Predicting Lake Erie Wave using XGBoost

Haoguo Hu[1], Andre van der Westhuysen[3], Philip Chu[2], and Ayumi Fujisaki-Manome[1,4]

[1]Cooperative Institute for Great Lakes Research, the University of Michigan, Ann Arbor
[2]National Oceanic and Atmospheric Administration, Great Lakes Environmental Research Laboratory
[3]IMSG at National Weather Service, National Centers for Environmental Prediction Environmental Modeling Center
[4]Climate&Space Science and Engineering, the University of Michigan

3rd international workshop on waves, storm surges, and costal hazards
University of Notre Dame
Oct 1 - 6, 2023

Outline

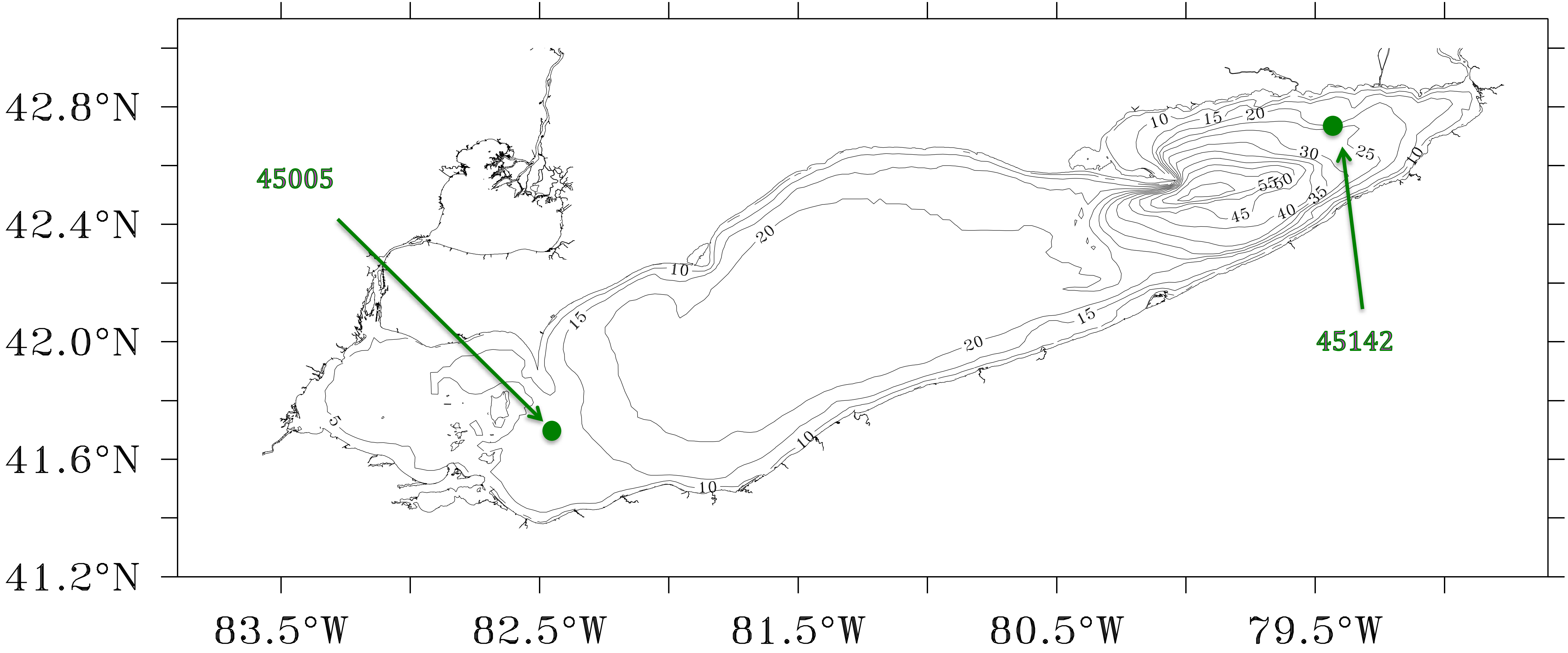
- Background and objectives
- WaveWatch III (WW3) and results and limitations
- XGBoost and results
- Conclusion

Limitations of physics-based model for wave forecast

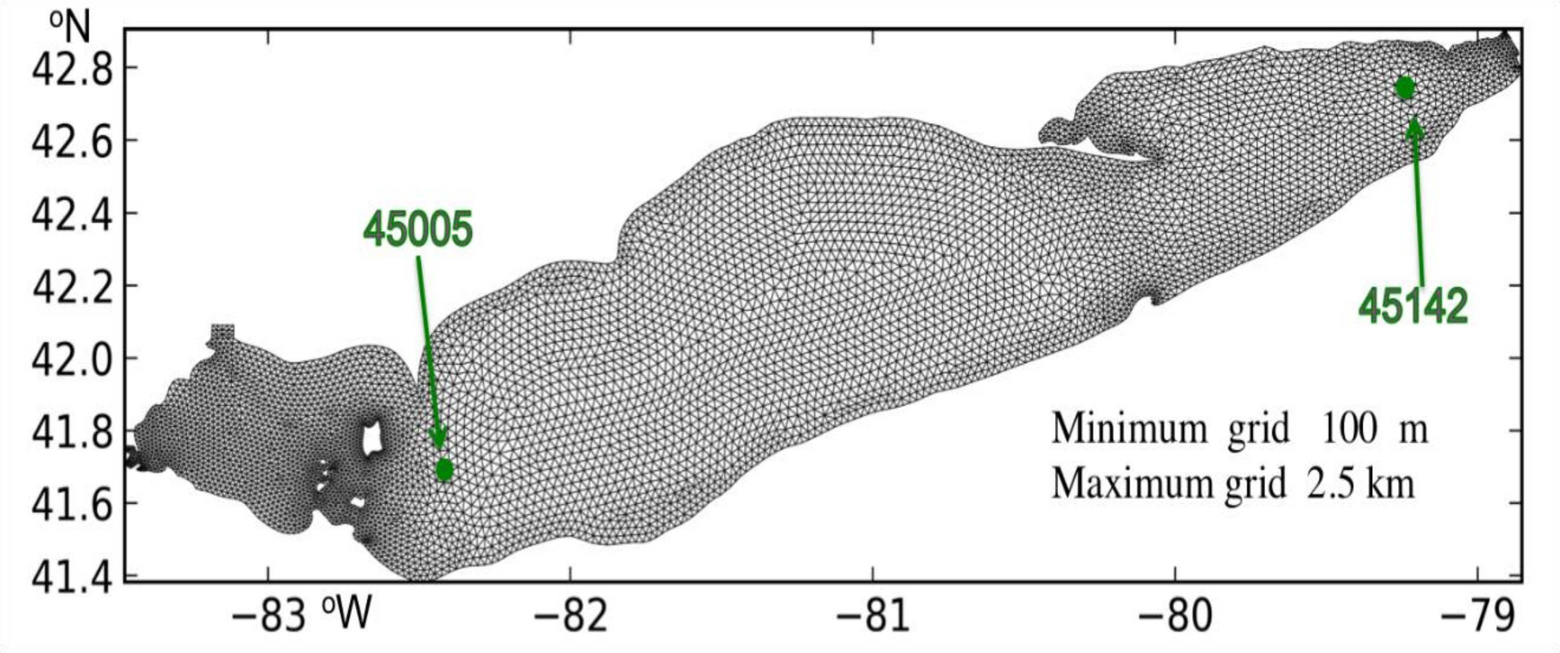
- Underestimate high wave during storm events
- Expensive computing time

WW3 Lake Erie case

Bathymetry and Buoy Stations



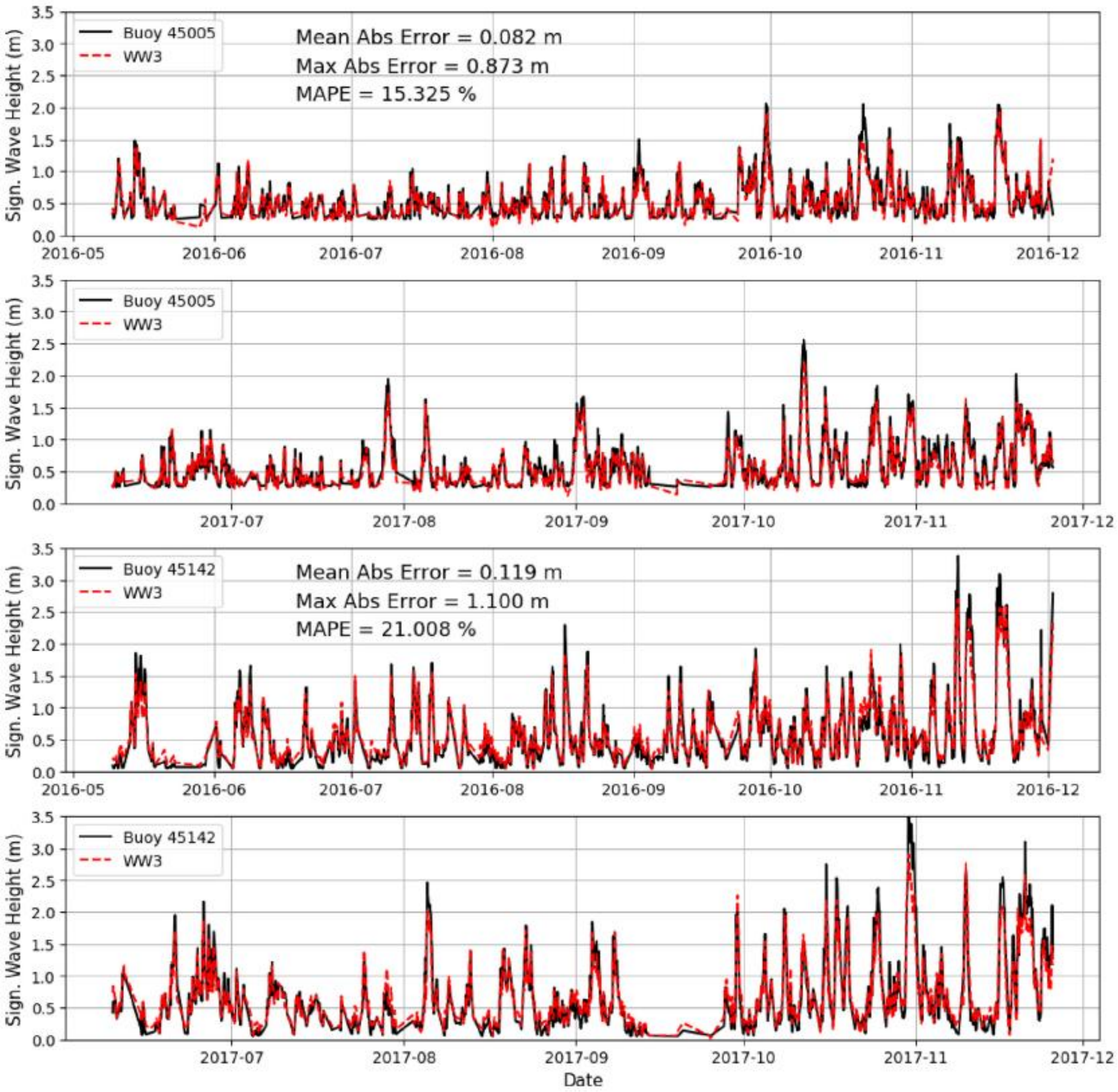
Unstructured grid for WW3



Node:6106, Cell:11509

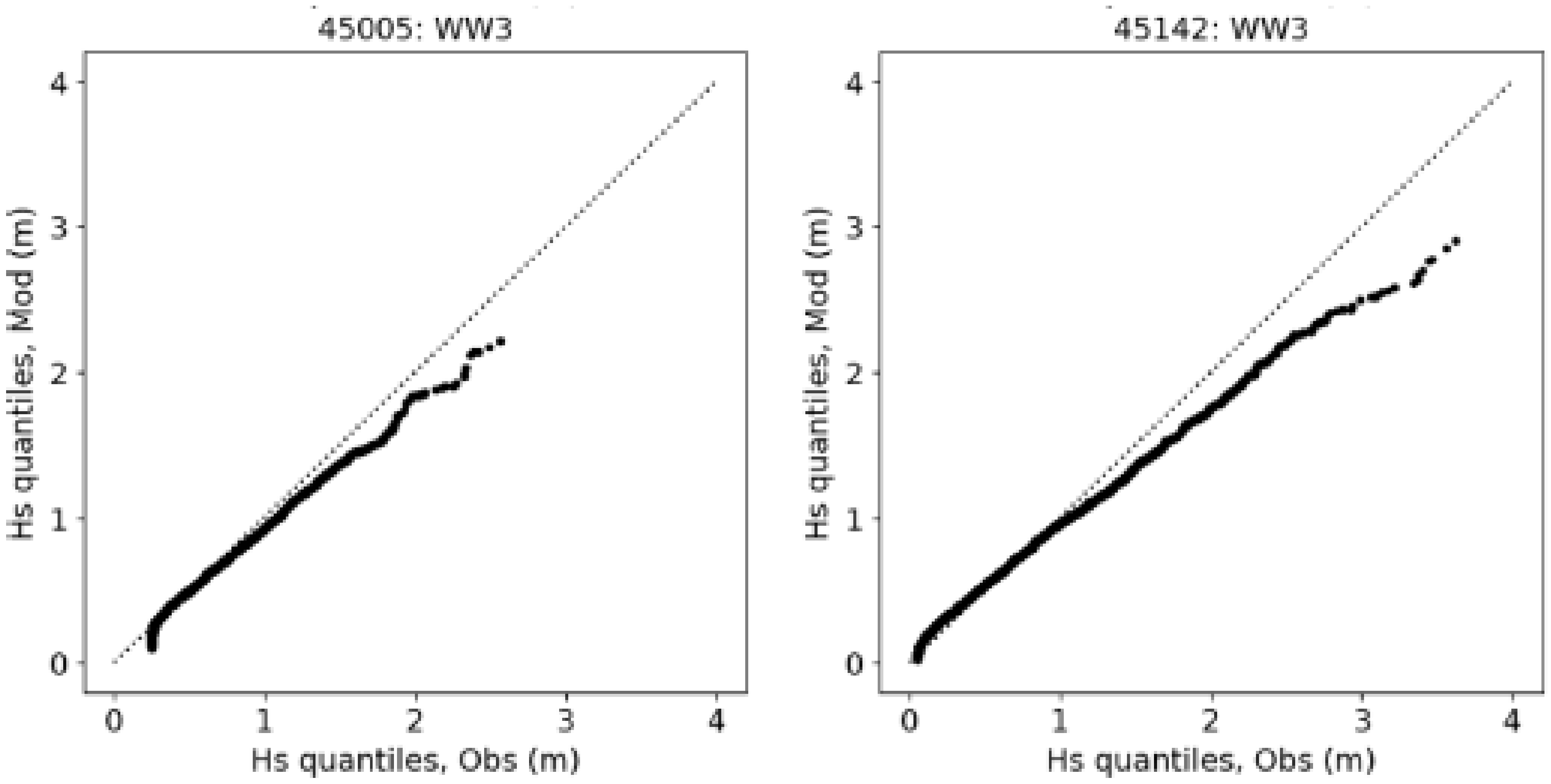
Wind forcing is interpolated from observations, including buoys and airport along the coast

Significant Wave Height (SWH) comparison: WW3 vs Observation



Computing time for one model year: 12 hours with 60 CPUs

SWH comparison: WW3 vs Observation

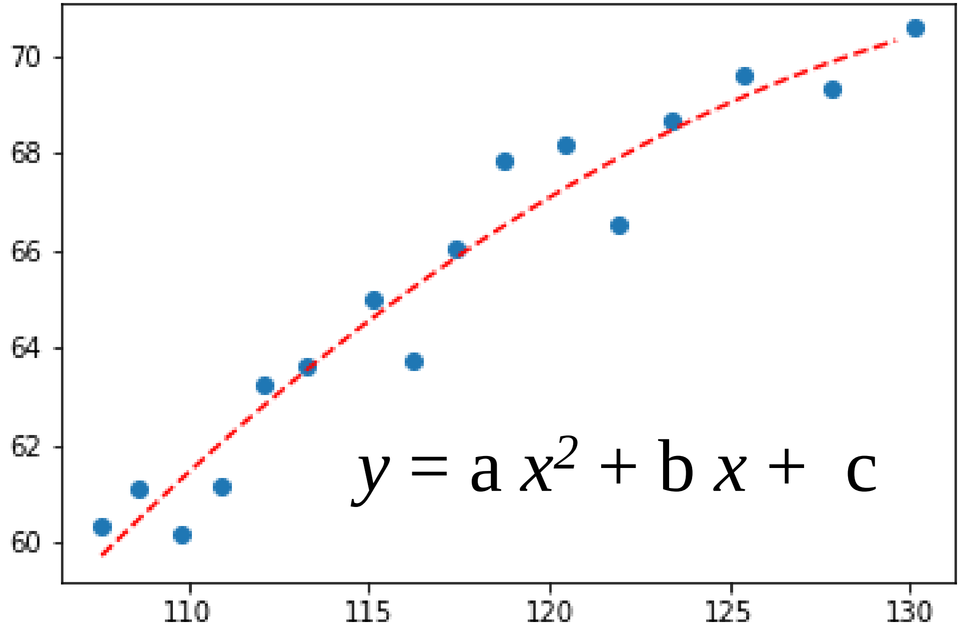


Limitations of WW3 for wave forecast

- Underestimate high wave during storm events
- Large computing time
- Tons of unused data. Data being used only for comparison, verification, and assimilation

- Machine learning is an attractive alterative approach

Data fitting and machine learning



Traditional data fitting:
Guess a function (here quadratic)
and then find a, b, c

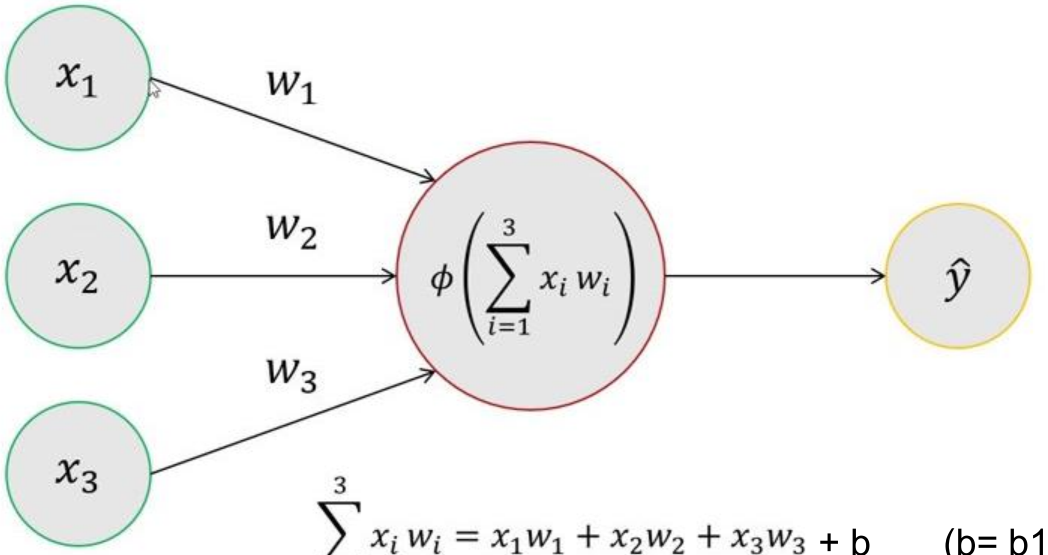
Machine/deep learning :
Don't have to guess a function,
Train a function to fit the data

How ? What is trained function
looks like ?

The training function in machine learning

A function expressed as sum of linear functions

$$
\begin{aligned}
Y = f(x) \;=\;& w_1 x_1 + b_1 \\
+\;& w_2 x_2 + b_2 \\
+\;& w_3 x_3 + b_3 \\
+\;& \ldots\ldots
\end{aligned}
$$



$$\sum_{i=1}^{3} x_i w_i = x_1 w_1 + x_2 w_2 + x_3 w_3 + b \qquad (b= b1+b2+b3+\ldots)$$

Taylor's theorem[4][5][6] — Let $k \geq 1$ be an integer and let the function $f : \mathbf{R} \to \mathbf{R}$ be k times differer

a function $h_k : \mathbf{R} \to \mathbf{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x -$$

and

$$\lim_{x \to a} h_k(x) = 0.$$

This is called the **Peano form of the remainder**.

Fourier Series ----- tides analysis

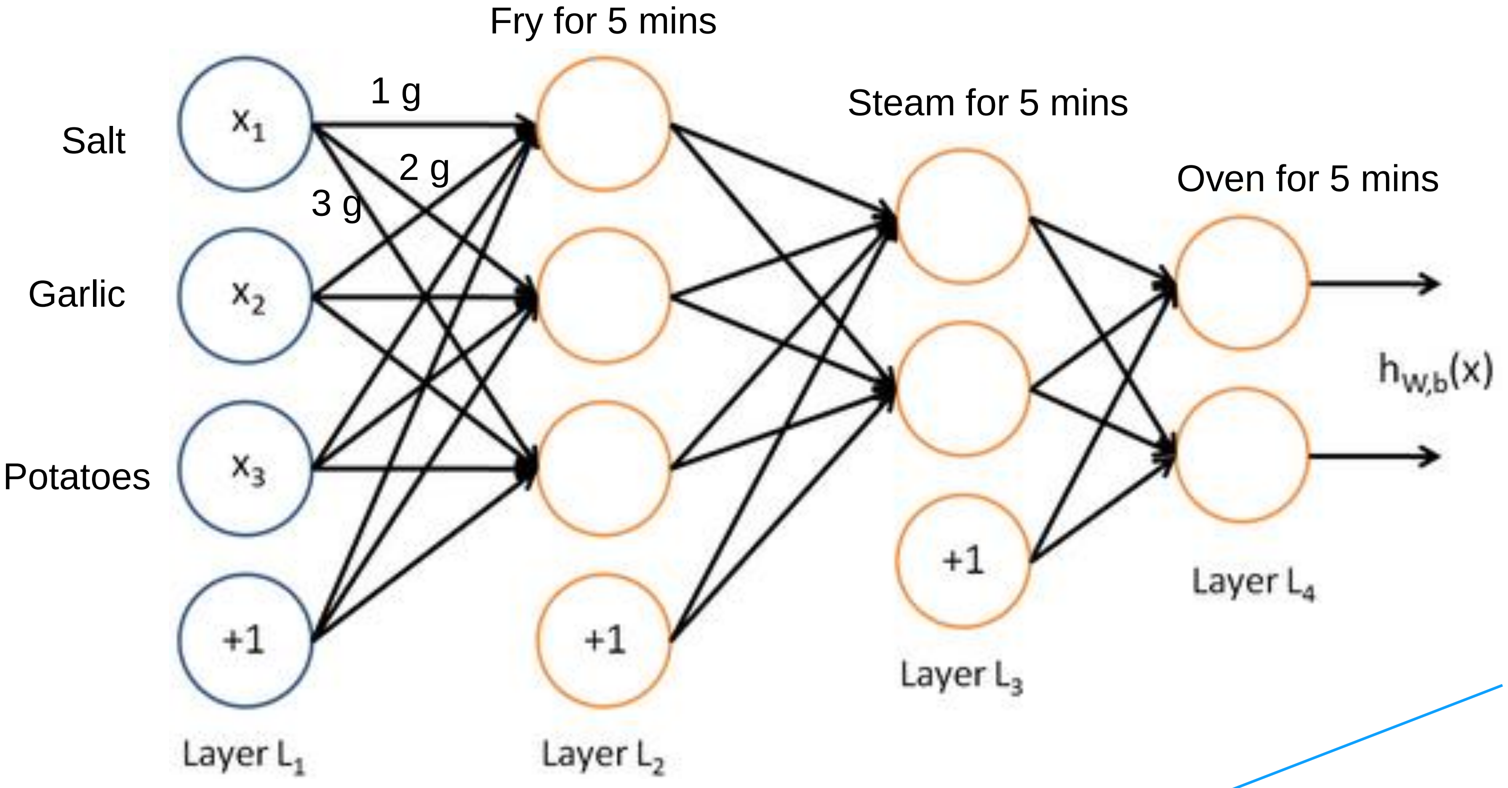
$$
\begin{aligned}
s(x) &= A_0 + \sum_{n=1}^{\infty} \left(A_n \cos(nx) + B_n \sin(nx)\right) \\
&= \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nx), \quad \text{for} \quad x - \pi \notin 2\pi\mathbb{Z}.
\end{aligned}
$$

To find the parameters – like backward engineering

Linear function, but with many variables and relationships

Deep learning = multiple layers

$$
\begin{aligned}
f(x) = {} & w_{11}\, x_1 + b_{11} + w_{12}\, x_1 + b_{12} \\
& + w_{21}\, x_2 + b_{21} + w_{22}\, x_2 + b_{22} \\
& + w_{31}\, x_3 + b_{31} + w_{32}\, x_3 + b_{23} \\
& + \ldots\ldots
\end{aligned}
$$



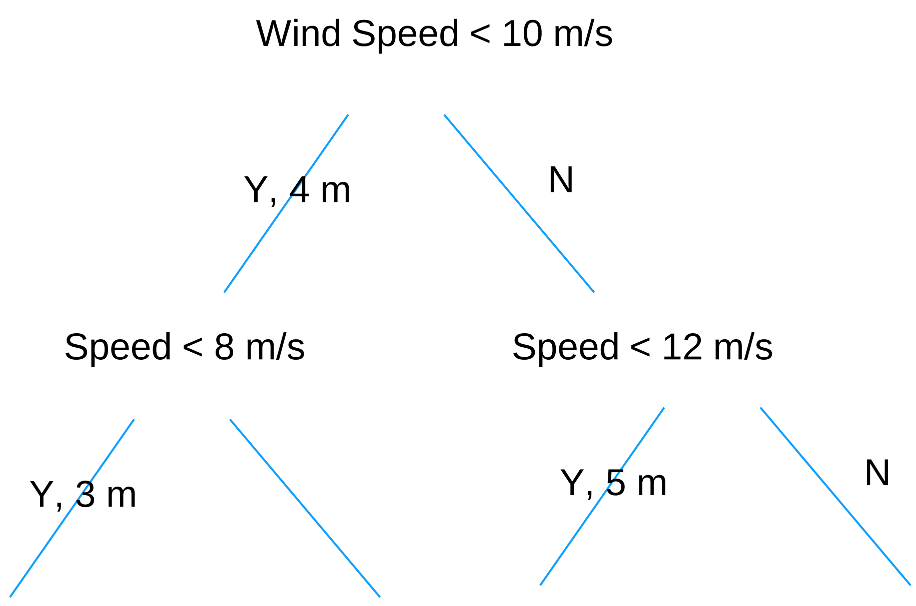
Final dish

Compare to

Tasty dish
(Data)

Keep change parameters with back-propagation

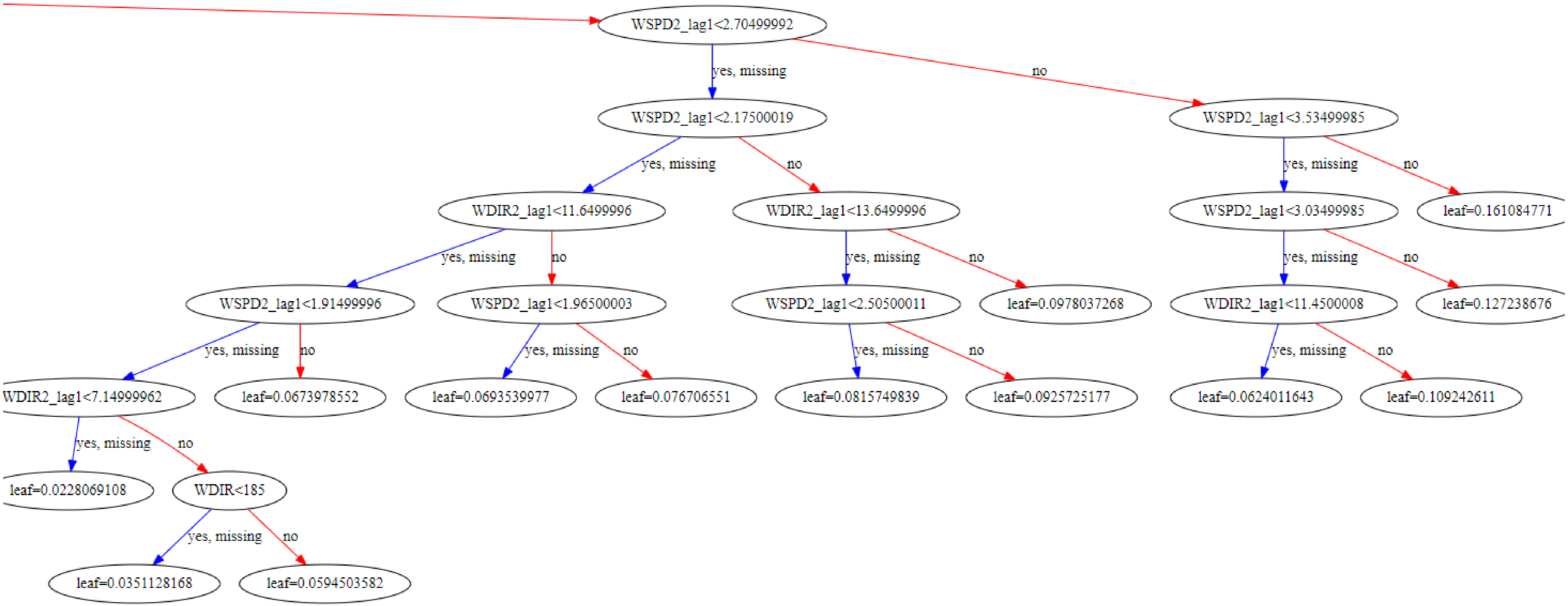
The training Function can also be a tree function



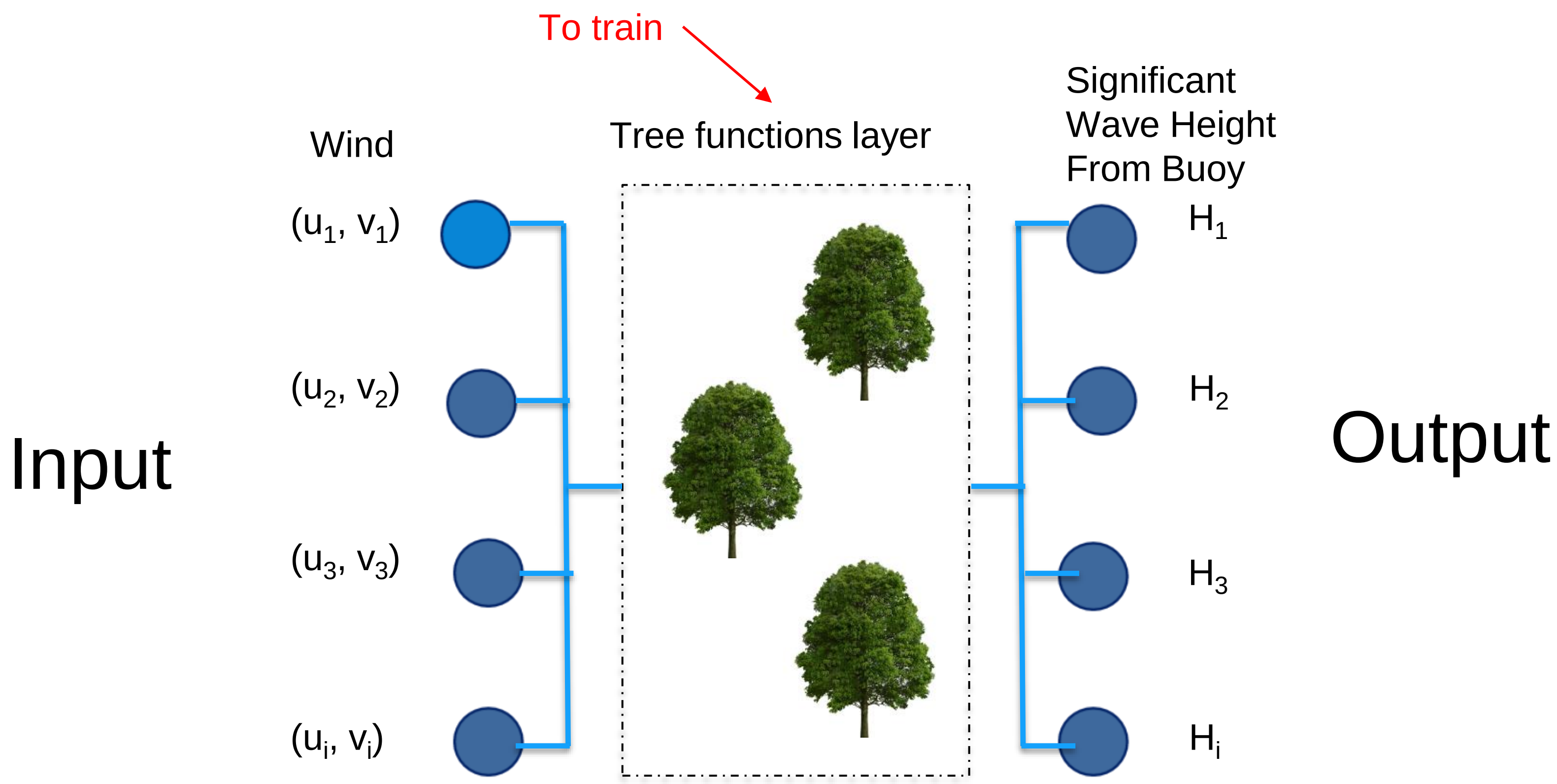
If input a 11 m/s wind, what the possible wave height might be?

A snapshot of the trees in Lake Erie case of XGBoost

How many trees and how many branches needed ? Solved automatically by XGBoost model



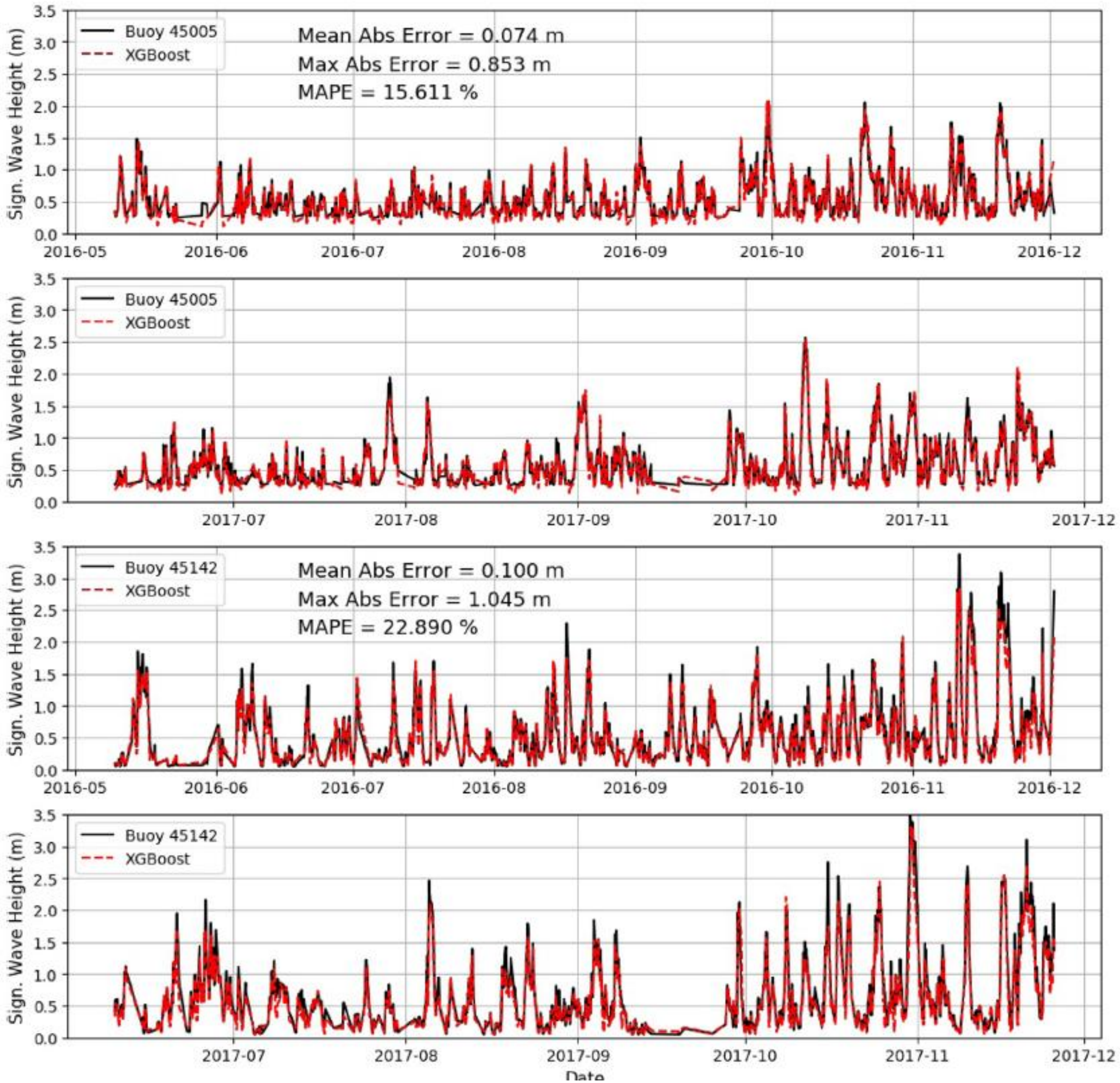
XGBoost (Extreme Gradient Boost tree) for Lake Erie wave prediction



XGBoost function trained using the same wind for WW3

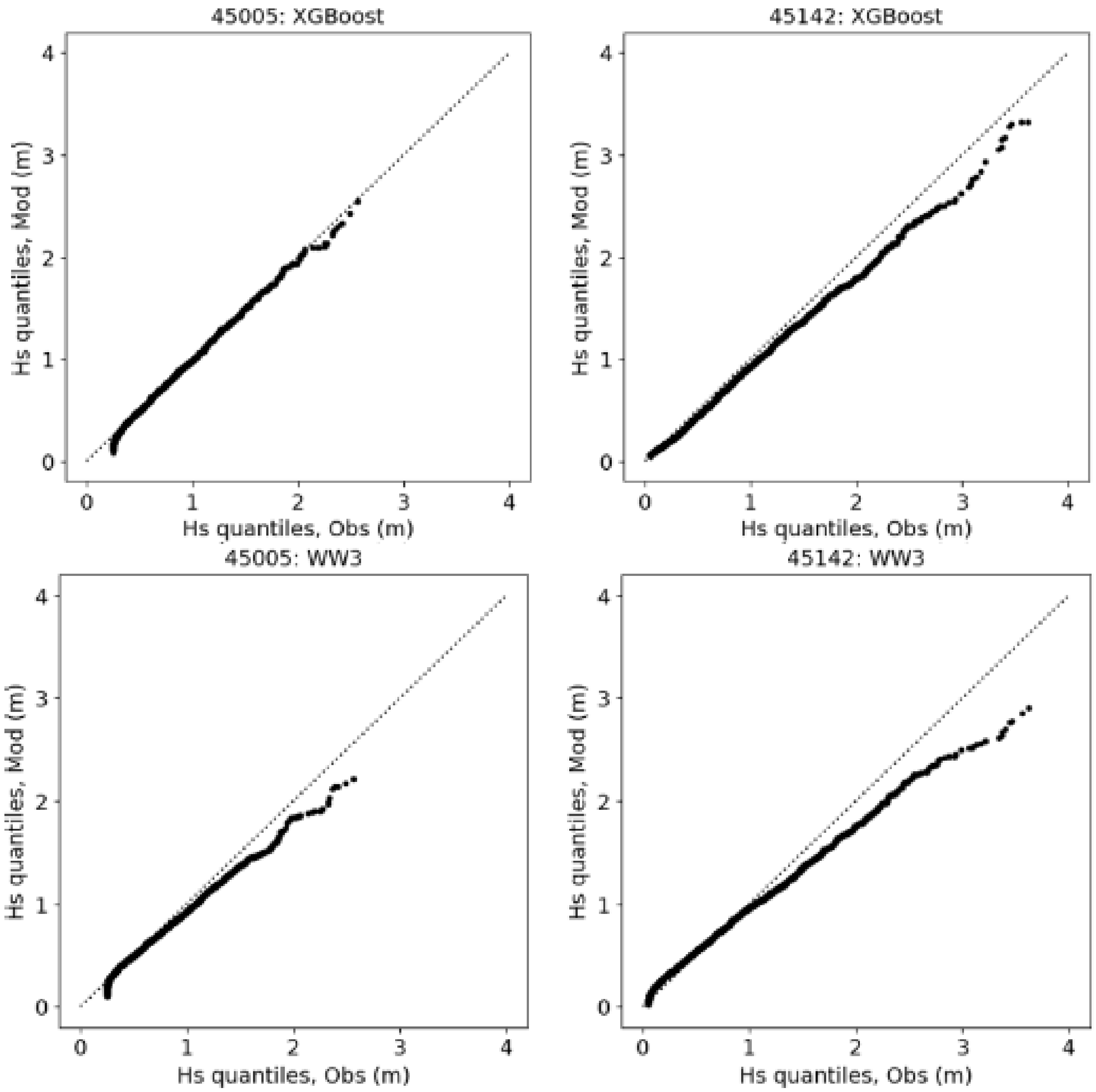
Training data: 1995-2015 buoy data. Prediction:2016-2017
training time: ~3 minutes with 1 CPU

SWH comparison: XGBoost vs Observation



SWH comparison: XGBoost vs WW3

Y axis: Model predicted wave height



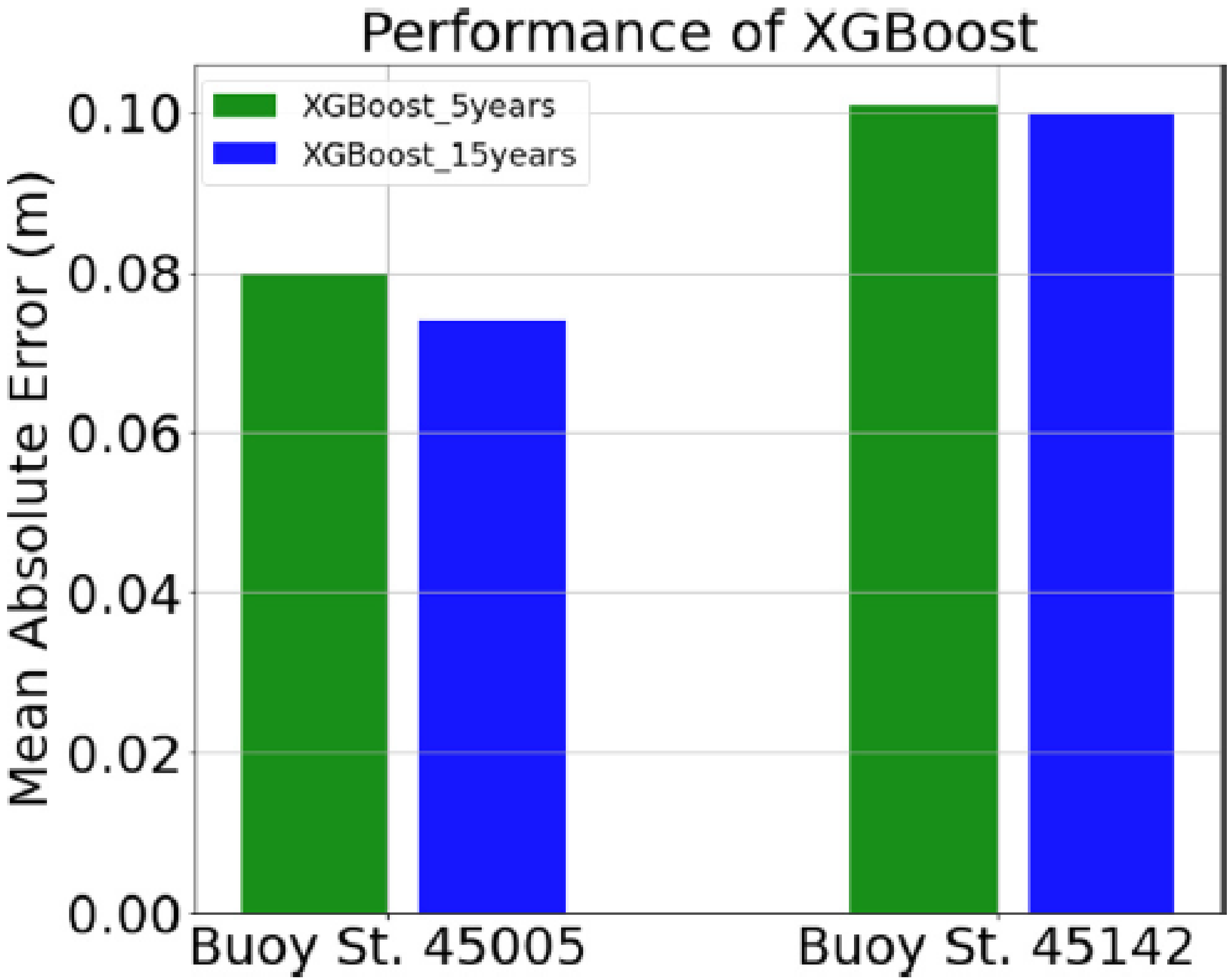
XGBoost

WW3

X axis: Observed wave height

How much data needed for XGBoost ?

Usually, the more data, the better. However, XGBoost behaves a little different



Conclusions

- XGBoost improve strong wave forecasts during storm events.
- XGBoost model perform well on both large and small sets of training data
- XGBoost need much mush less computing time than WW3

Future study

- Improve wave prediction through incorporating machine learning methods with physics-based models

References

Hu, H., A. Van der Westhuysen, P. Chu, and A. Fujisaki-Manome (2021). Predicting Lake Erie wave heights and periods using XGBoost and LSTM. Ocean Modelling, Volume 164.

.

Thank you !