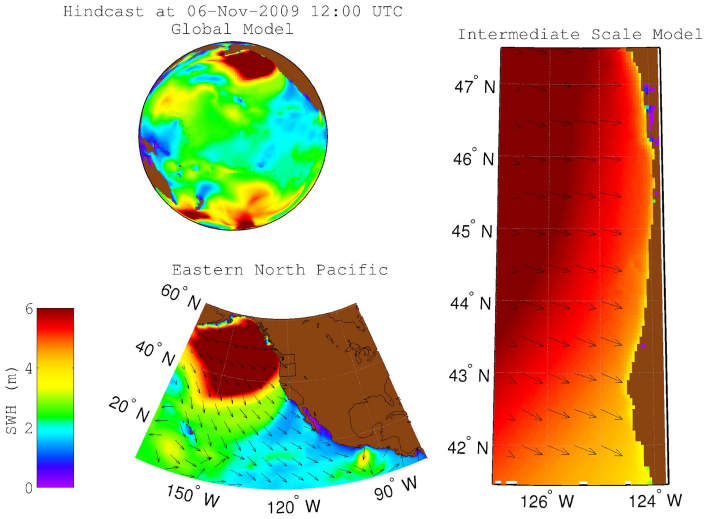
- Classification
- Regression

Objective Function (Regression)

- Mean Absolute Error
- Least Squares Error

Data

- Observations from NDBC Buoy 46050
- WW3 ST2 24-hour forecast horizon, years 2012-2015
- Forced by Global Forecast System (GFS) wind



Garcia-Medina et al, 2014

Experimental Set Up

Train

Summer (April 1-September 31)

- 10,158 points
- Years 2012-2014

Winter (Jan 1- March 31, October 1 - Dec. 31)

- N = 10, 092 points
- Years 2012-2014

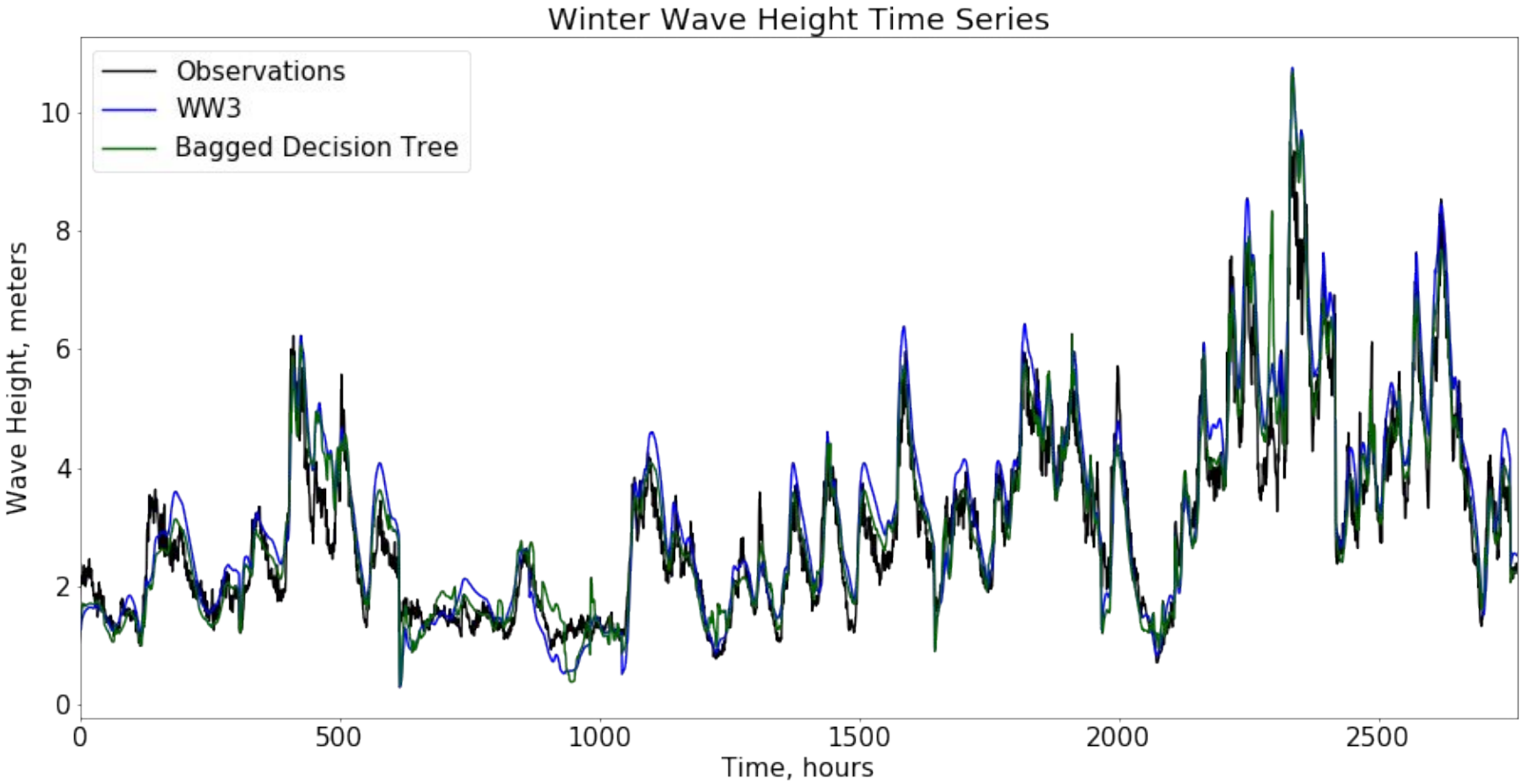
Test

Summer 2015

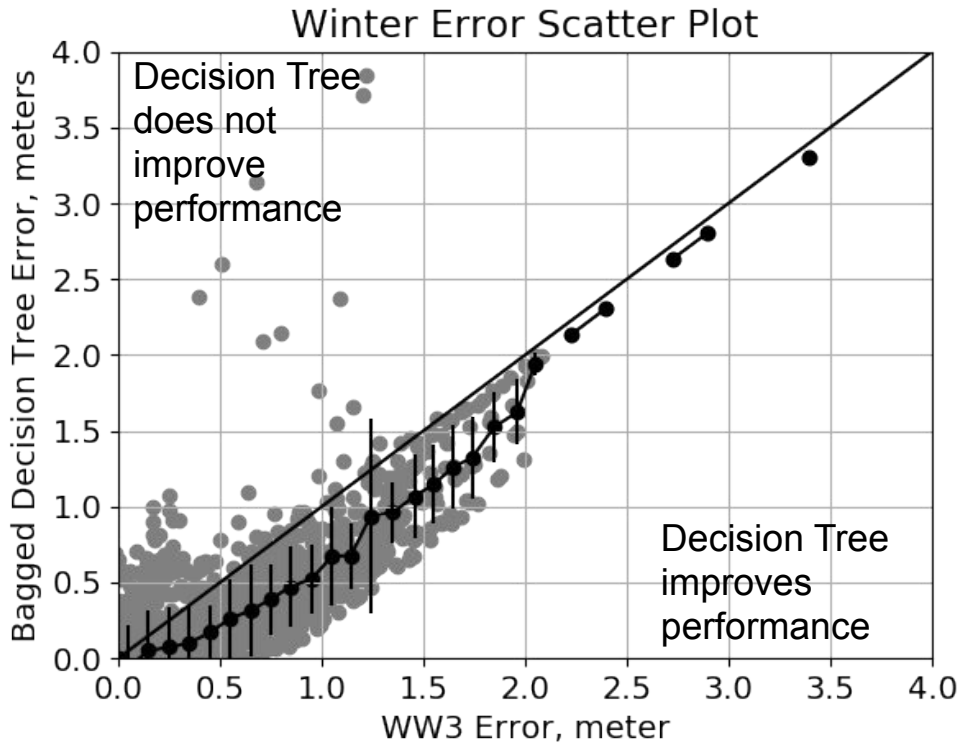
- 2015
- 3,531 points

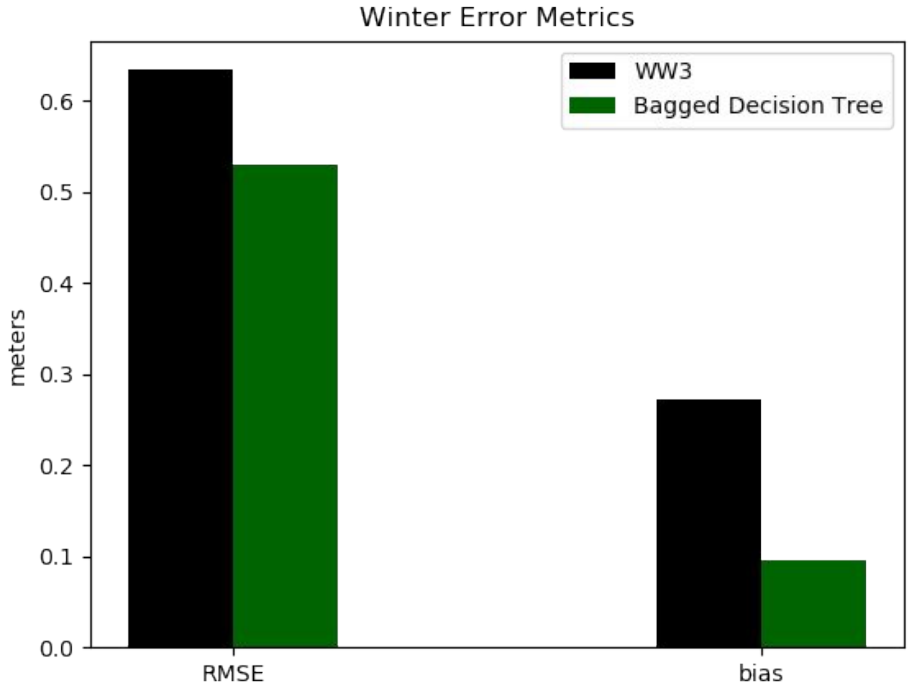
Winter 2015

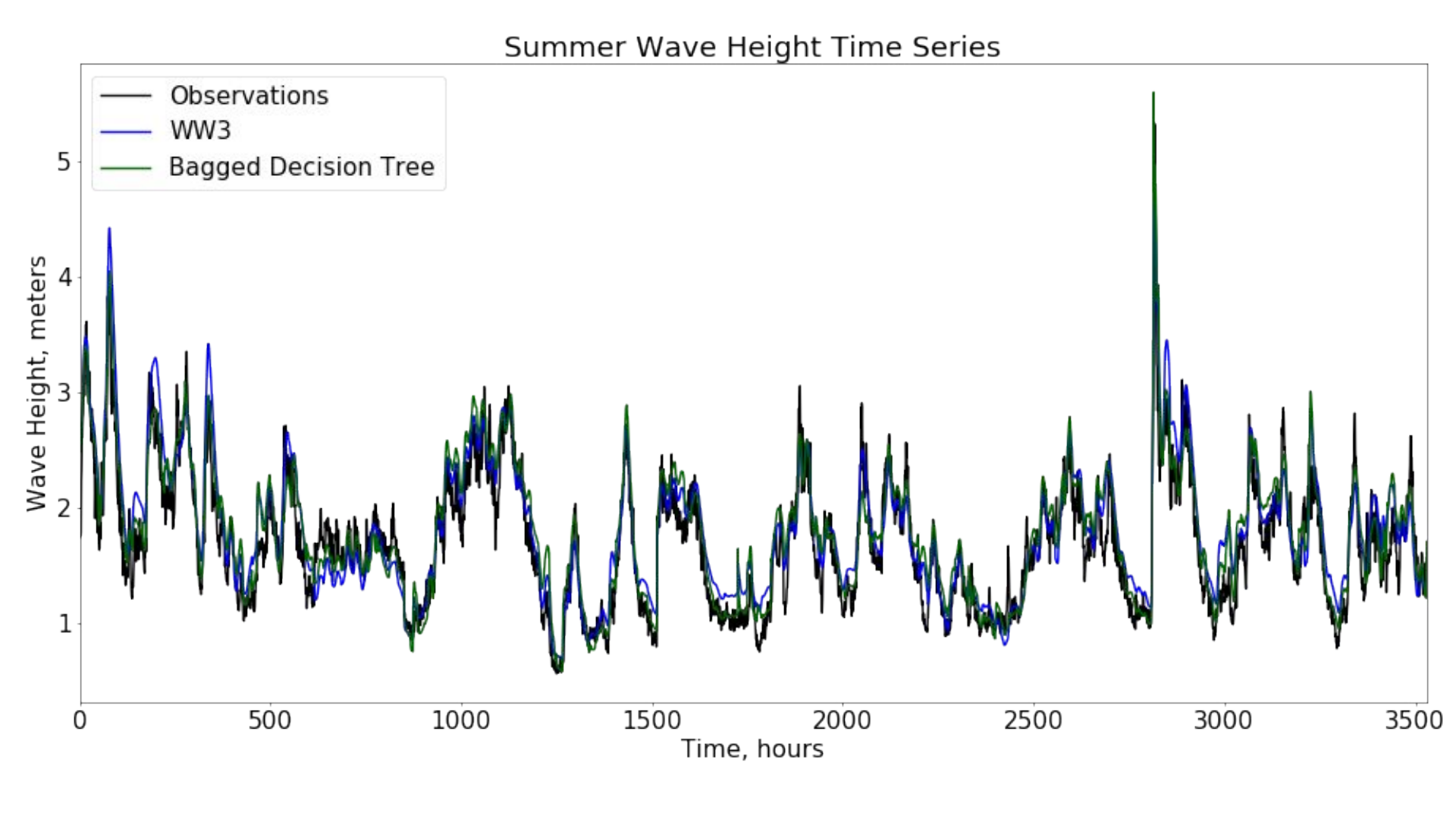
- 2015
- 2,767 points


Winter Wave Height Time Series
Observations
WW3
Bagged Decision Tree
Wave Height, meters
0
2
4
6
8
10
0
500
1000
1500
2000
2500
Time, hours

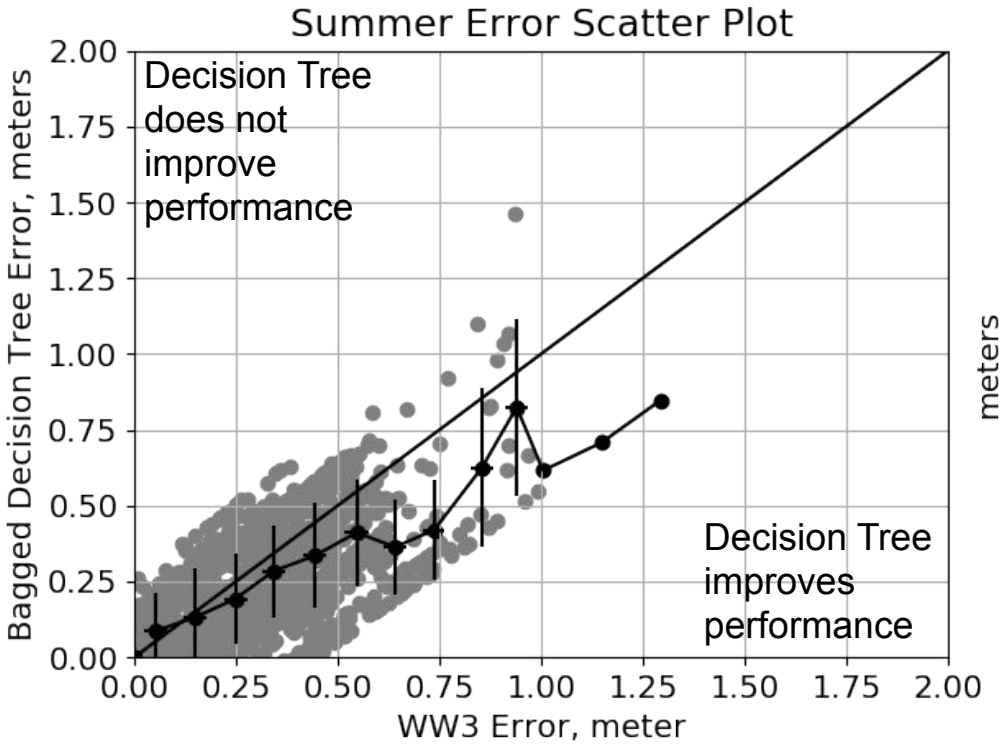
Results: Winter

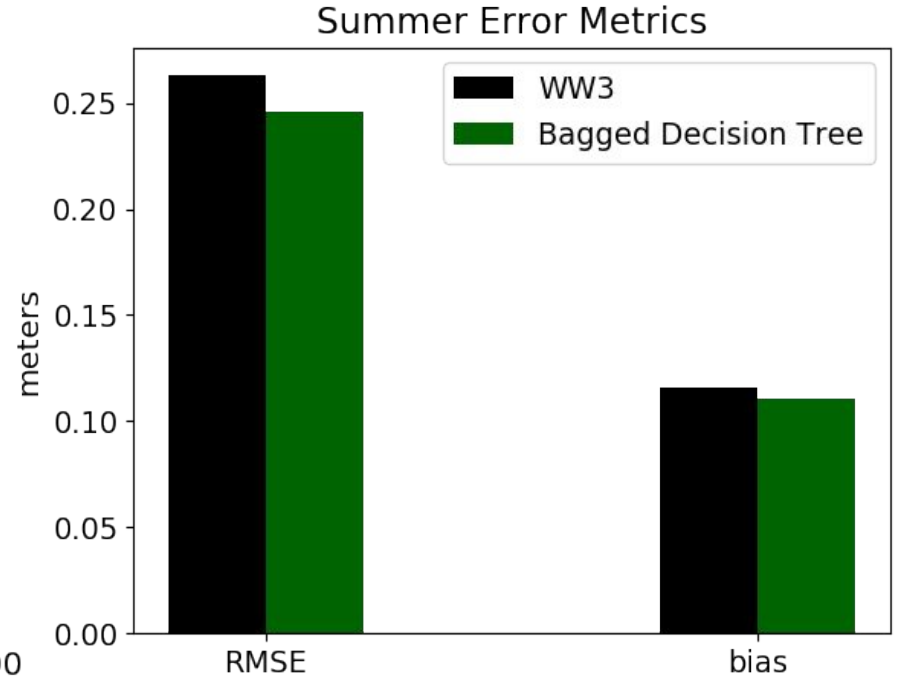





Summer Wave Height Time Series
Observations
WW3
Bagged Decision Tree
Wave Height, meters
1
2
3
4
5
0
500
1000
1500
2000
2500
3000
3500
Time, hours

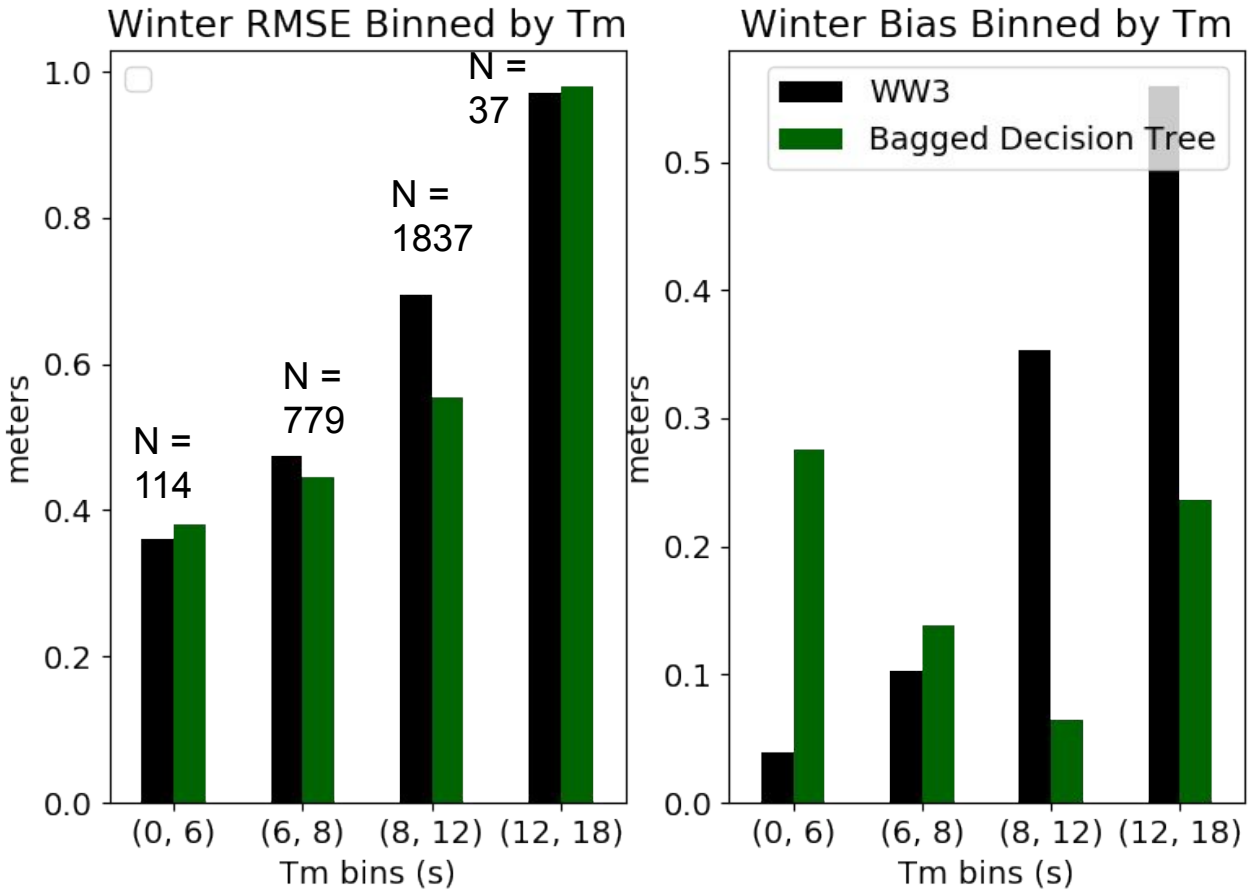
Results: Summer





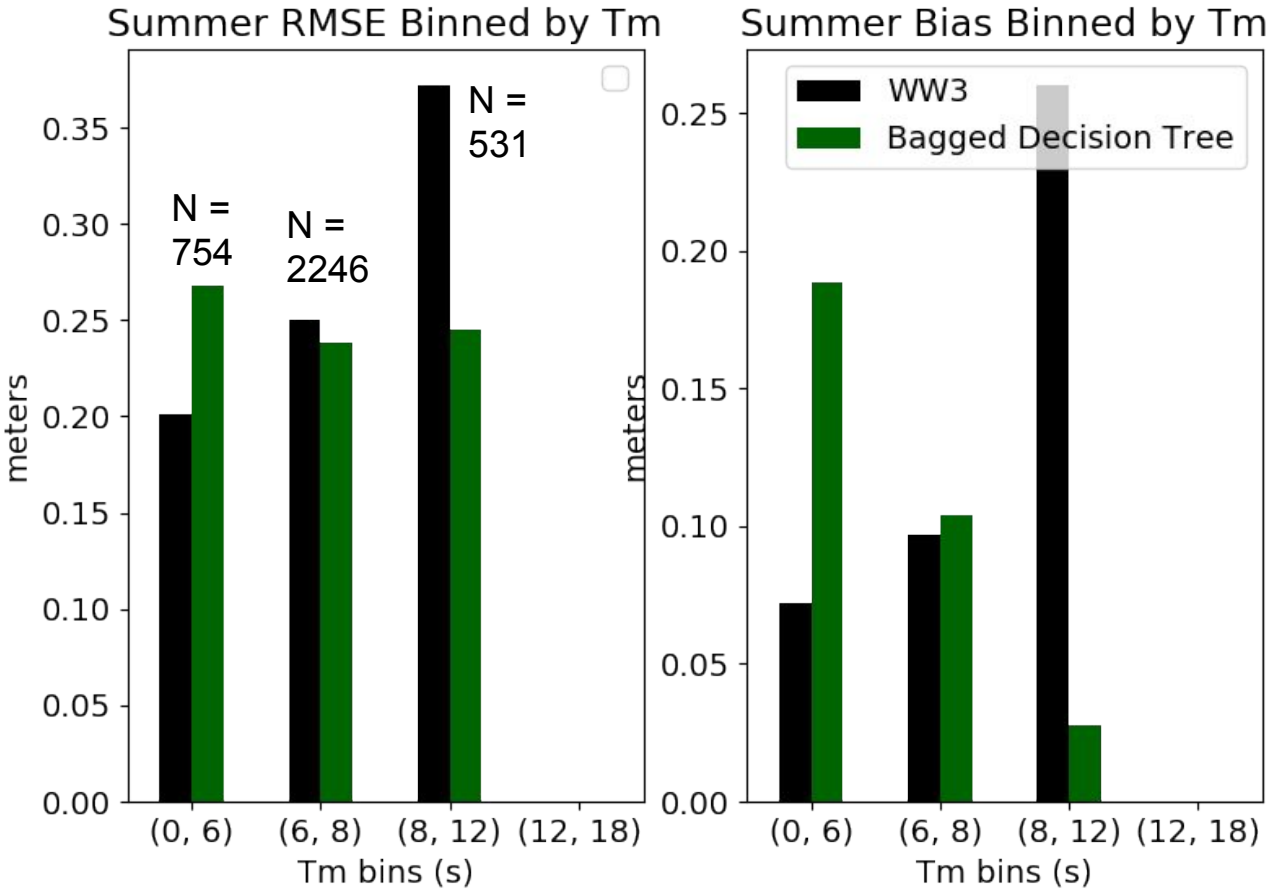
Results: Winter

- Decision Tree improves performance in mid-range periods (6-12s)
- Improves bias more than RMSE



Results: Winter

- Decision Tree improves performance in mid-range periods (6-12s)
- Improves bias more than RMSE



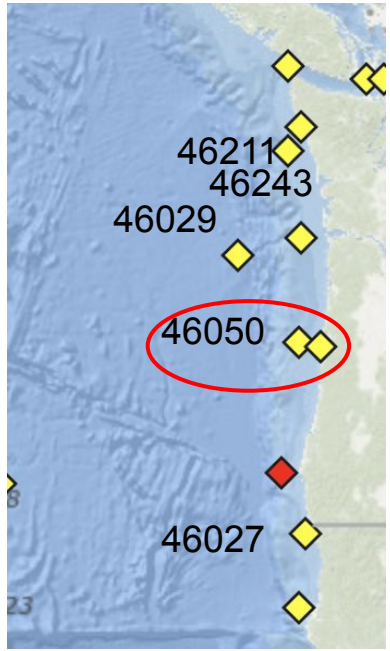
Can this be applied at different locations?

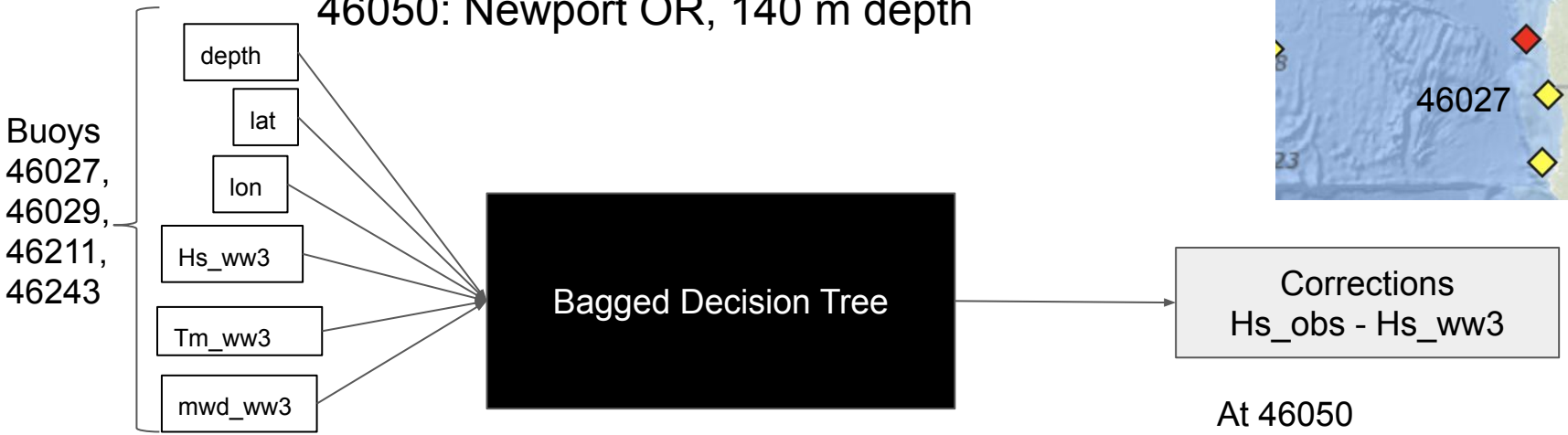
Train with data from other buoys:

46027: Crescent City, CA, 46 m depth
46029: Columbia River, OR, 134 m depth
46211: Greys Harbor, WA, 40 m depth
46243: Clatspot Spit, OR, 24.4 m depth

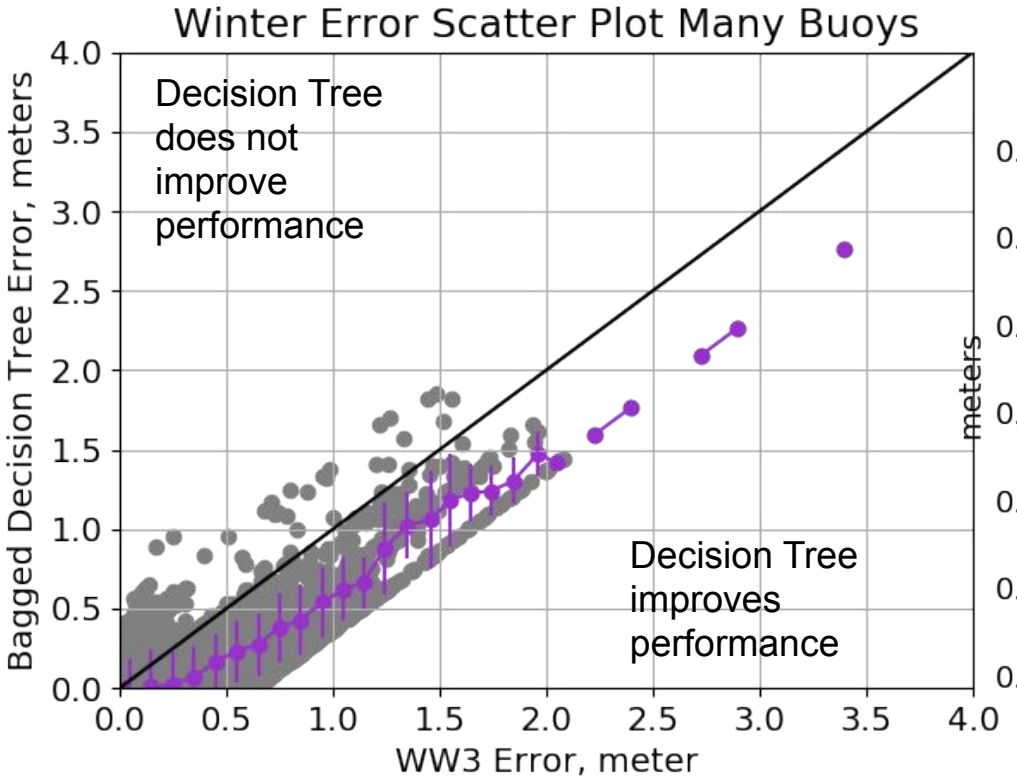
Test on

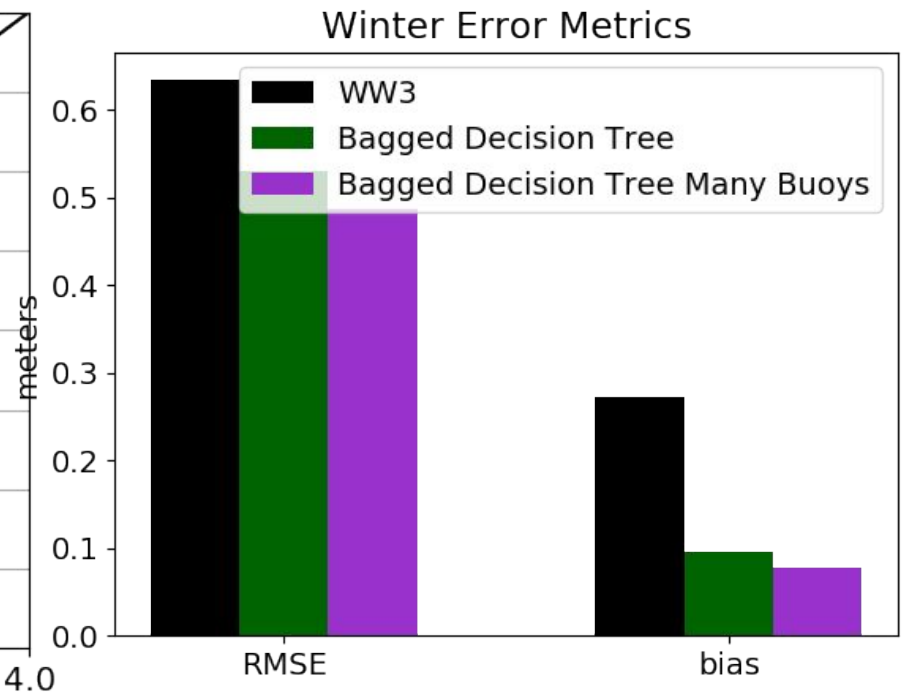
46050: Newport OR, 140 m depth



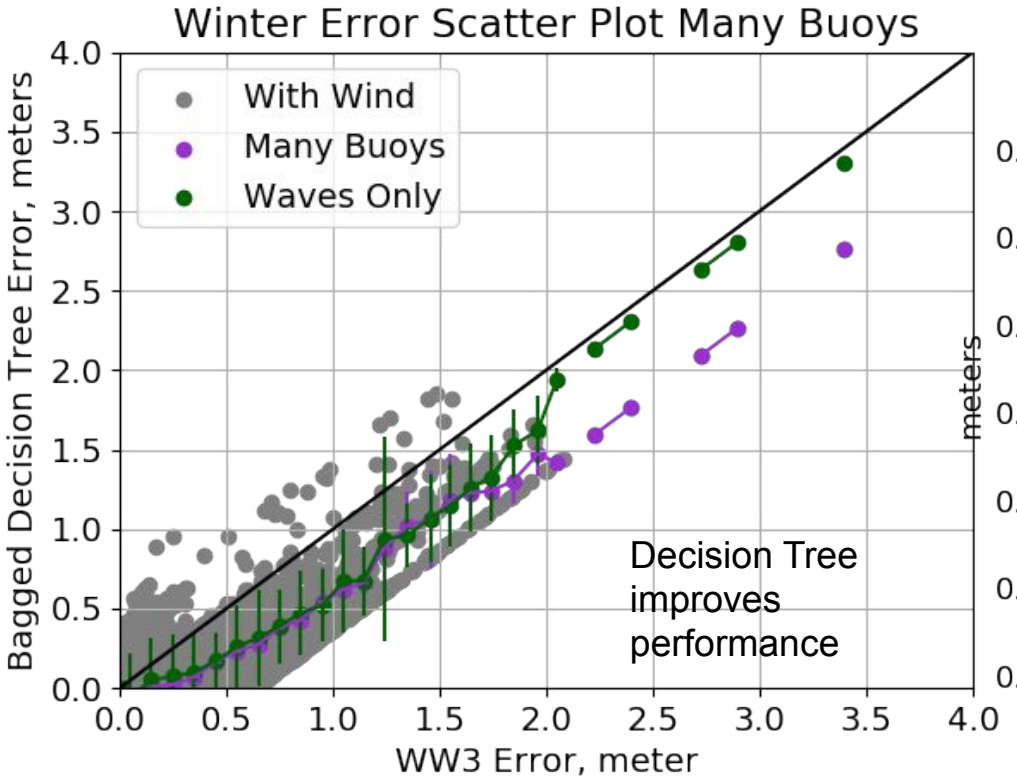


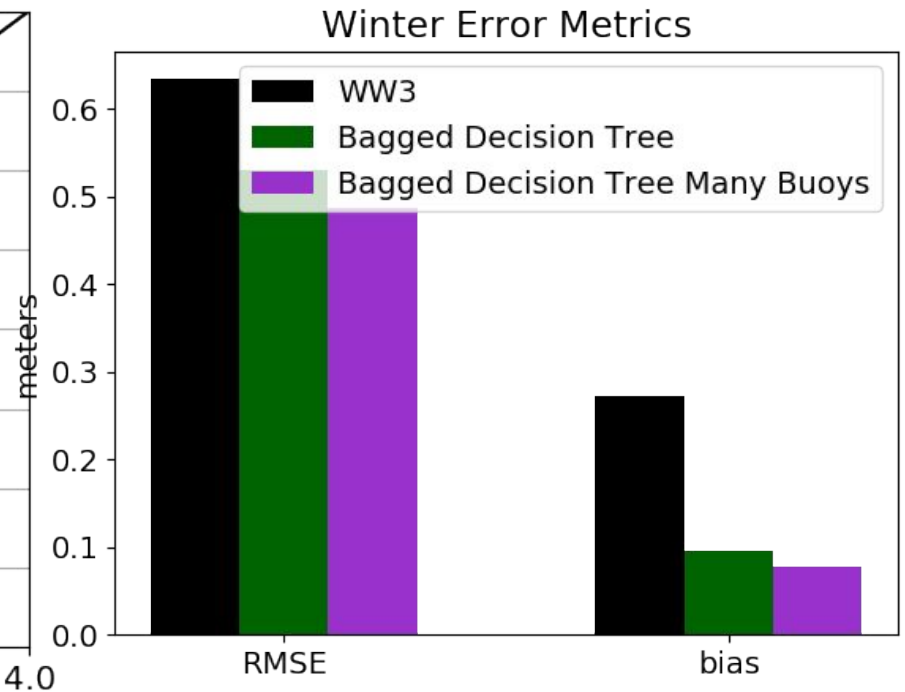
Results: Winter Many Buoys





Results: Winter Many Buoys

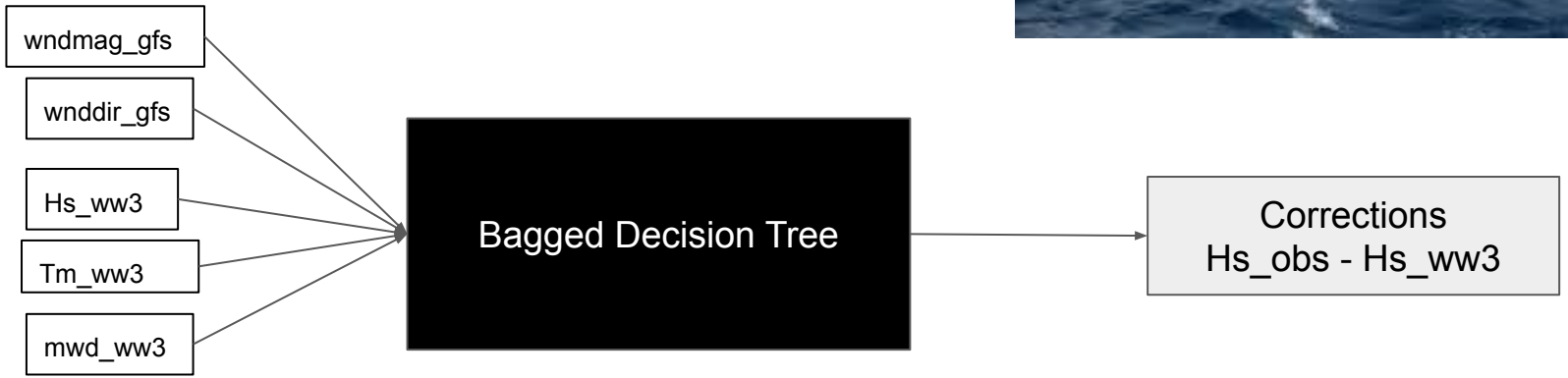




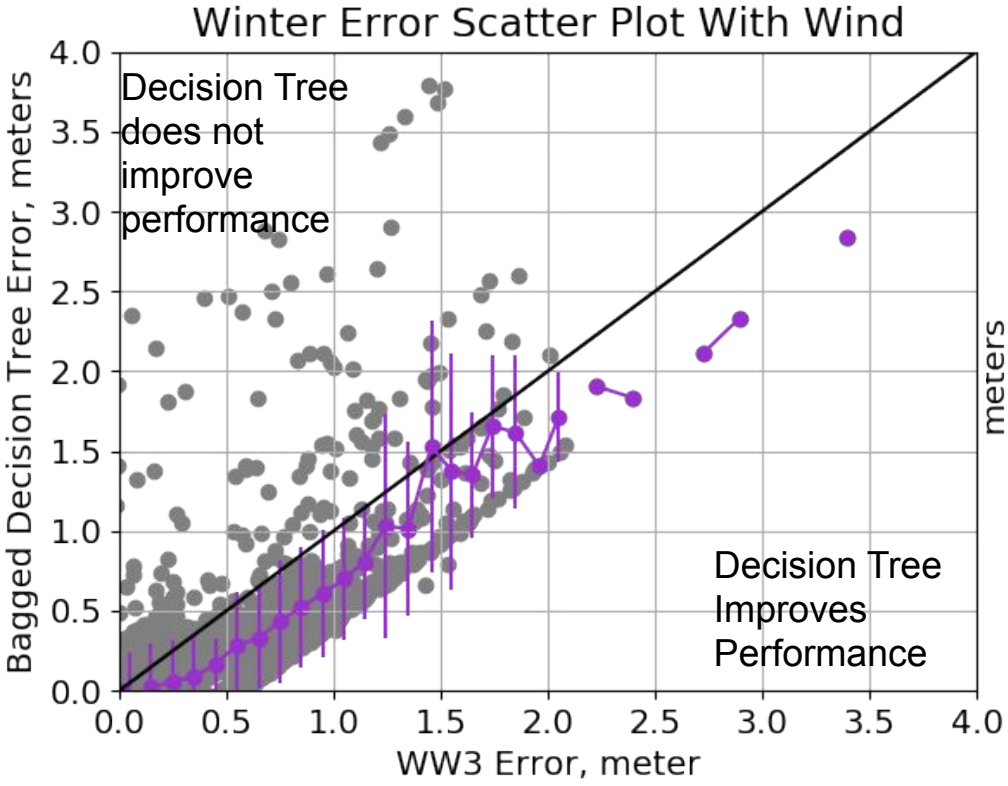
Can performance be further improved when more environmental information is included?

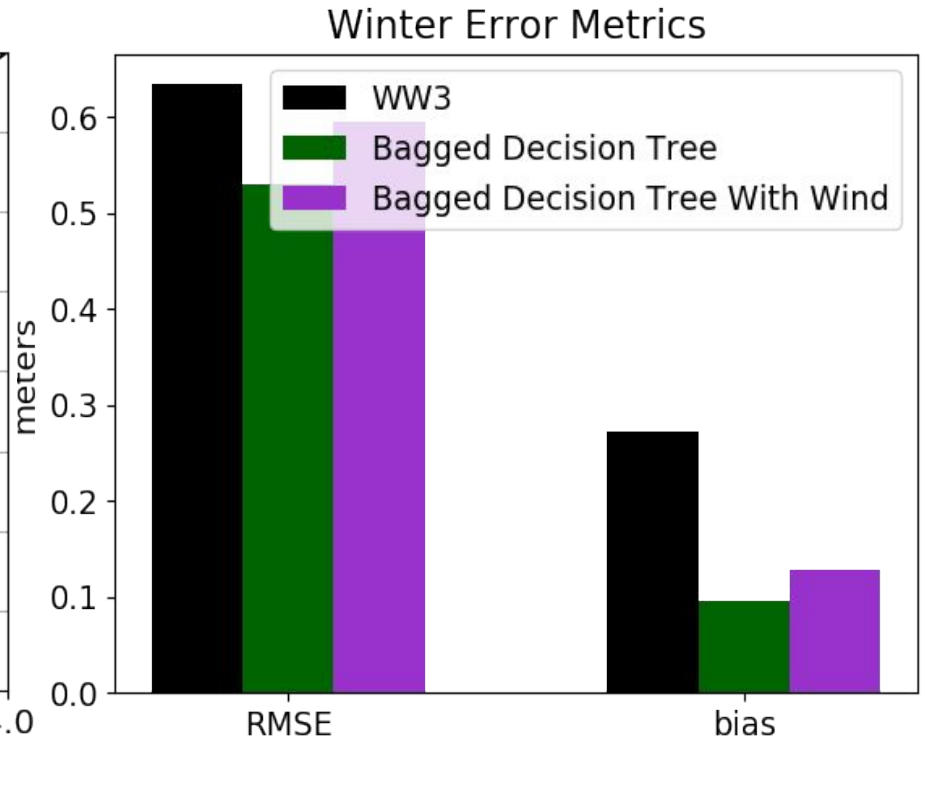
Information about wind forcing
- Wind direction
- Wind magnitude

Data Source: Global Forecast System (GFS) 24 hr forecast

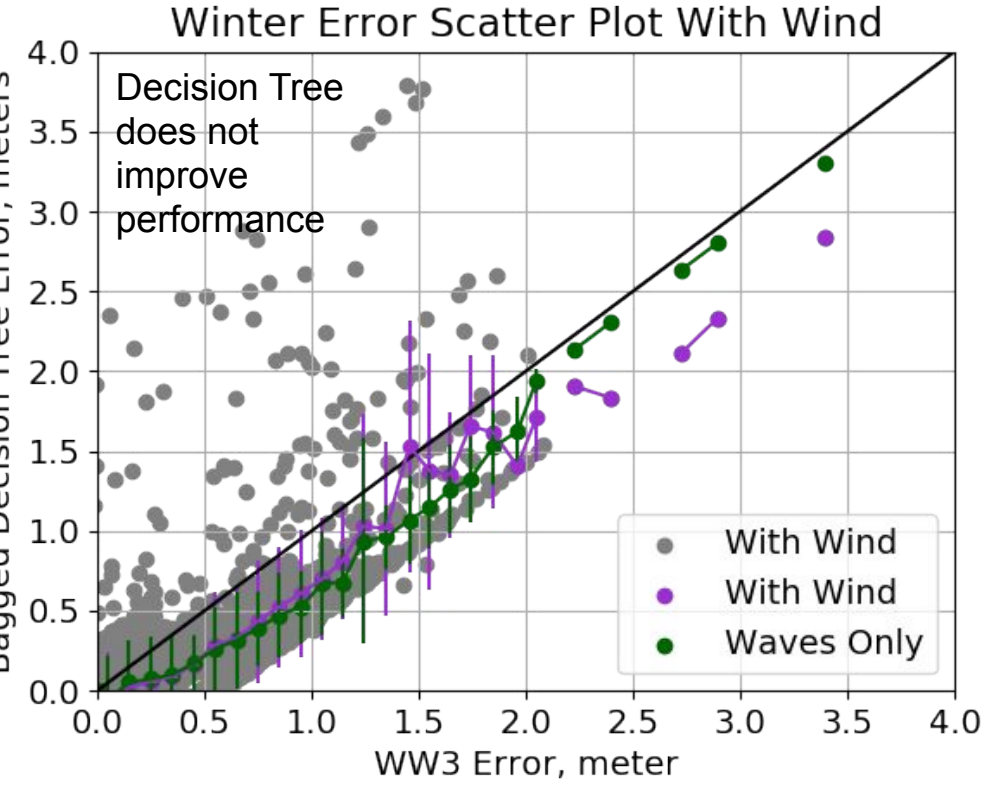


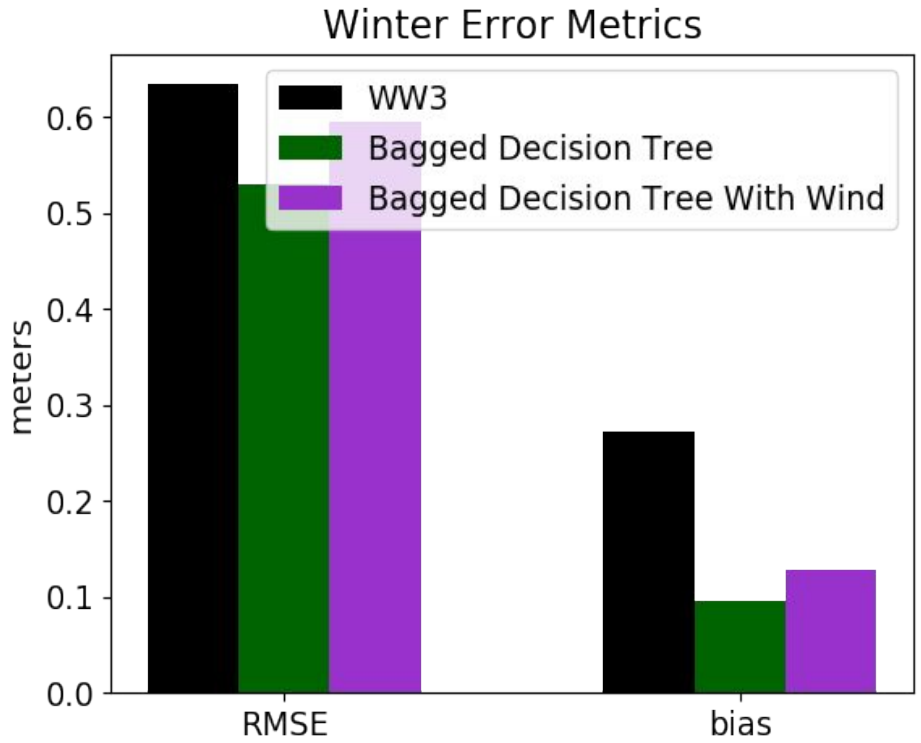
Results: Winter With Wind



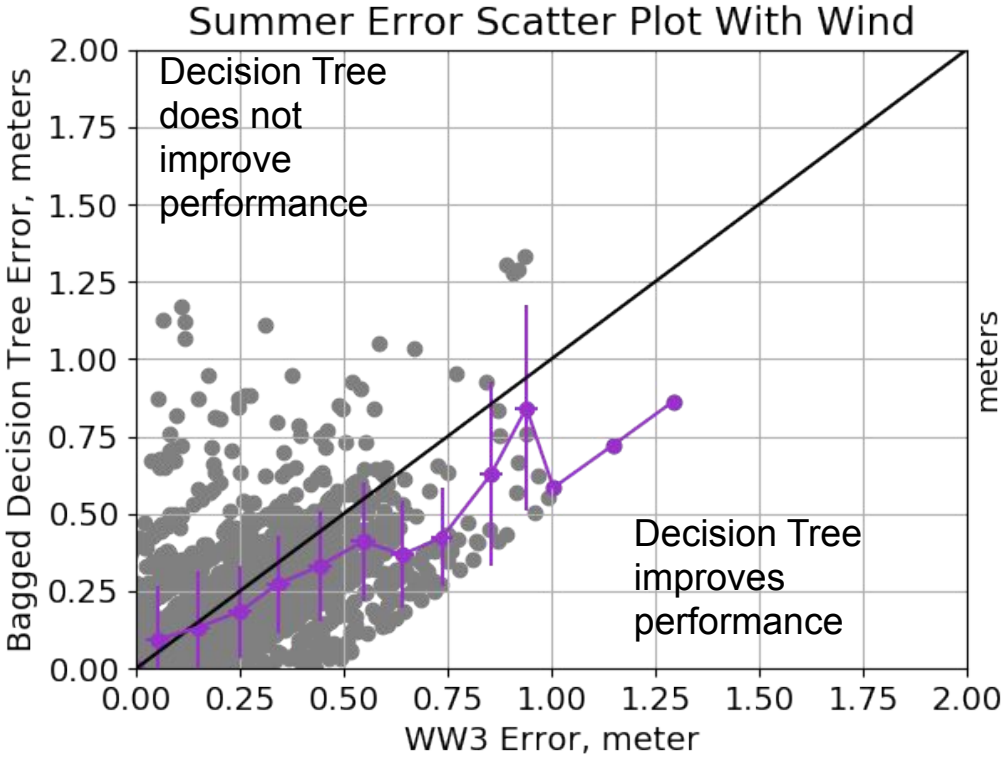


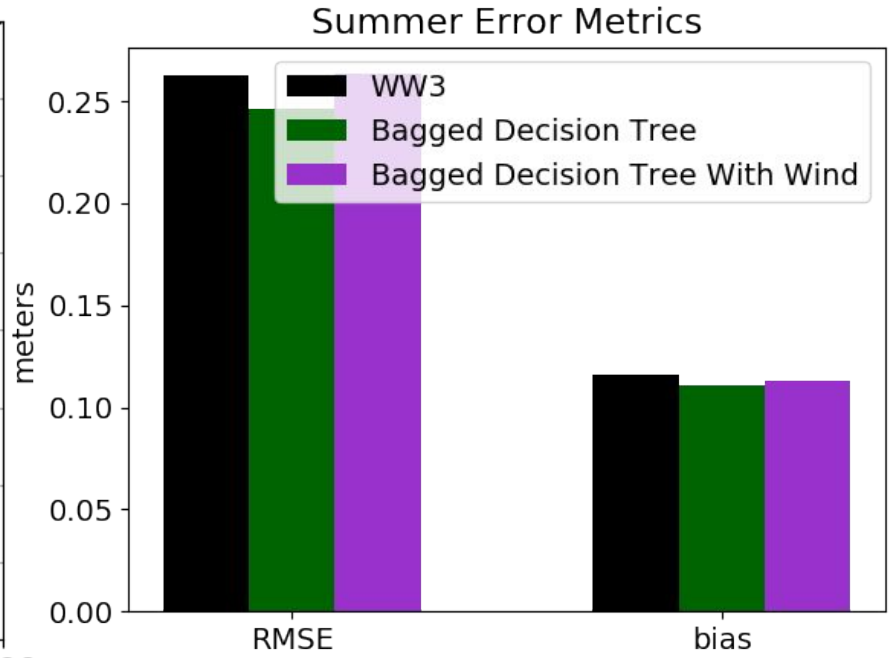
Results: Winter With Wind



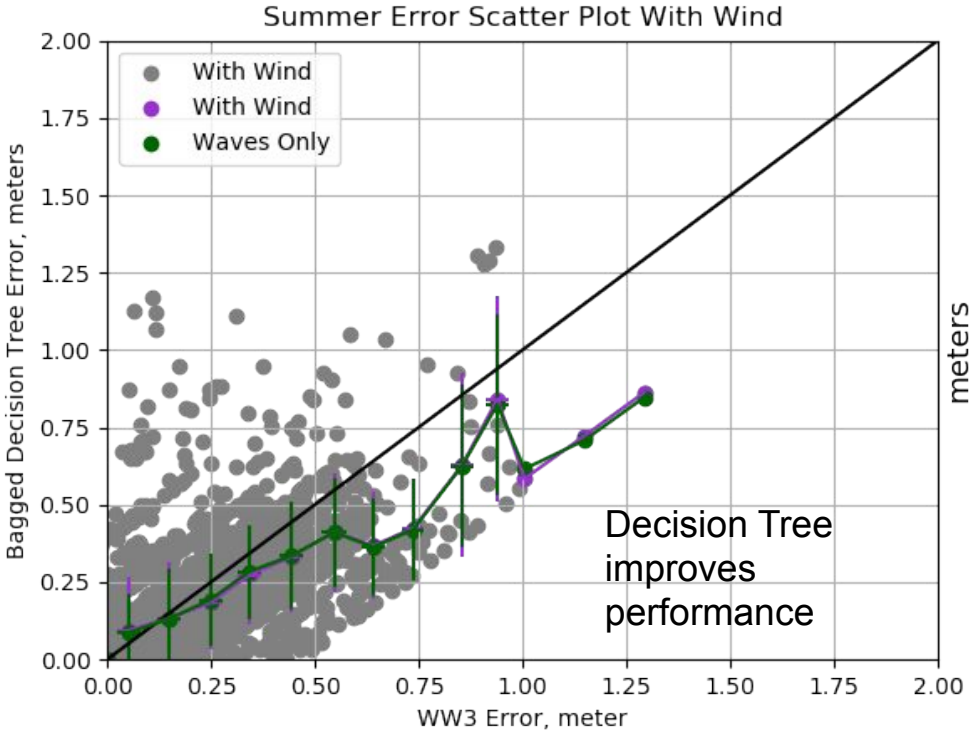


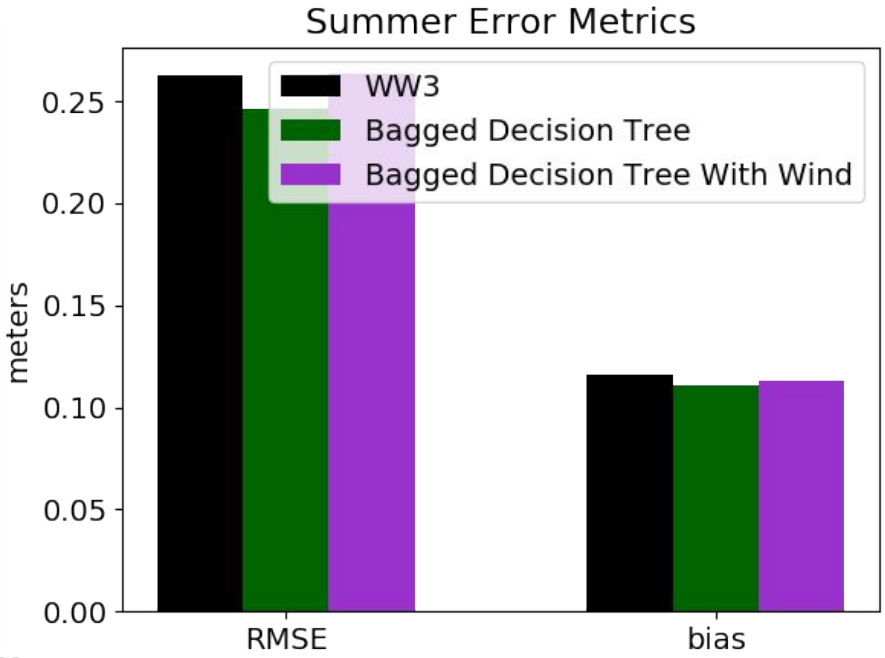
Results: Summer With Wind



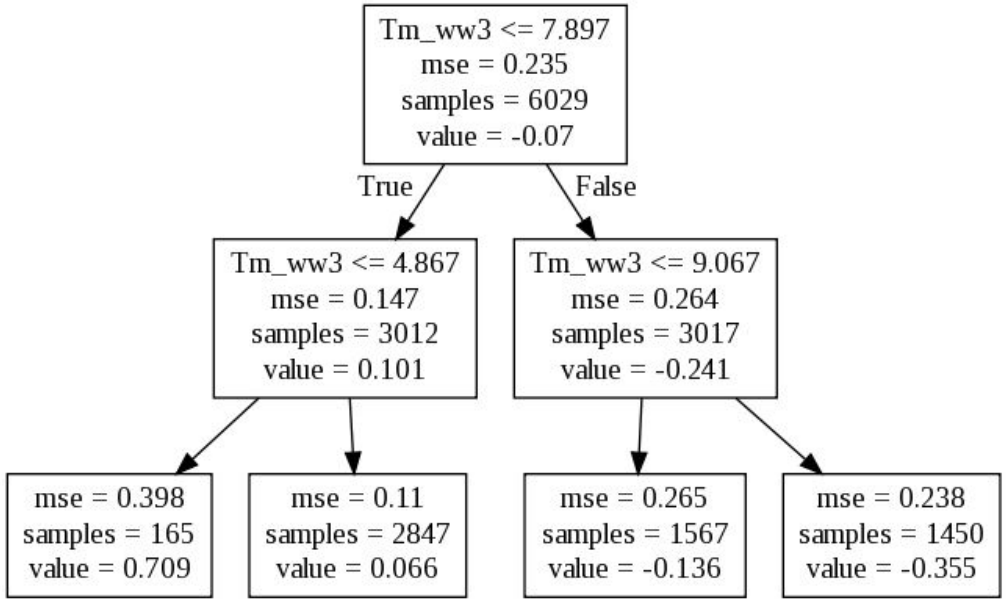


Results: Summer With Wind



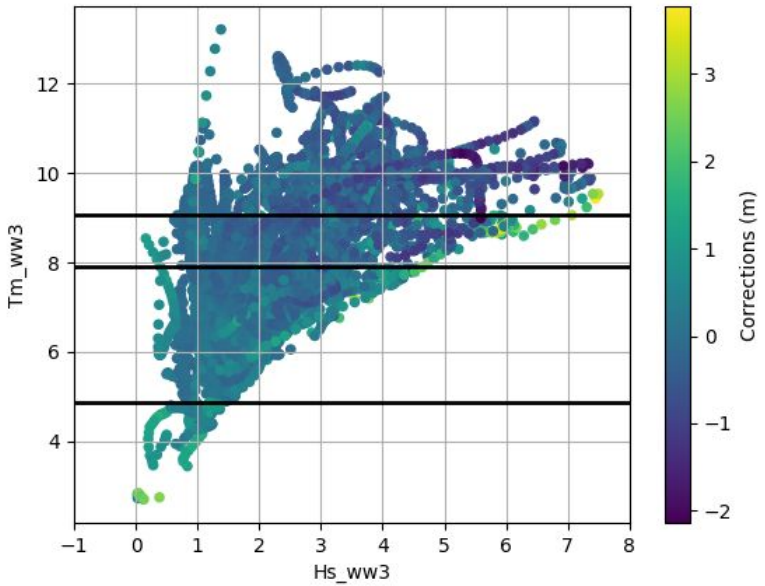


Feature Importances



- In this case, DT only splits on Tm, not Hs

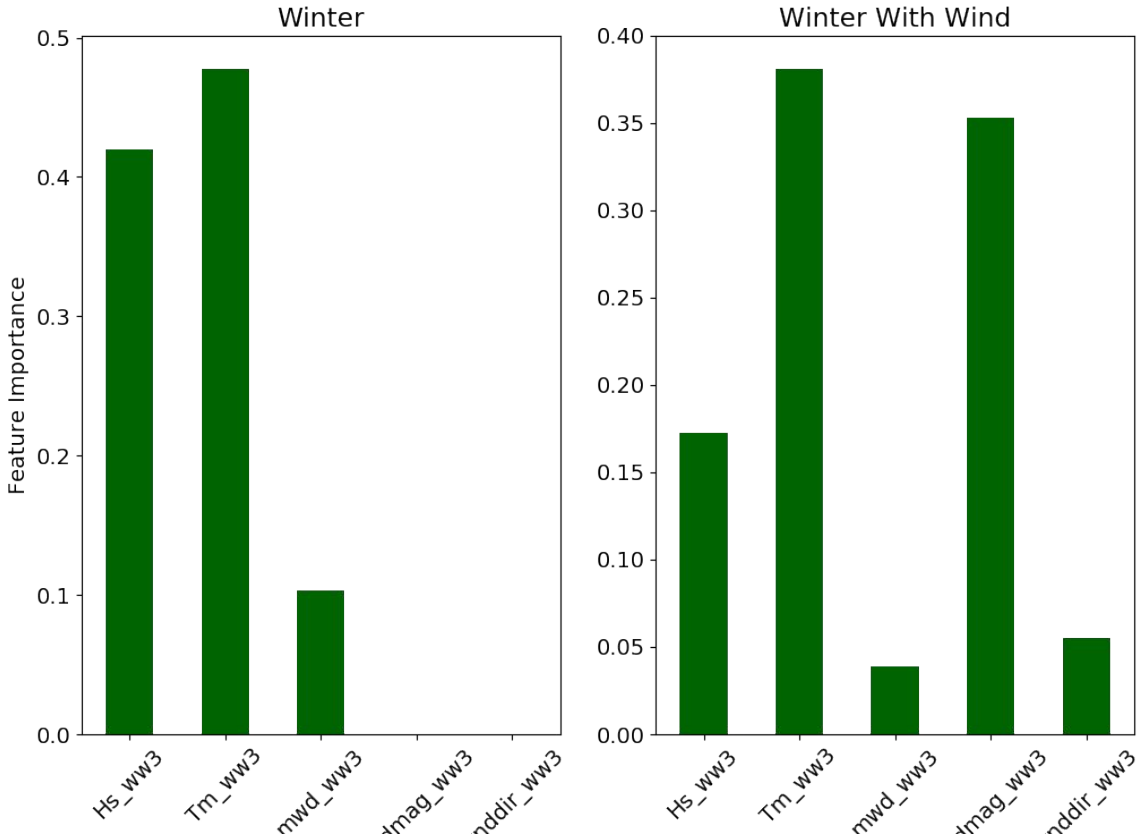
Feature Space



- Tm more important feature (feature importance of 1)

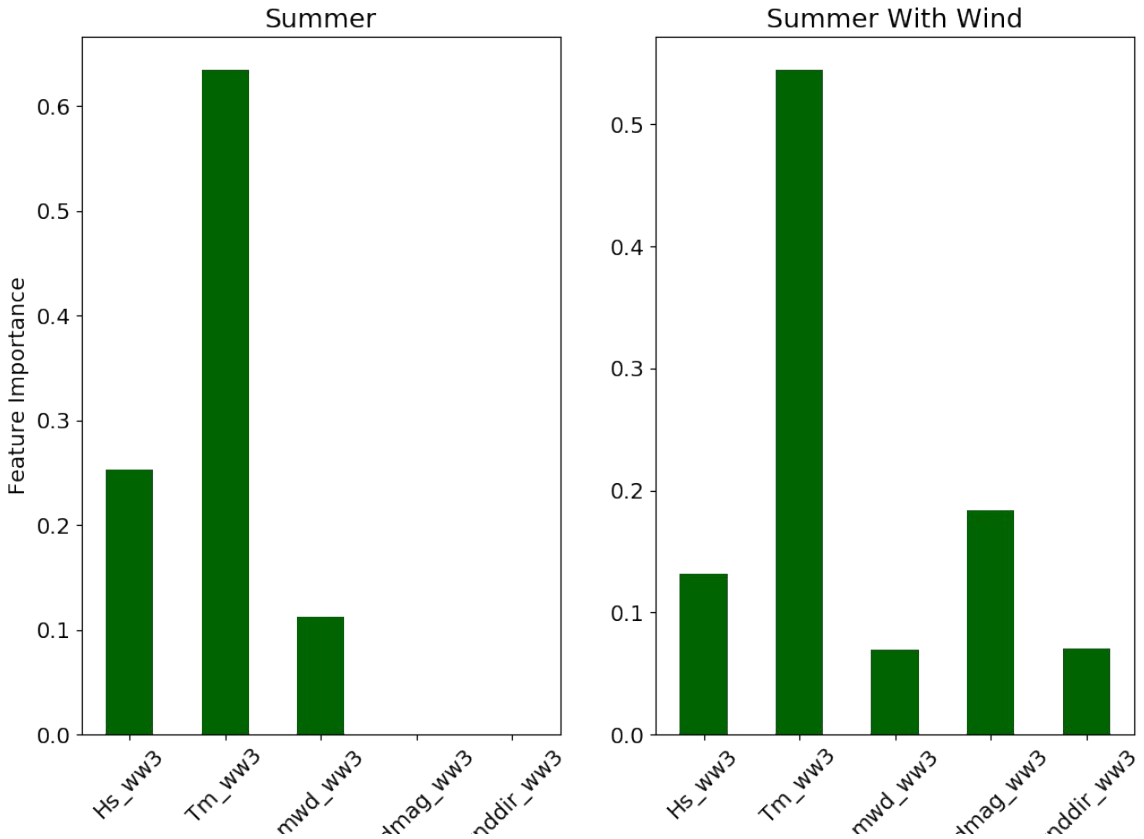
Feature Importances: Winter

- Mean period most important feature, Hs second
- When wind is added, wind magnitude becomes second most important feature

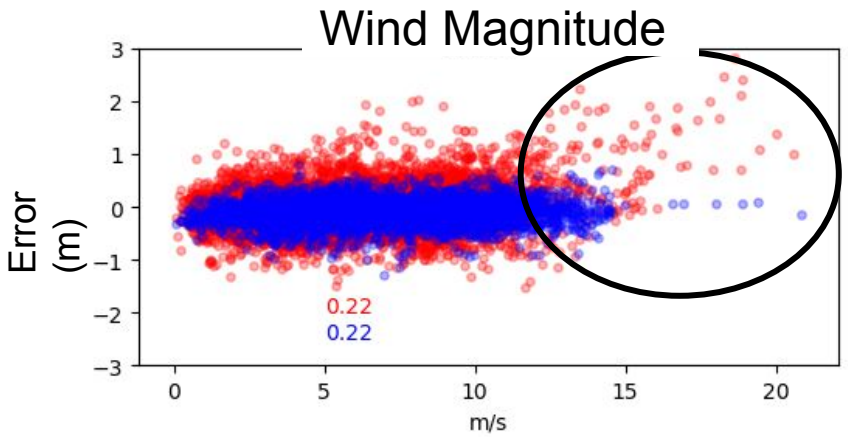


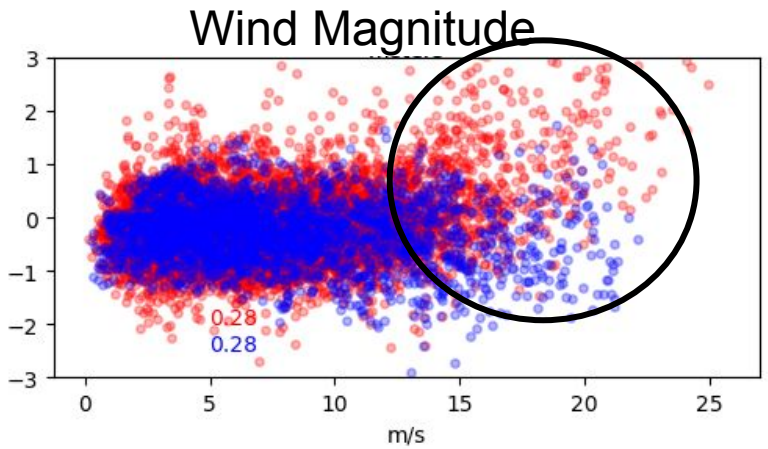
Feature Importances: Summer

- Mean period most important feature, Hs second
- When wind is added, wind magnitude becomes second most important feature



Training Data (2012-14) vs Testing Data (2015)





- Training data
- Testing data

Hypothesis: The training wind magnitude values result in DT trained on environmental conditions that do not arise in test conditions

For best performance, provide a training set that is representative of the testing conditions - best to give it median conditions

Conclusions

General:

- Random forest able to find systematic error within numerical model output
  - Evidenced by decreased RMSE and bias of post processed time series
- Performs better for winter data than for summer
  - Hypothesis that greater variability in the data provides more opportunity for the decision tree to partition the data space

WW3 Systematic Errors:

- Mean period is the most important parameter for Random Forest to partition the data space
  - error correlation higher for longer period than shorter period

Performance:

- Train it on data which is representative of the testing data
- Don’t extrapolate method out of training conditions