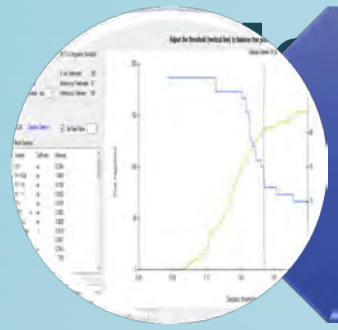
Developing A Predictive Model for Your Beach

Module 2

great lakes observing system

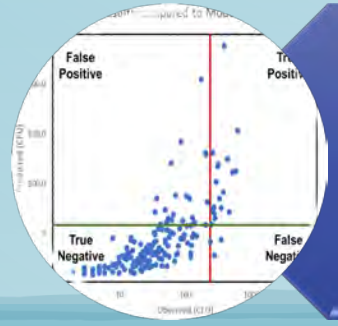
UNIVERSITY OF WISCONSIN SEA GRANT INSTITUTE

science for the sustainable use of Great Lakes resources



Which model type to use?

GBM explained



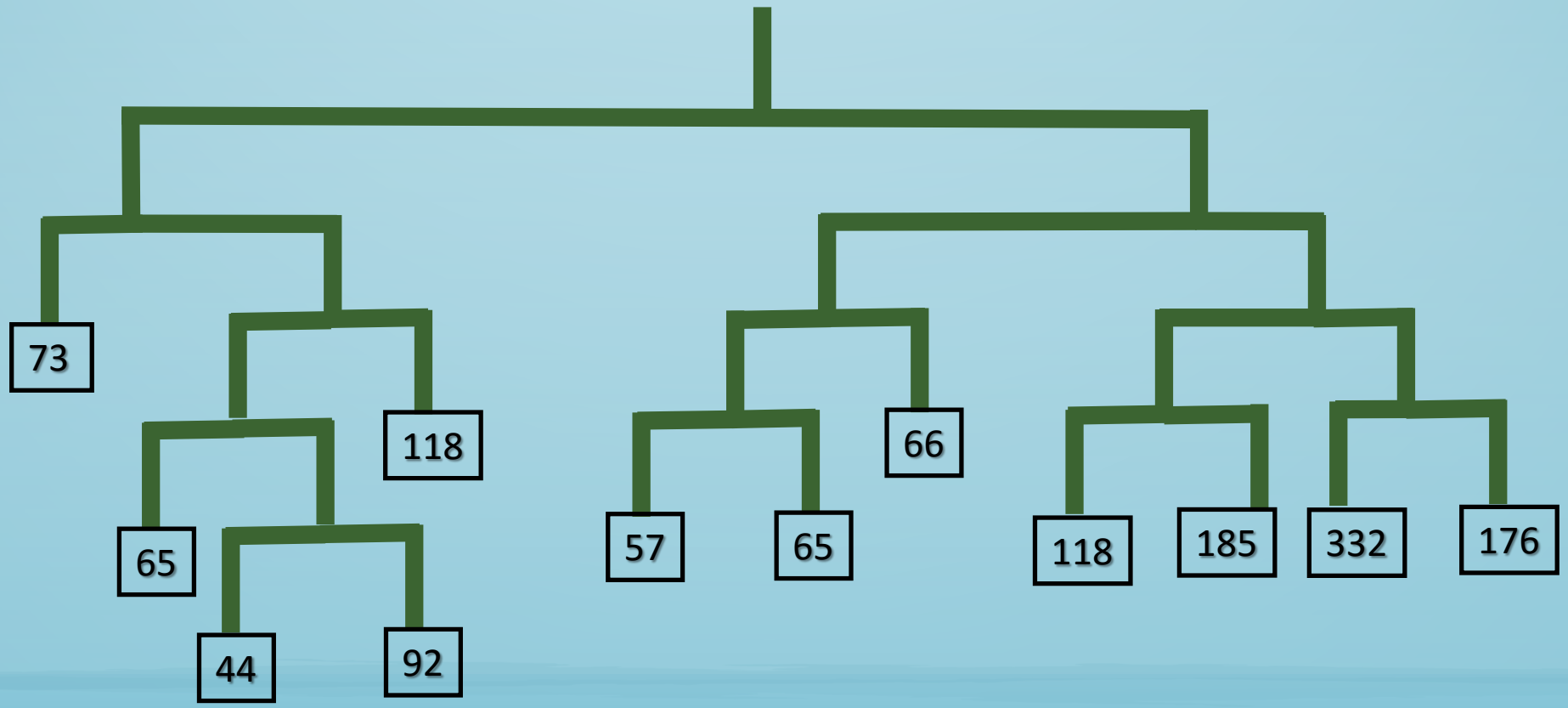
Evaluating: Was it a good model?

Virtual Beach Modeling Methods

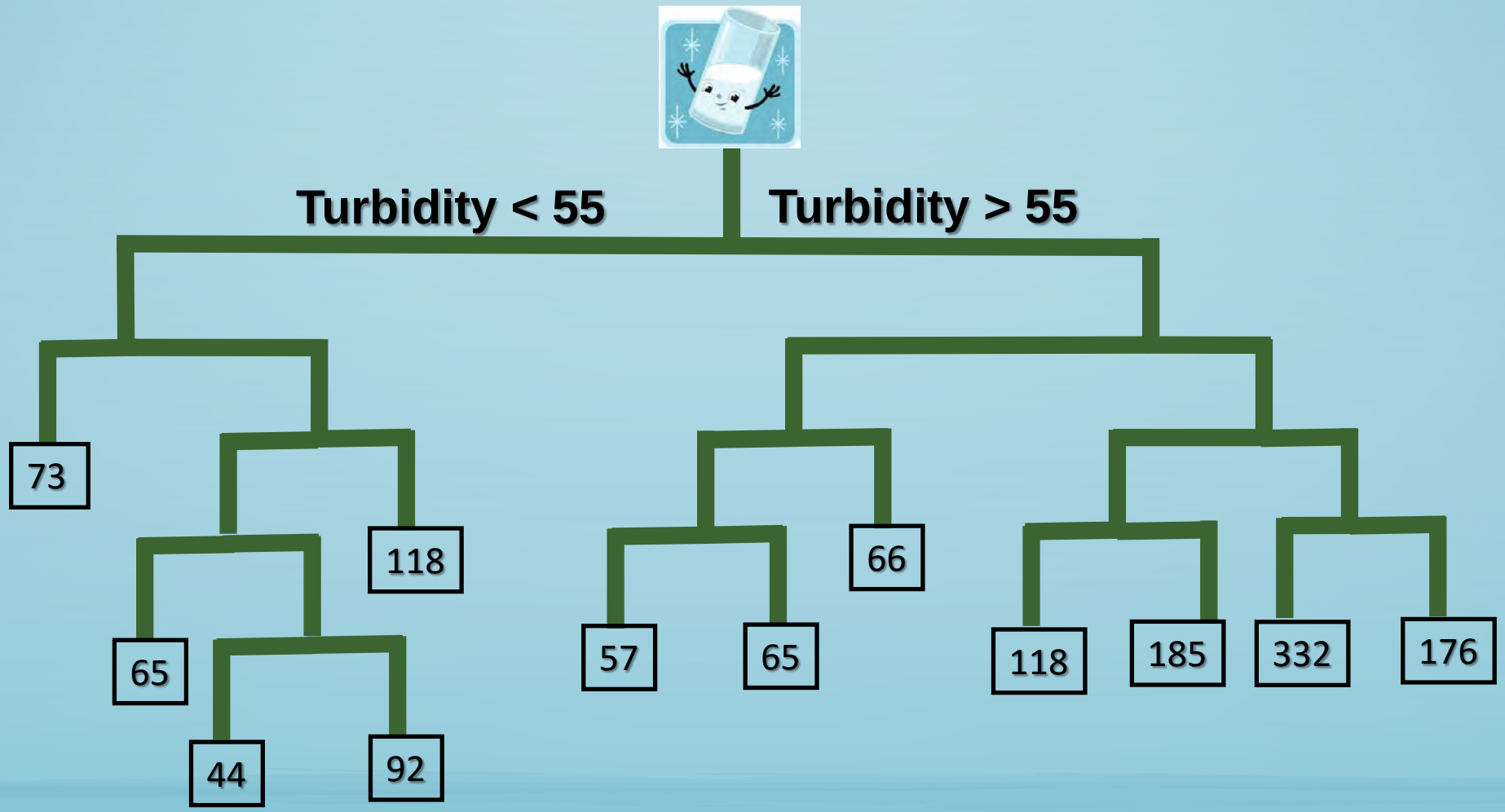
	MLR	PLS	GBM
	Multiple Linear Regression	**Partial Least Squares**	**Gradient Boosted Machines**
Model type	Linear	Linear	Non-linear
Pros	• Straightforward equation • Clear relationships between variables	• Inherently accounts for collinearity • Uses cross validation to select the best model	• Does not assume linear relationships • Uses cross validation to select the best model
Cons	• More involved model-building process. • Assumes linear relationship between variables. • Prone to over-fitting.	• Assumes linear relationship between variables.	• “Black-boxy”

Regression Trees

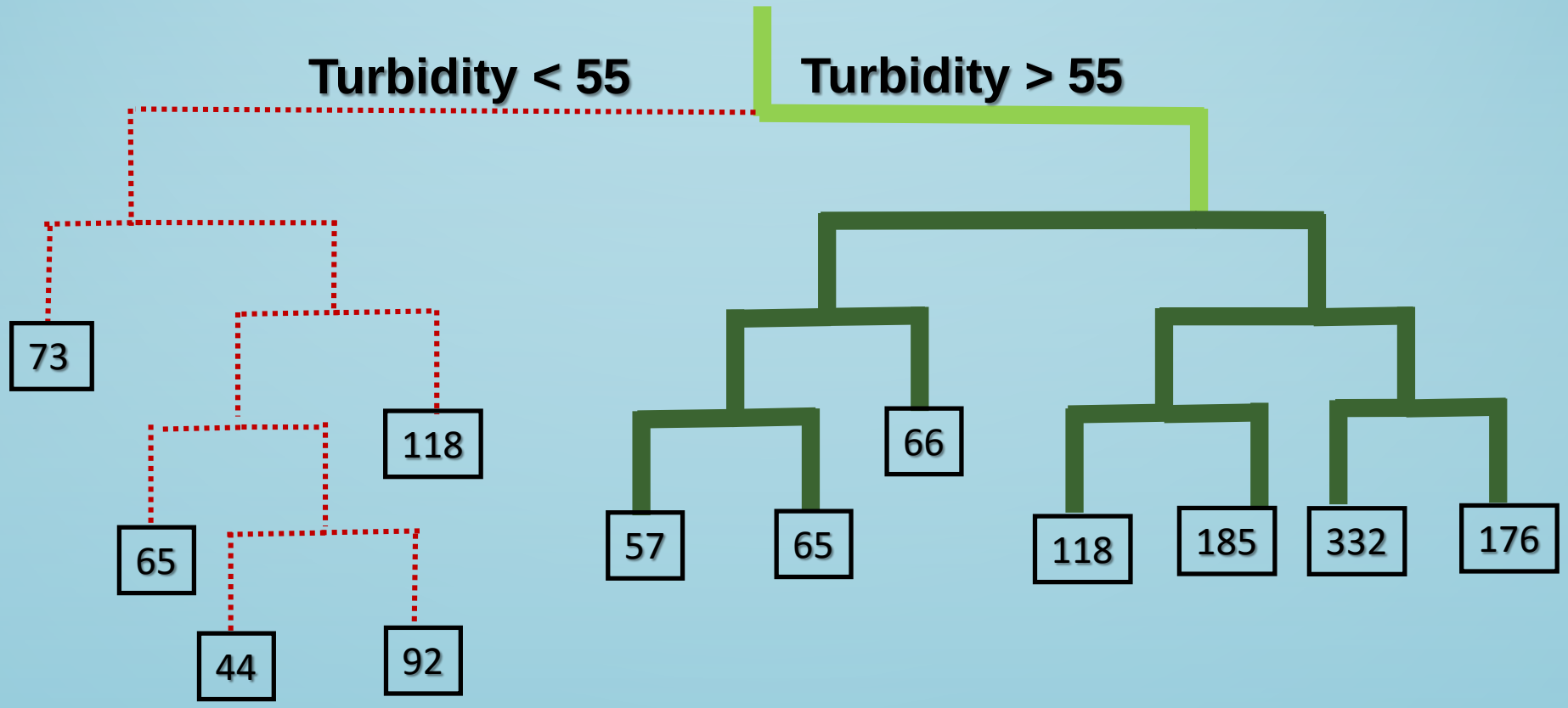
Modeling Continuous Data



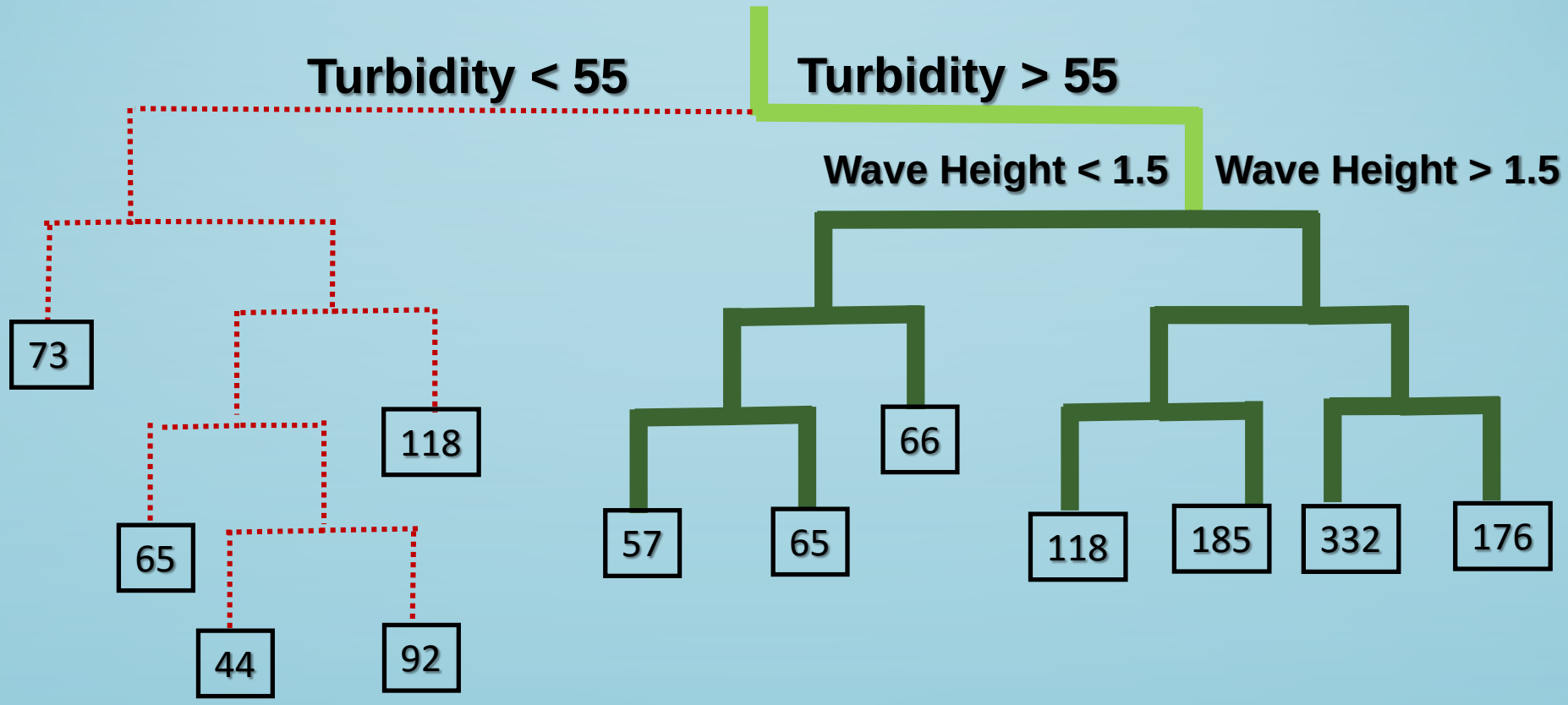
Regression Trees



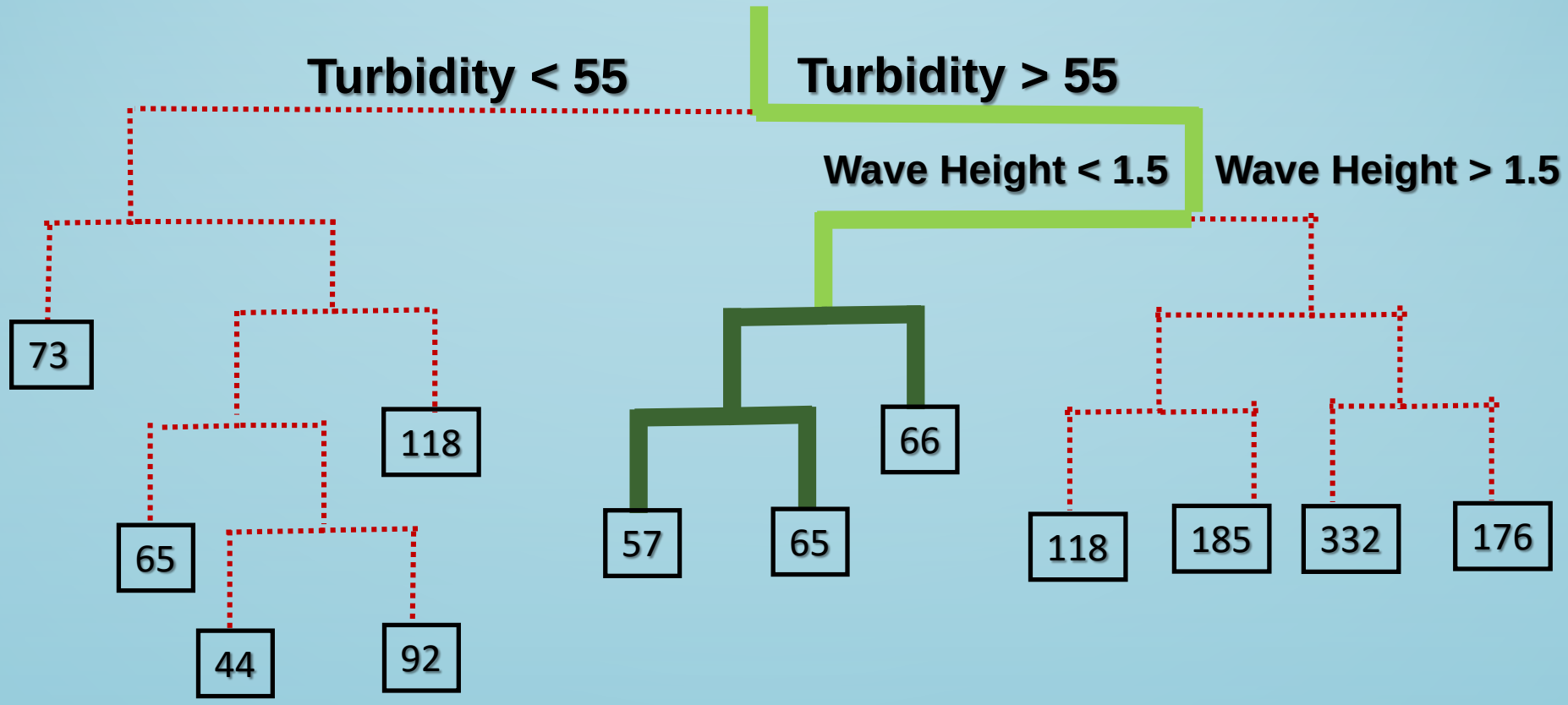
Regression Trees



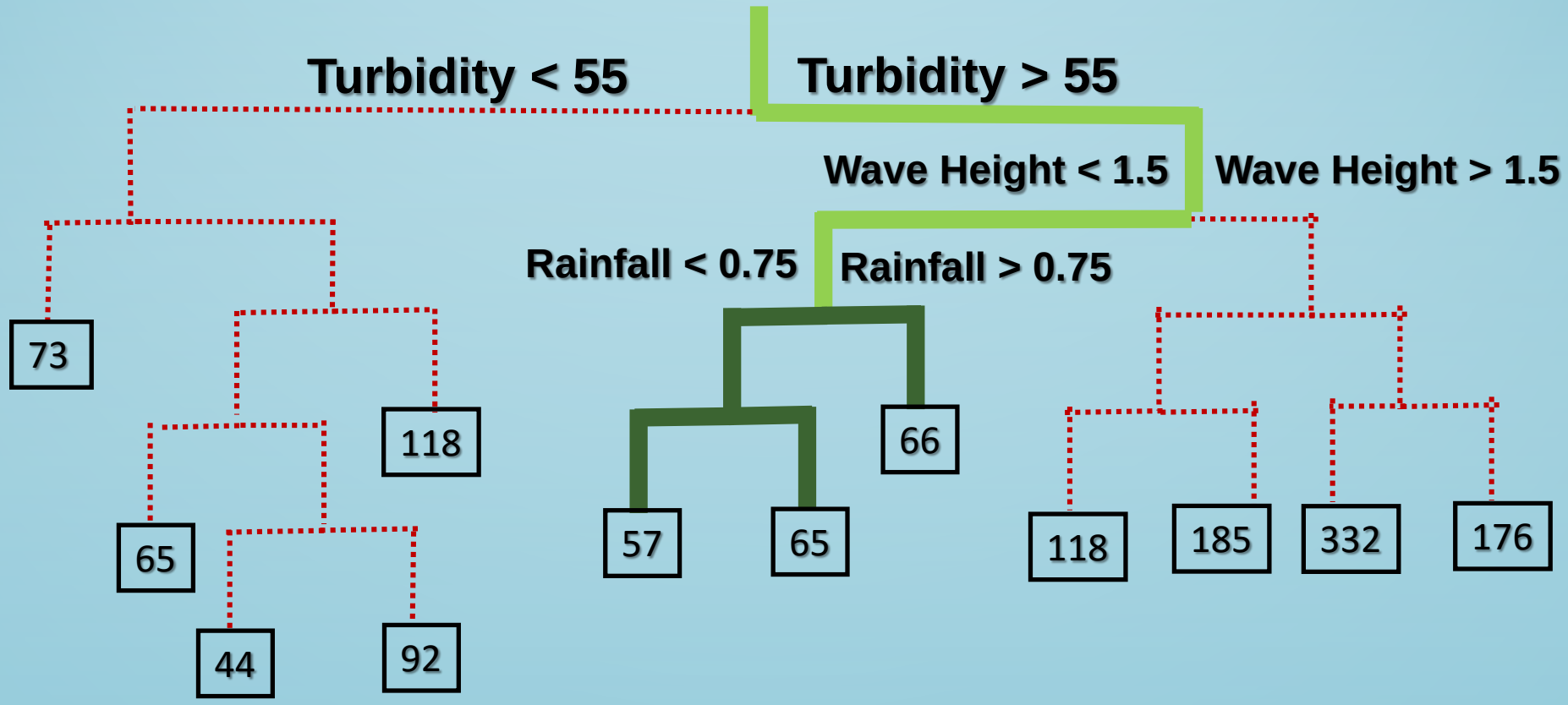
Regression Trees



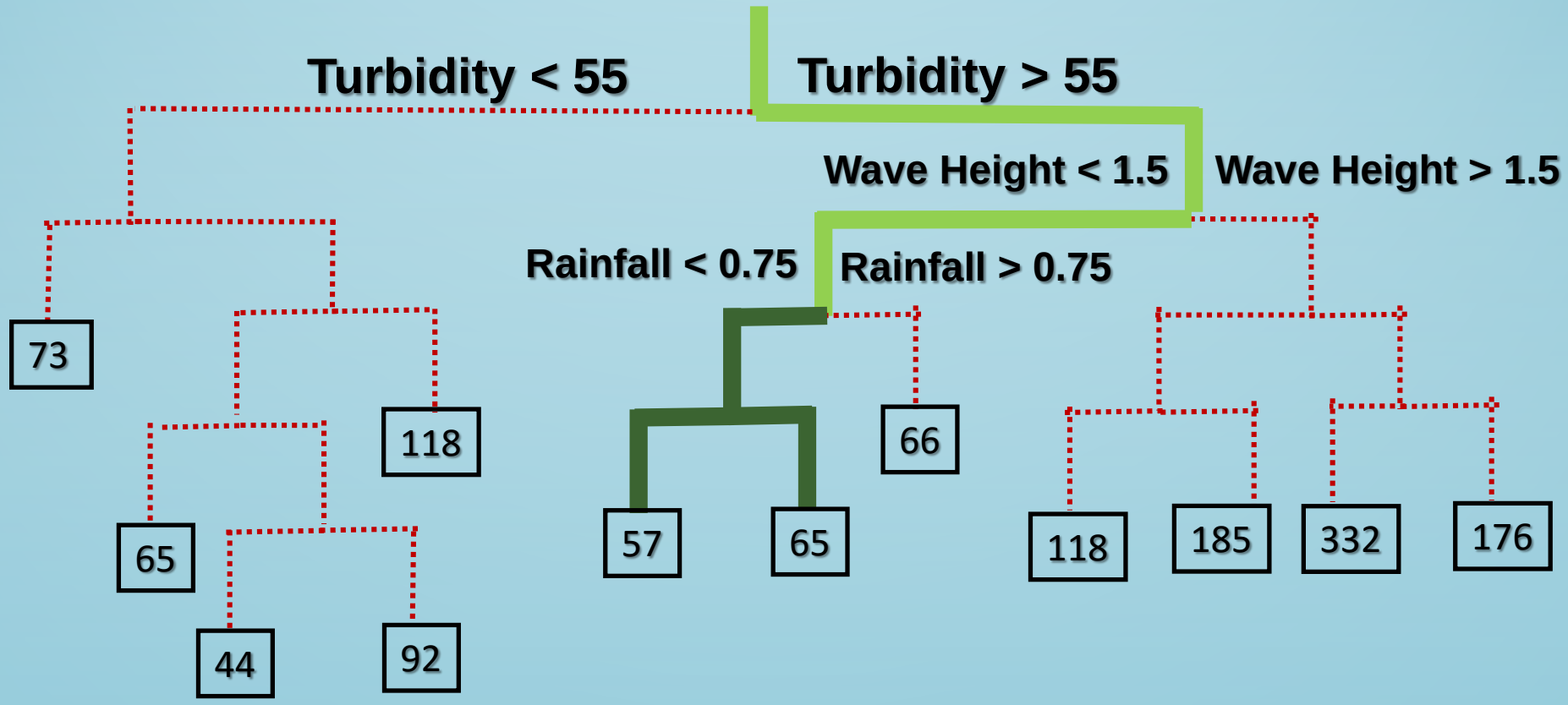
Regression Trees



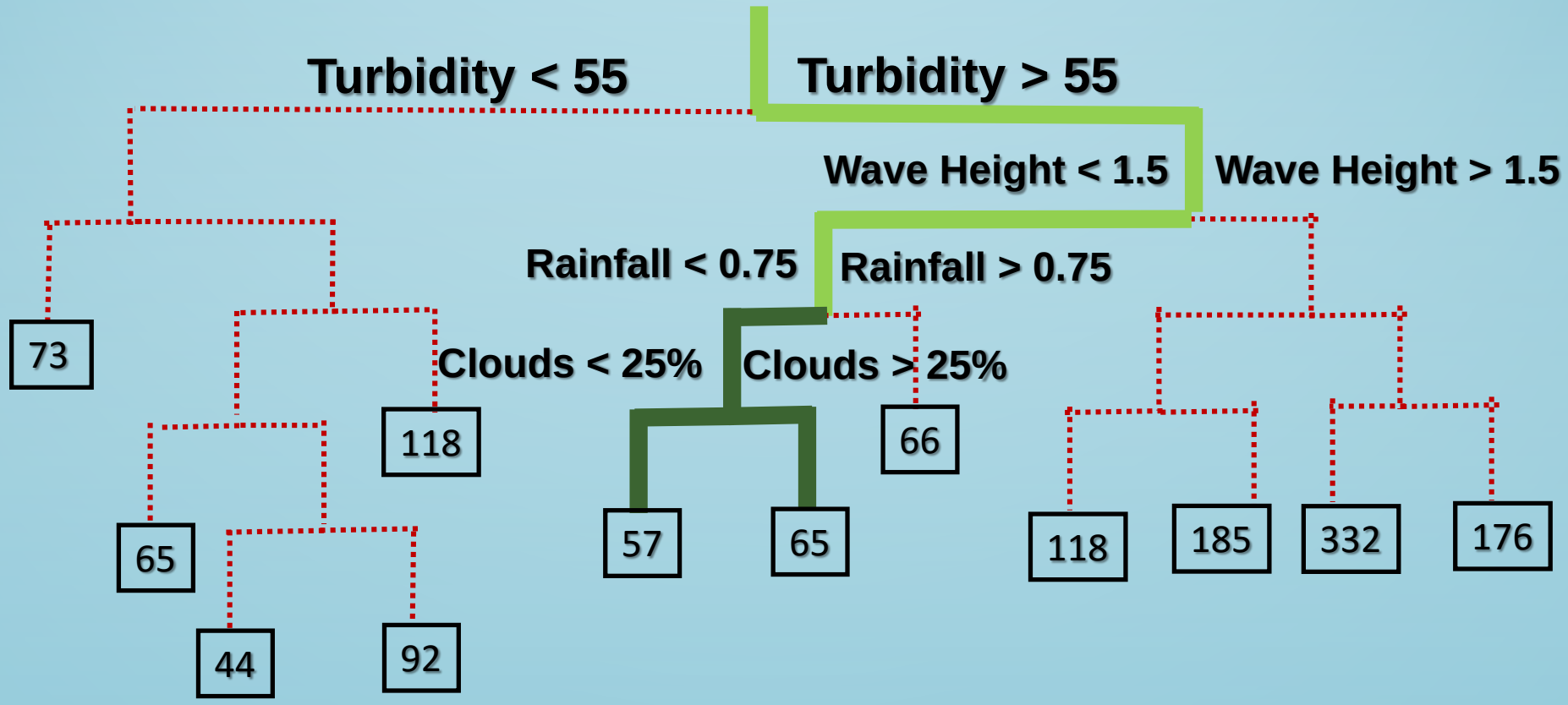
Regression Trees



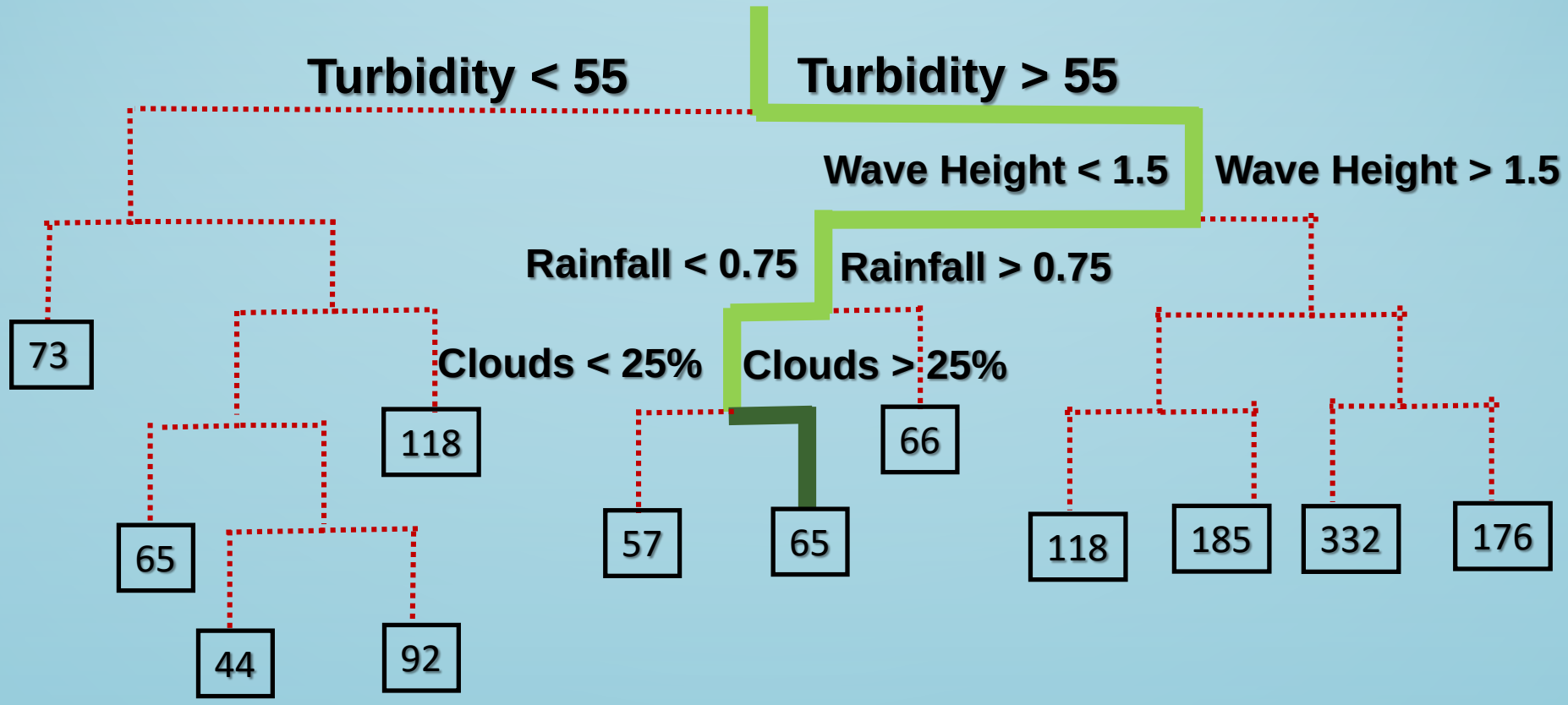
Regression Trees



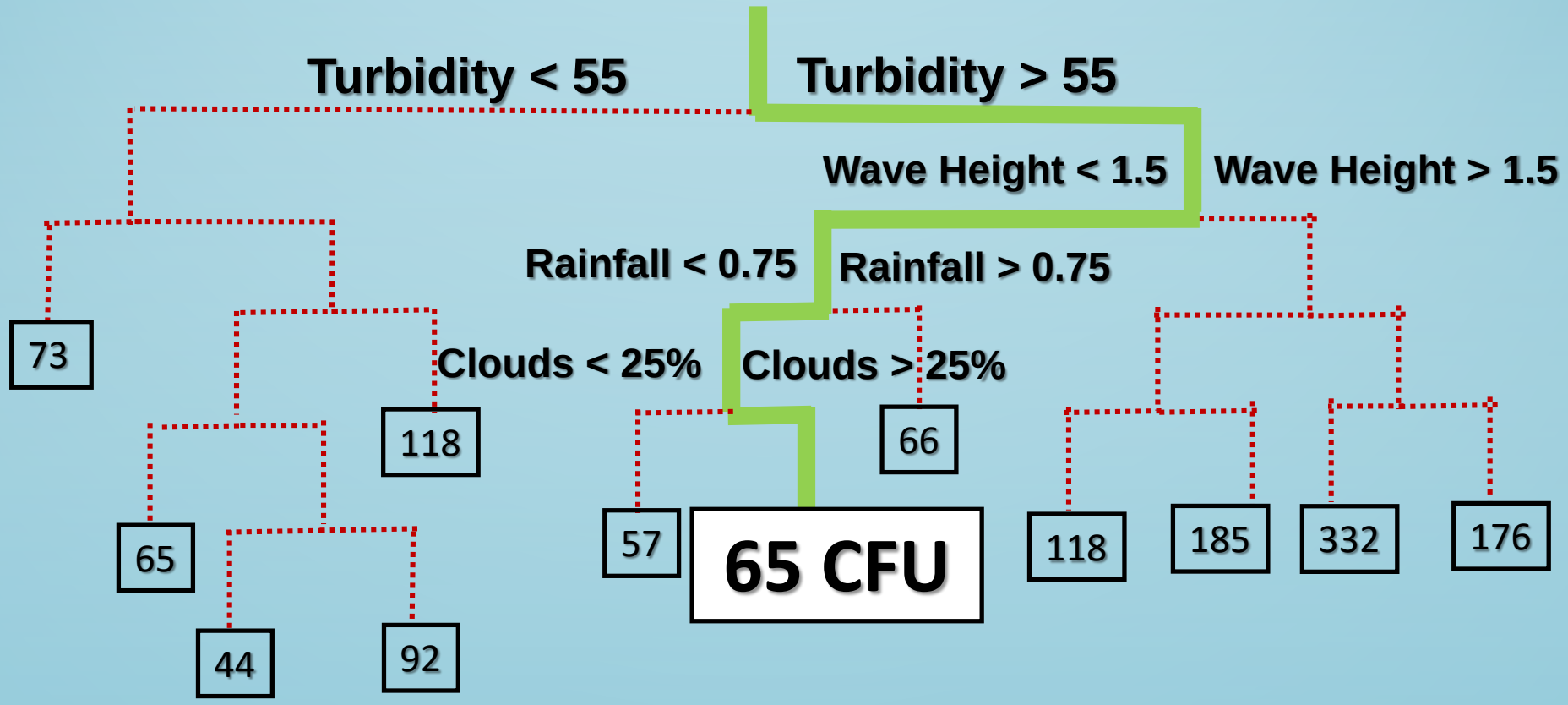
Regression Trees



Regression Trees



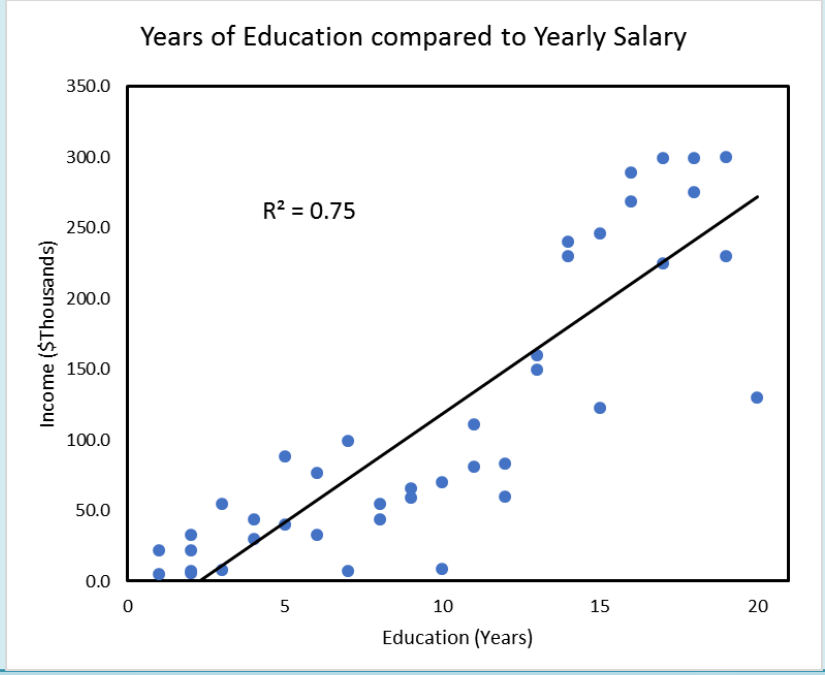
Regression Trees



Evaluating Models

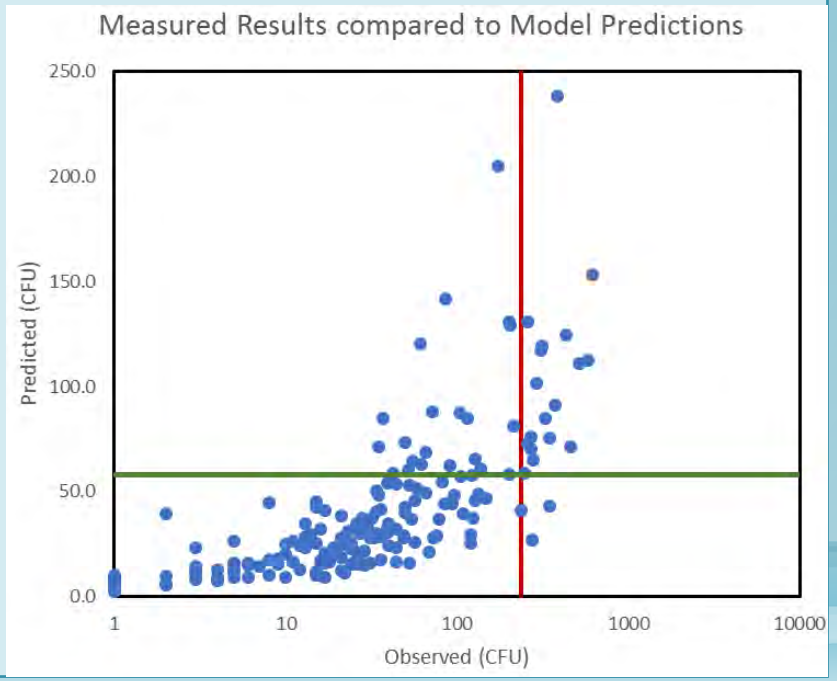
Goodness of Fit

- How well response variation is explained by variation in independent variables
- R^2 …



Predictive Power

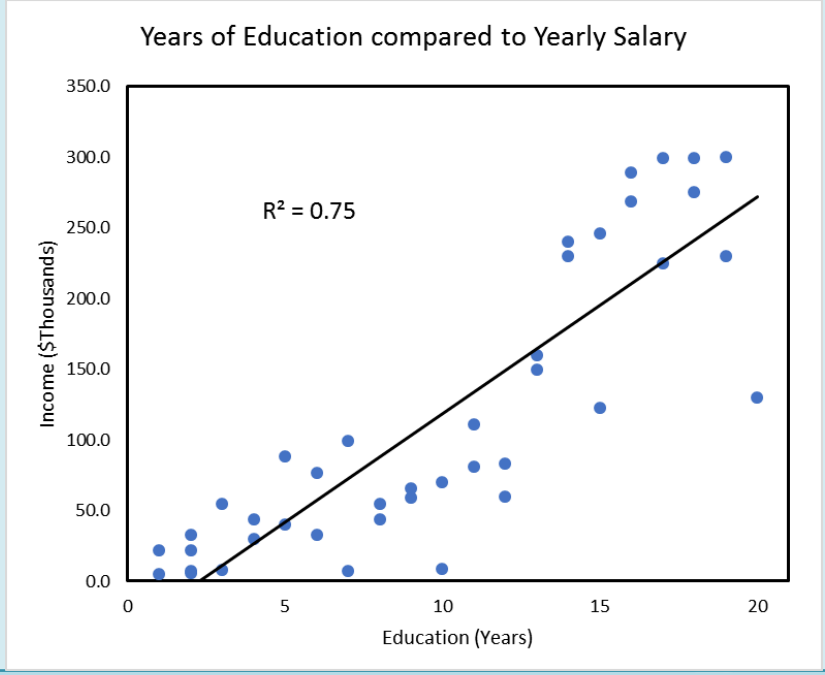
- ability to make predictions for data the model has not seen before
- Sensitivity, Specificity…



Evaluating Models

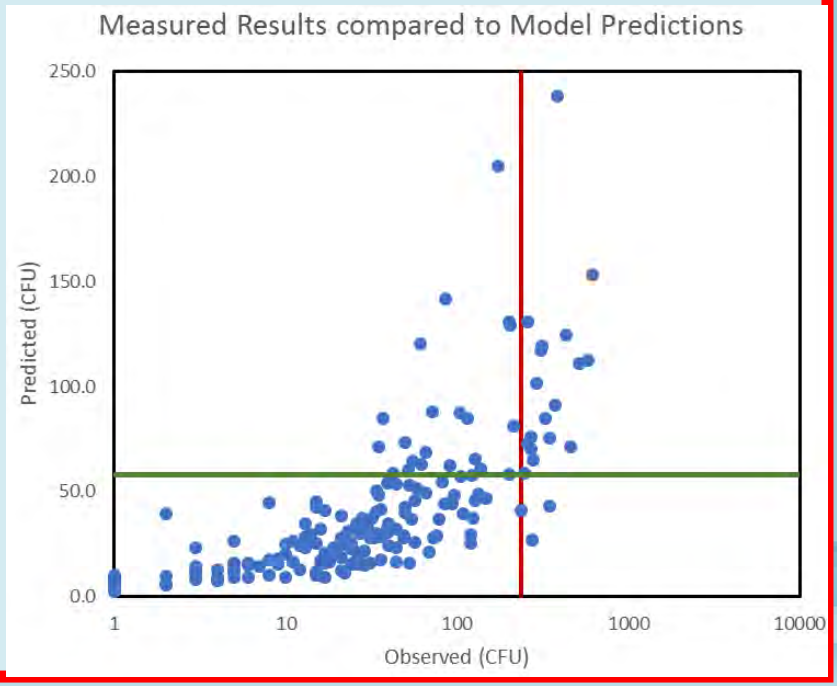
Goodness of Fit

- How well response variation is explained by variation in independent variables
- R^2 …

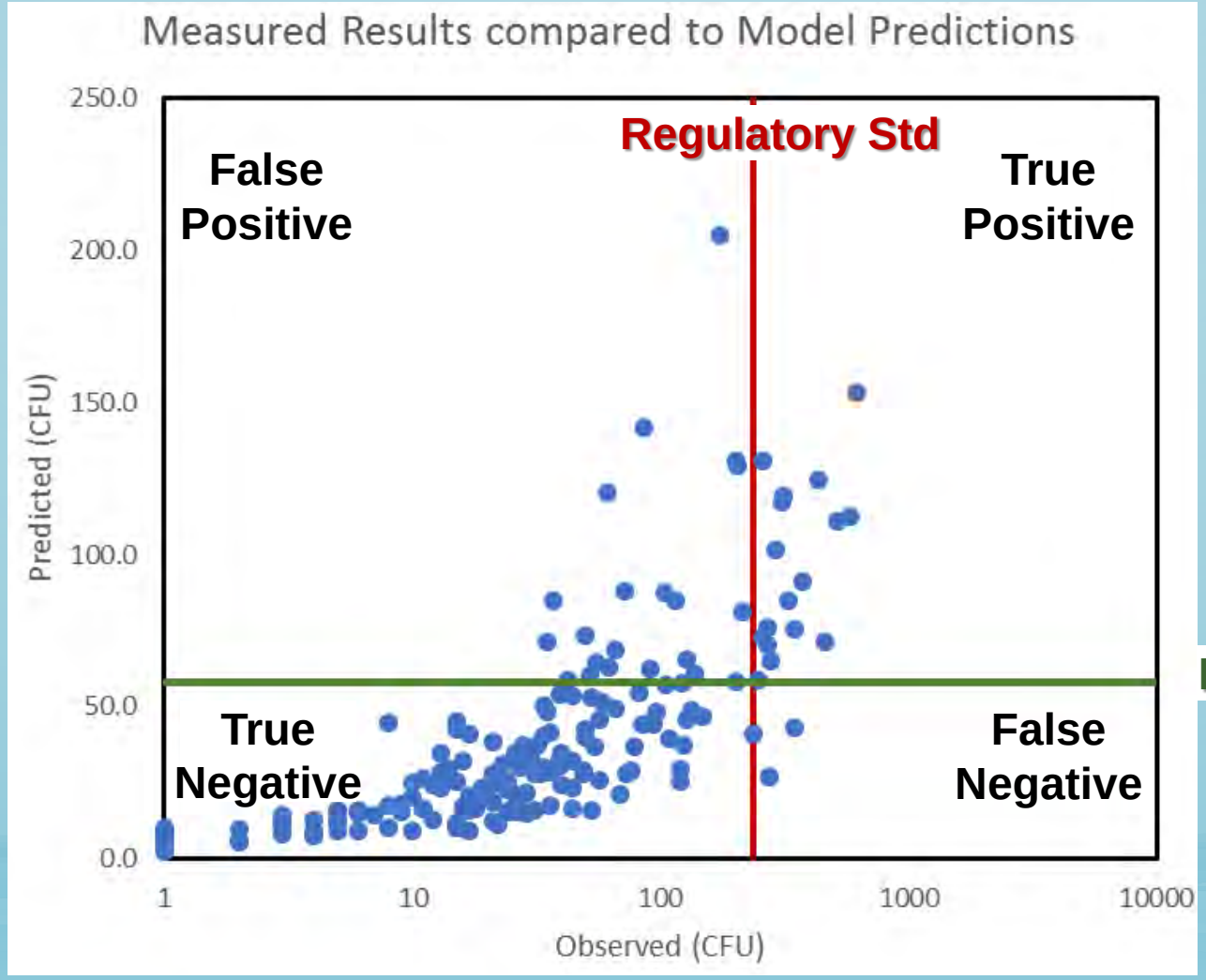


Predictive Power

- ability to make predictions for data the model has not seen before
- Sensitivity, Specificity…



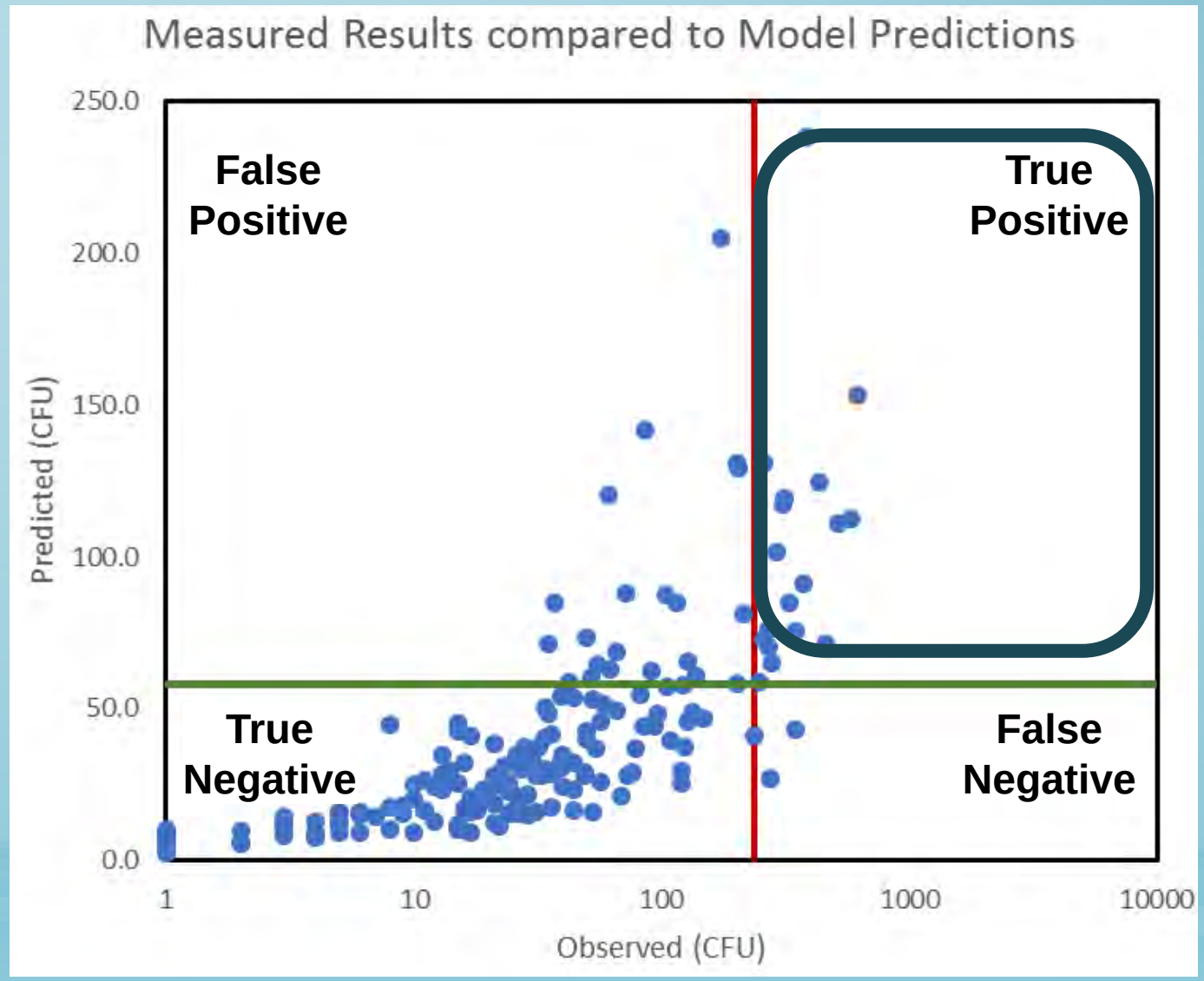
Evaluate: How well will models predict exceedances over standard?



Sensitivity

% of all results that were observed to be over standard correctly predicted by the model

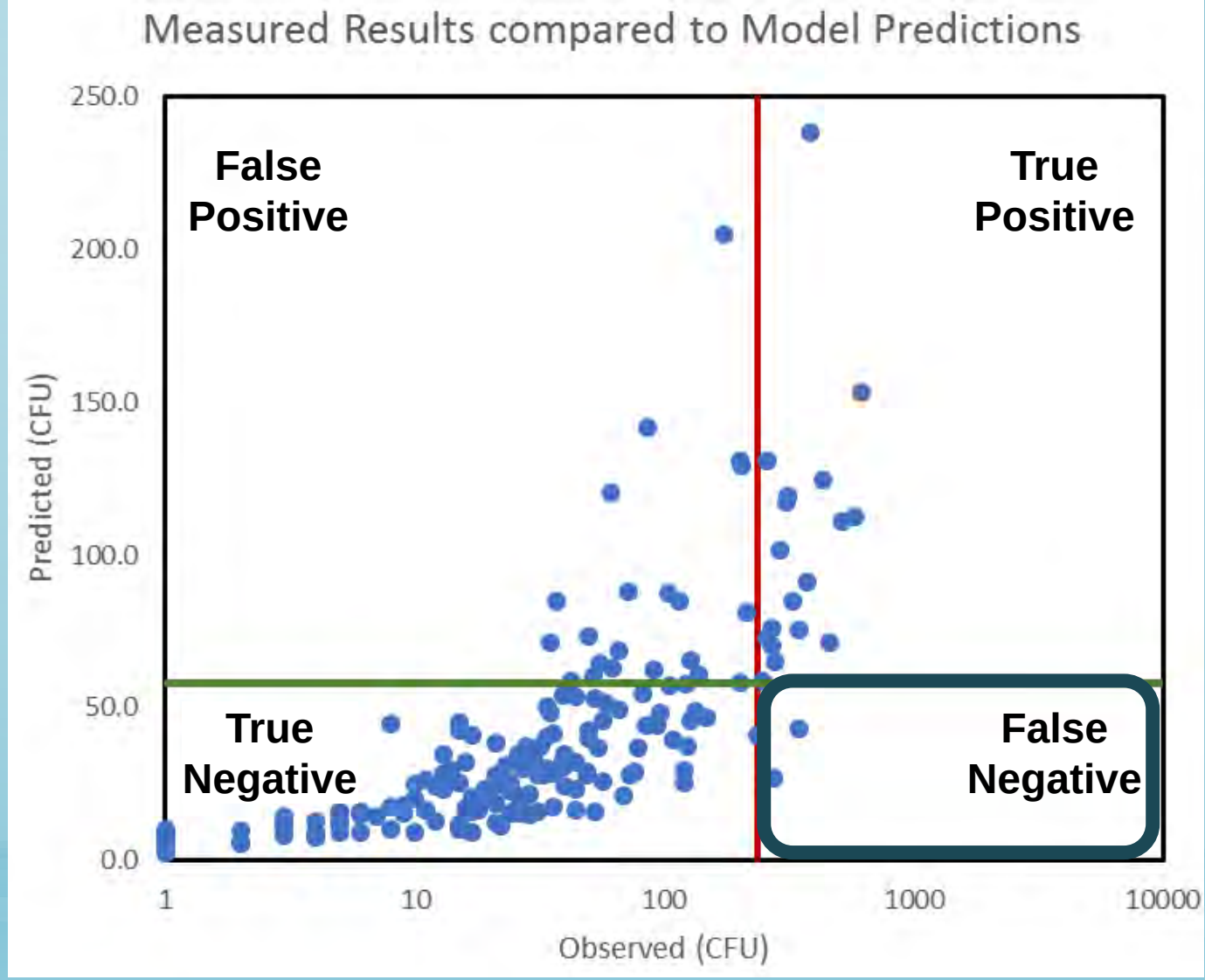
$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$



Sensitivity

% of all results that were observed to be over standard correctly predicted by the model

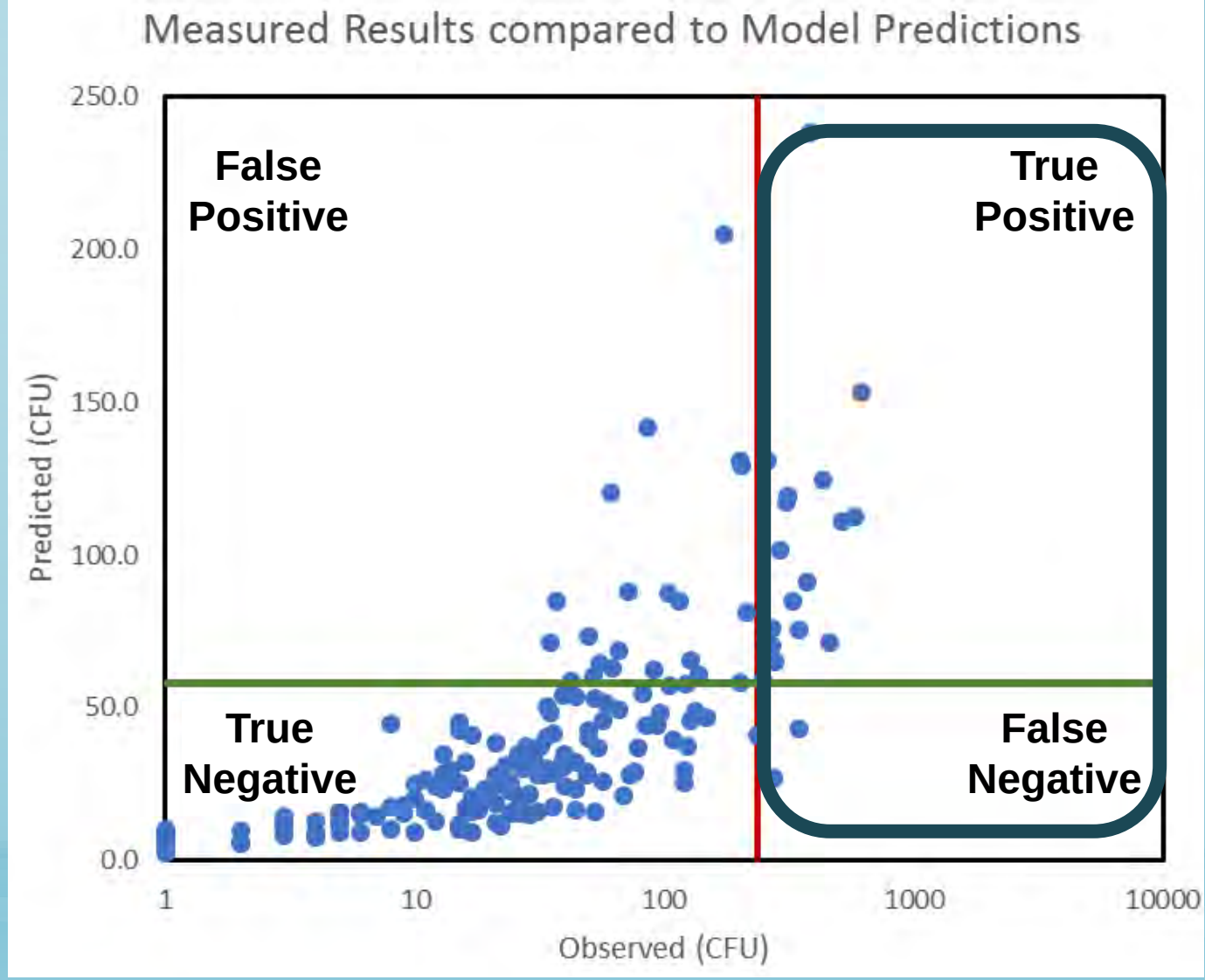
$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$



Sensitivity

% of all results that were observed to be over standard correctly predicted by the model

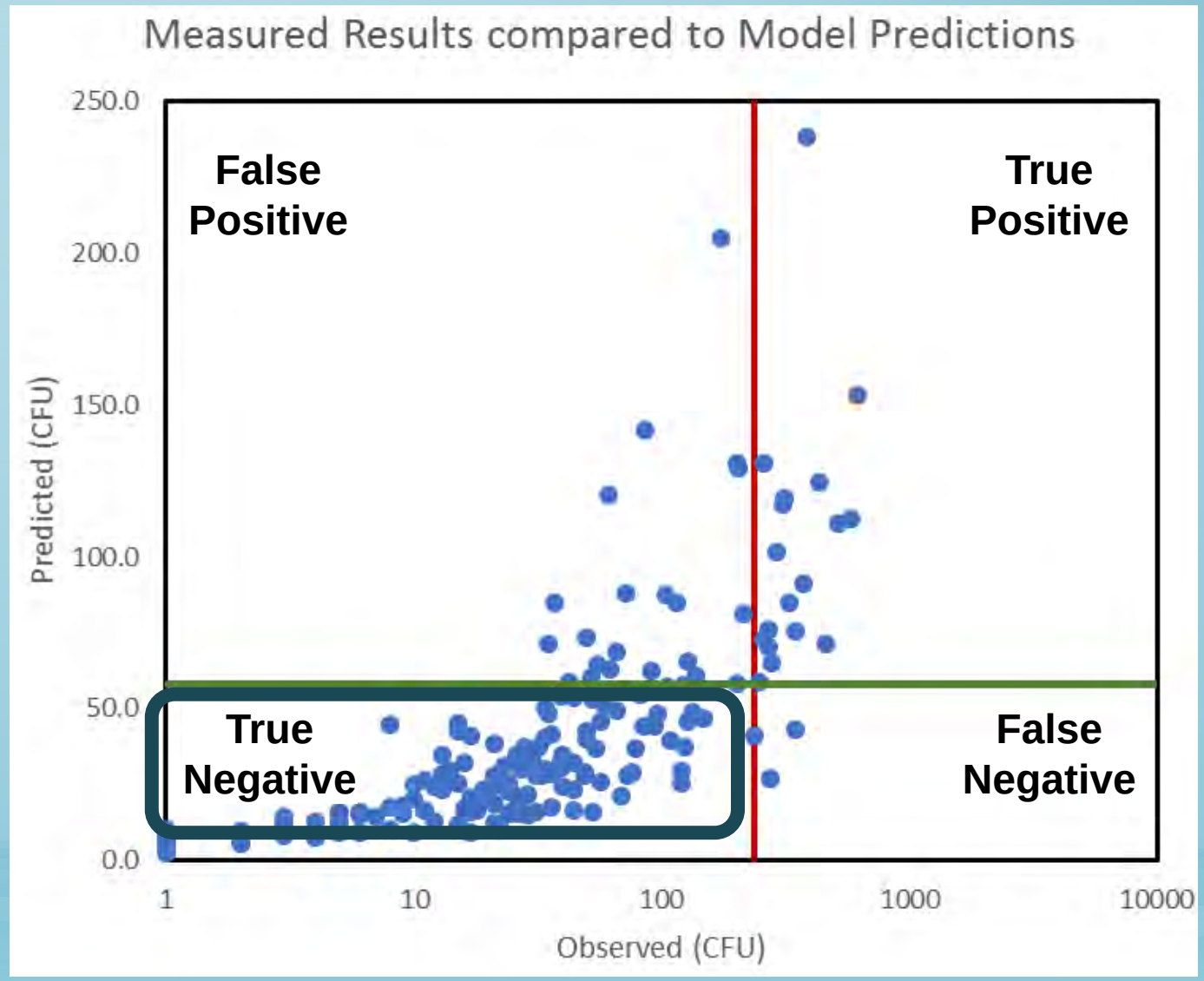
$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$



Specificity

% of all results that were observed to be under standard correctly predicted by the model

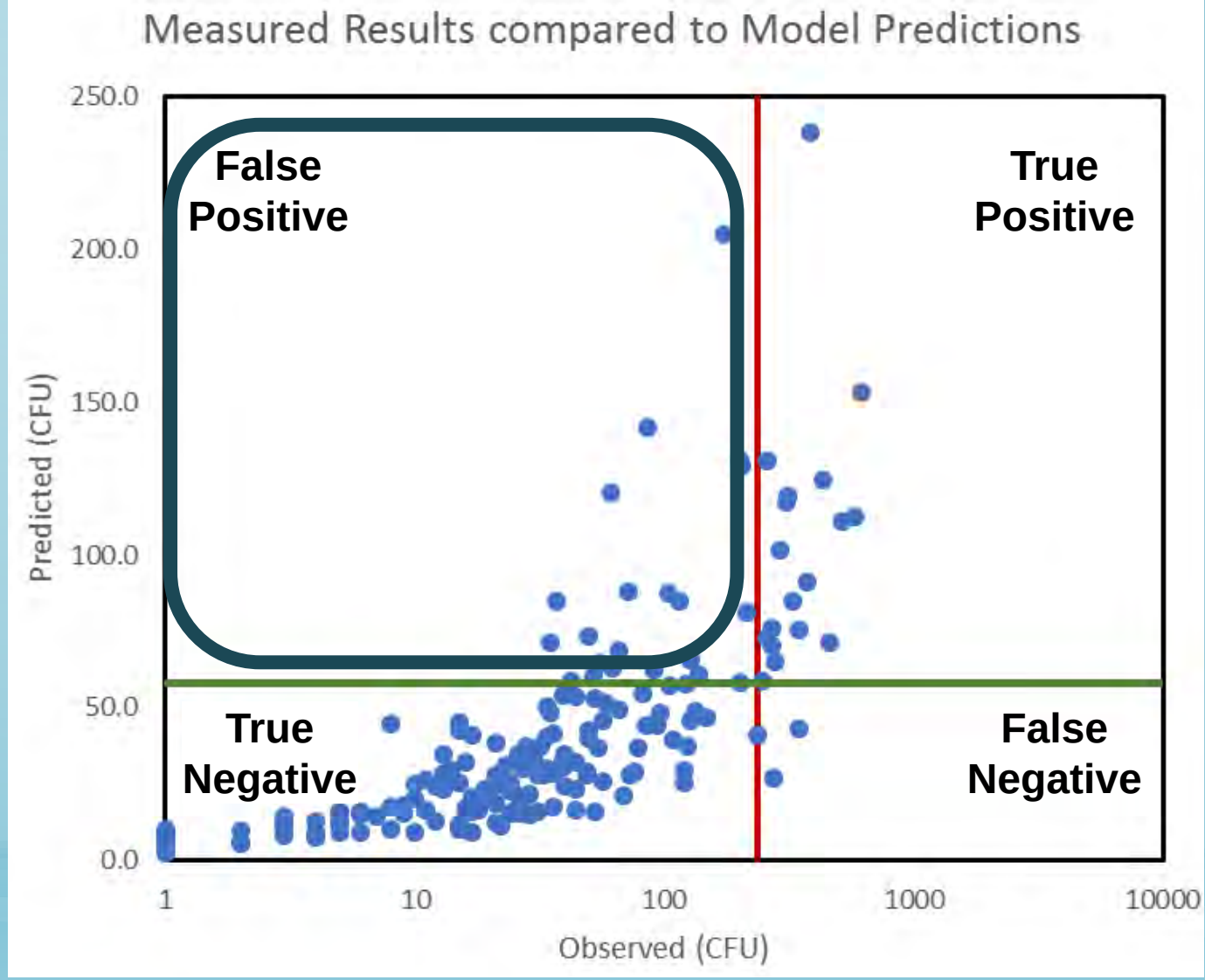
$$\frac{\text{True Negatives}}{\text{True Negaitives + False Positives}}$$



Specificity

% of all results that were observed to be under standard correctly predicted by the model

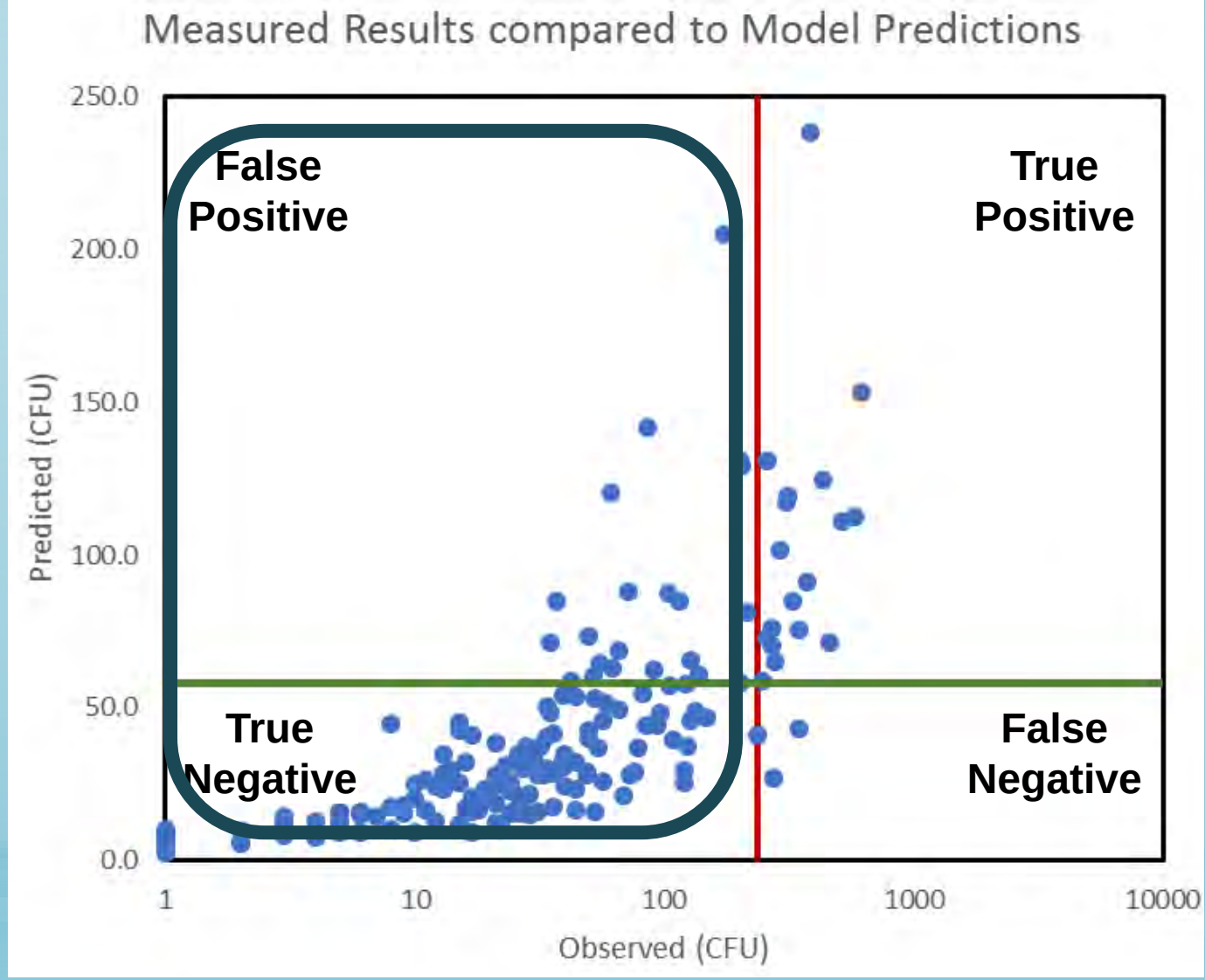
$$\frac{\text{True Negatives}}{\text{True Negaitives + False Positives}}$$



Specificity

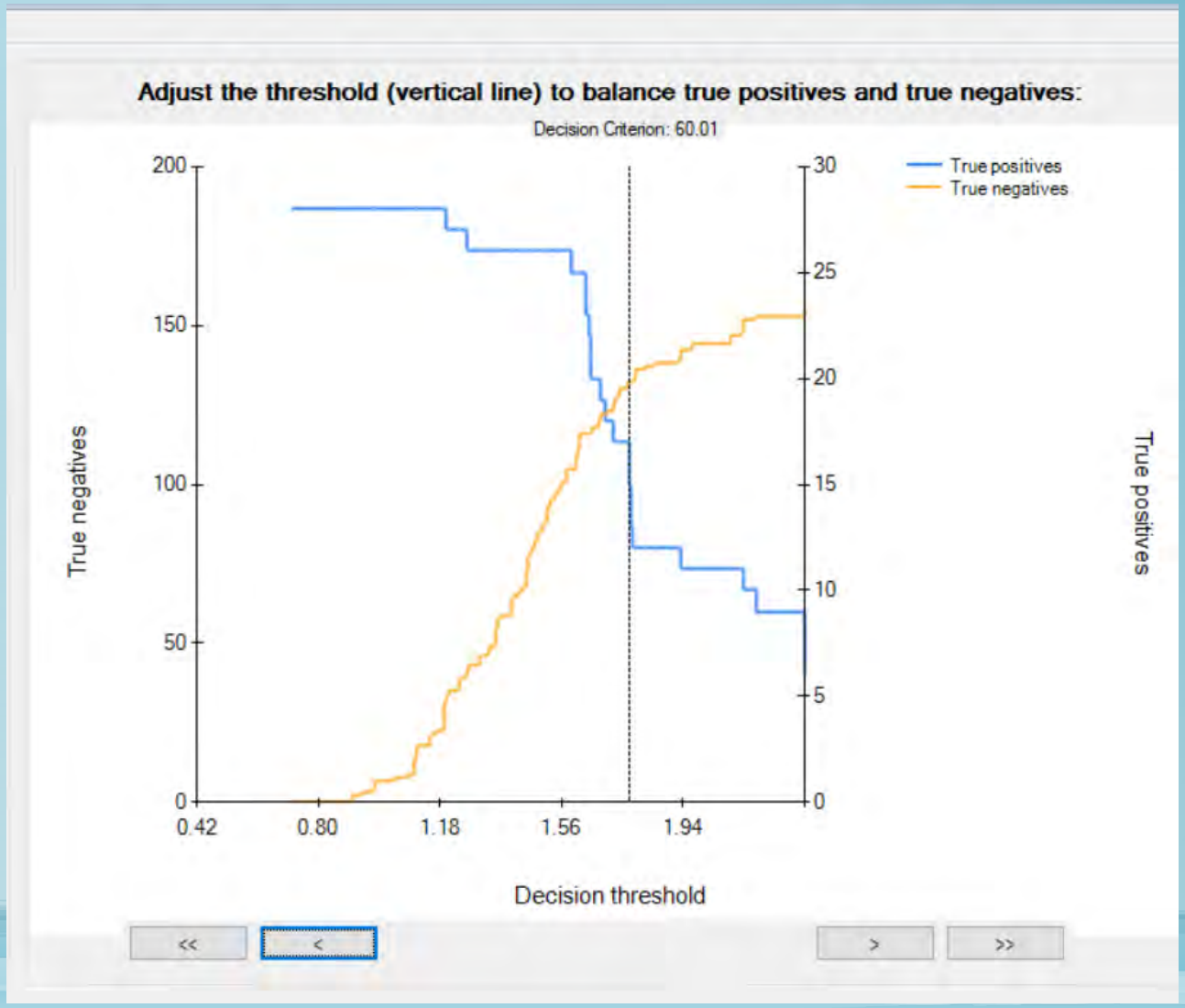
% of all results that were observed to be under standard correctly predicted by the model

$$\frac{\text{True Negatives}}{\text{True Negaitives + False Positives}}$$

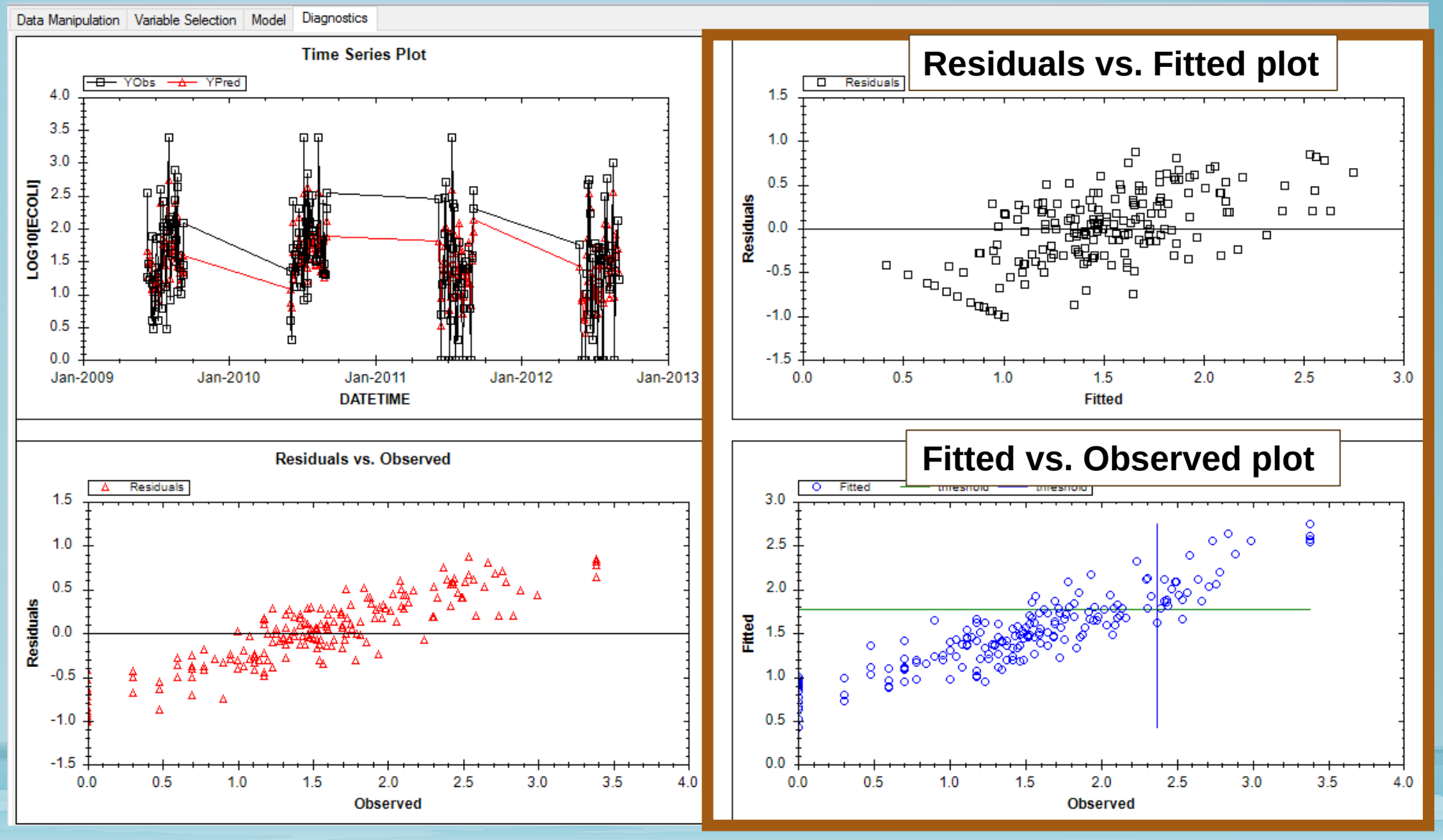


Decision Criterion

- The prediction thresholds that determine whether an actual exceedance of a regulatory standard occurred
- Model specific
- “Converts model predictions to real-world values”



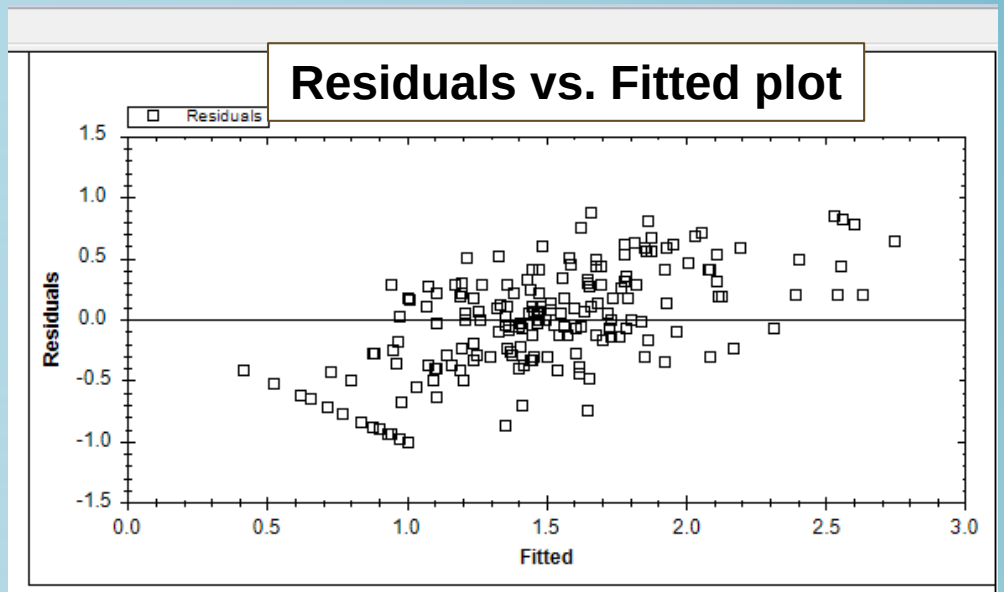
Diagnostics



Diagnostics

Mental Check:

Data points with residuals greater +/- 3.0 considered to be too influential



Model values of *E. Coli* v. observed values (Log_{10}). The closer to the 1:1 line, the better