Chesapeake Bay HydroML: Advanced Streamflow And Water Quality Predictions For The Chesapeake Bay Watershed

Kim Van Meter, Chaopeng Shen, Xueting Pu, Elham Mahmod Por
Pennsylvania State University

Machine-Learning Solutions

Problems with Current Approaches

- Older versions of CBP watershed models were not designed to work at fine spatial scales
- Process-based models run at fine scale are slow and computationally expensive

Opportunities with Machine-Learning Approaches

- Data Integration - Can harness different data types, large amounts of input data
- Can handle fine-scale geospatial data
- Computational efficiency – less computationally intensive
- Provide data-driven insights that capture complex, nonlinear relationships, reducing need for detailed process understanding

Introduction

Overall Objective

Explore and develop machine-learning approaches for eventual integration into the CBP modeling framework

Kim Van Meter
Department of Geography
Penn State University

Xueting Pu
Department of Civil & Environmental Engineering
Penn State University

Chaopeng Shen
Department of Civil & Environmental Engineering
Penn State University

Elham Mahmod Por
Department of Civil & Environmental Engineering
Penn State University

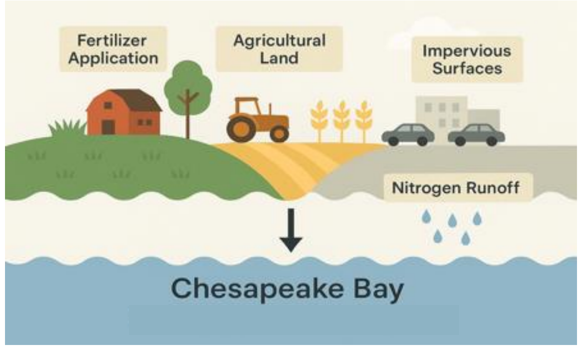
TEAM

Current Nutrient-Modeling Goals

- Develop Random Forest Models to Predict Nutrient Concentrations and Loads at a Monthly Time Scale

- Use the Random Forest/Machine-Learning Frameworks to Inform Development and Refinement of Land-to-Water Factors

- Assess the Effectiveness of Newer High-Resolution Land-Use and Geomorphological Data for Prediction



What is a Random Forest Model???

A decision tree = a flowchart of yes/no questions that ends in a prediction.

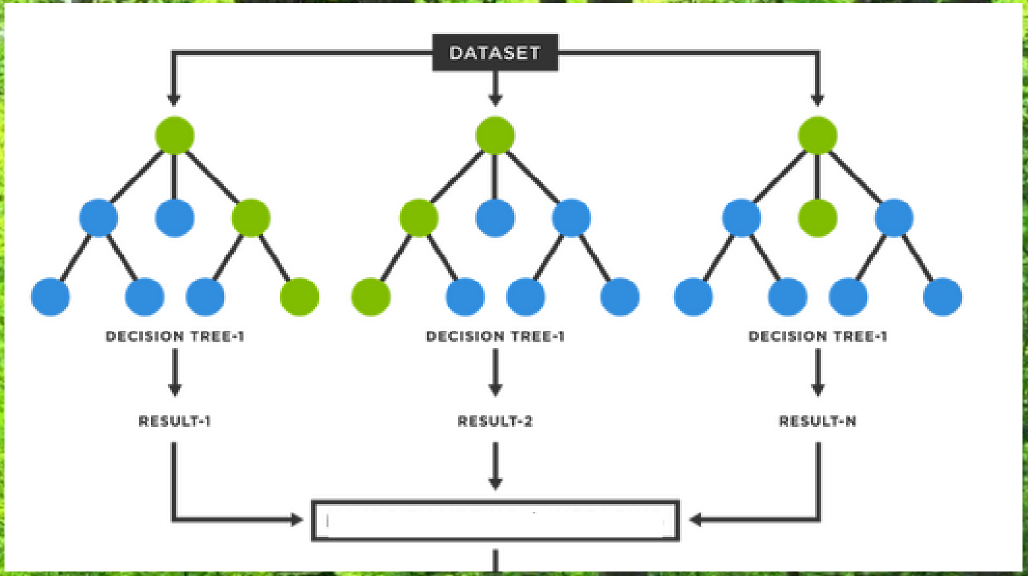
Potential Problem: one tree can “overfit” and make poor predictions on new data.

Random forest = many decision trees, each trained on a random slice of the data and features.

Each tree makes a prediction; the forest combines them (vote or average).

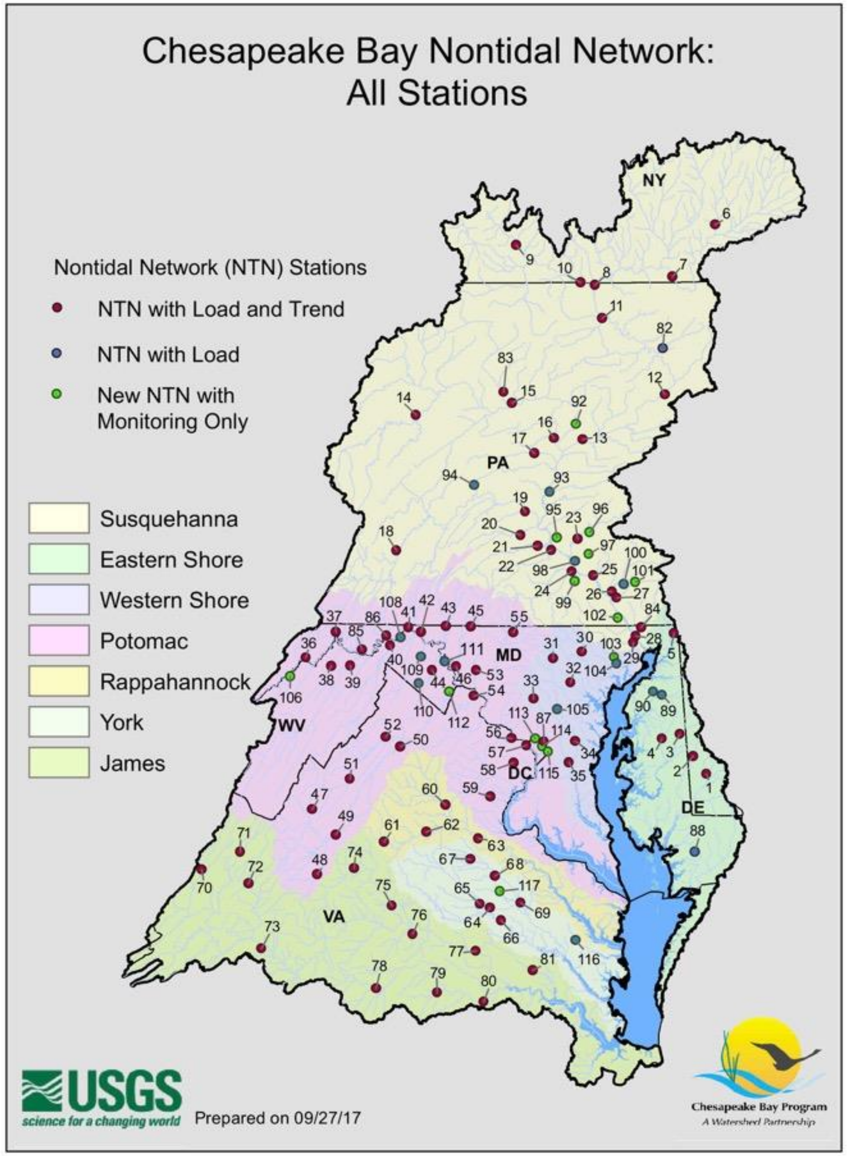
Result: more accurate, stable, and less likely to overfit.

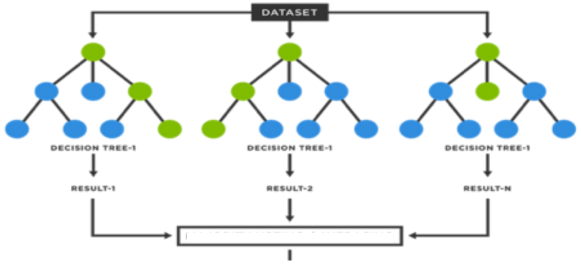
Bonus: can show which features matter most for the predictions.



Chesapeake HydroML

Random Forest Modeling





Total Nitrogen
Total Phosphorus
Phosphate (dissolved P)

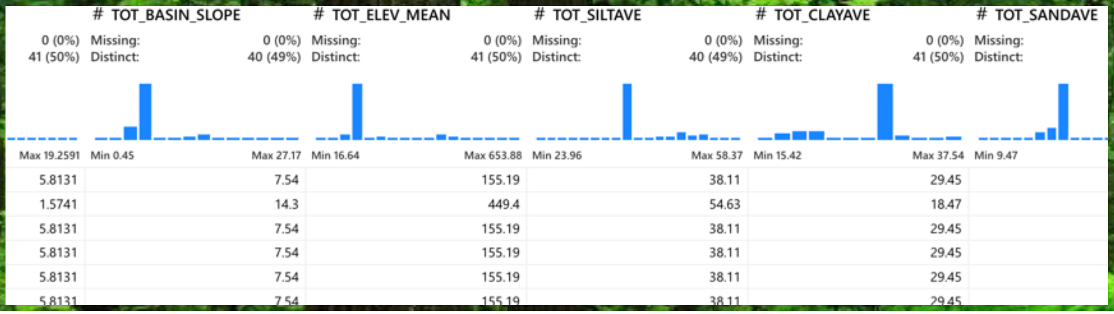
Model Structure

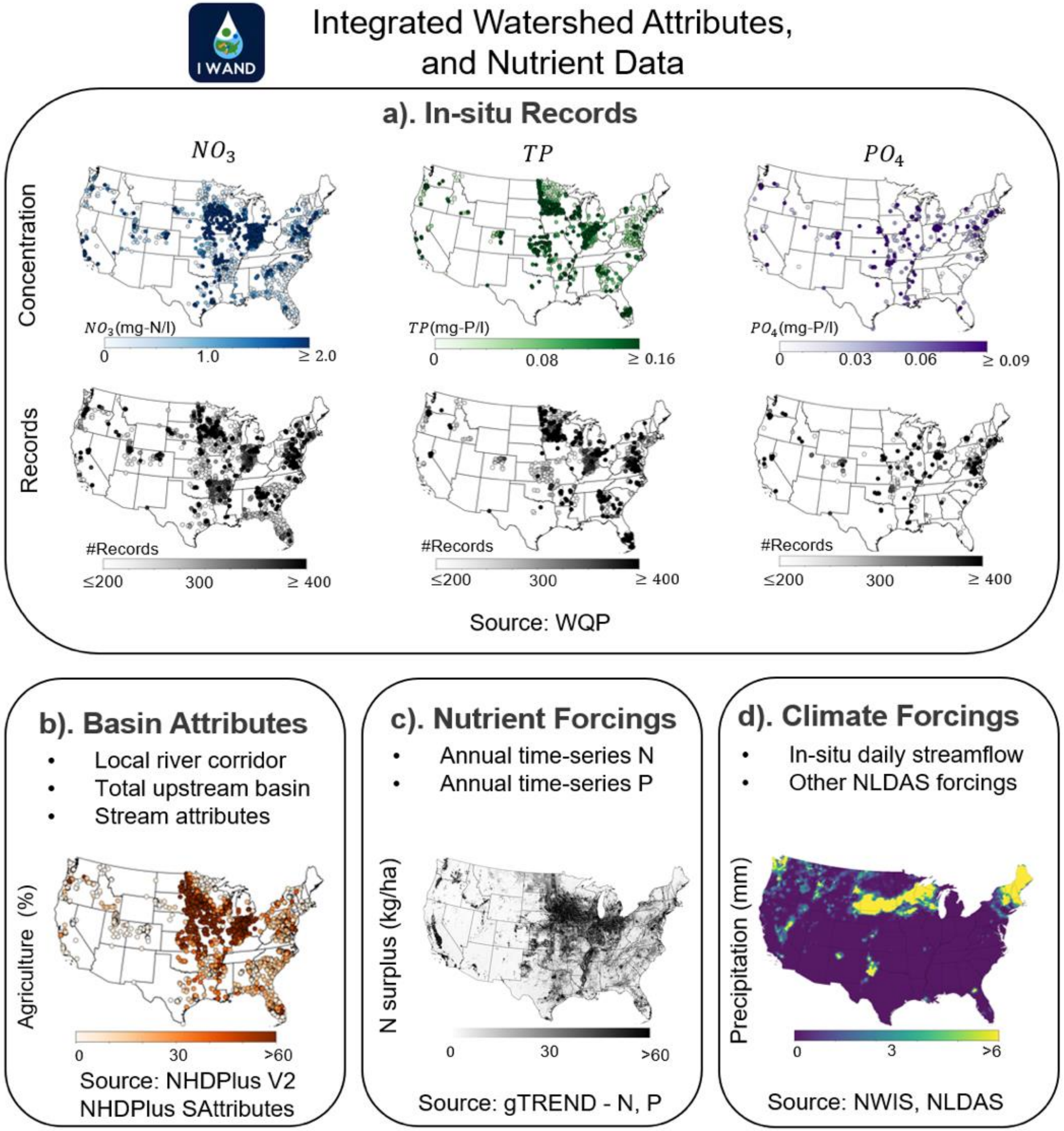
Watershed/Stream Characteristics

- Land Use/Land Cover
- Soil Type
- Geology
- Geomorphology
- Watershed Area
- Stream Order
- Etc.

Forcings

- Streamflow (monthly)
- Precipitation (monthly)
- Temperature (monthly)
- Other climate forcings (monthly)
- Nutrient Inputs (annual), from gTREND dataset

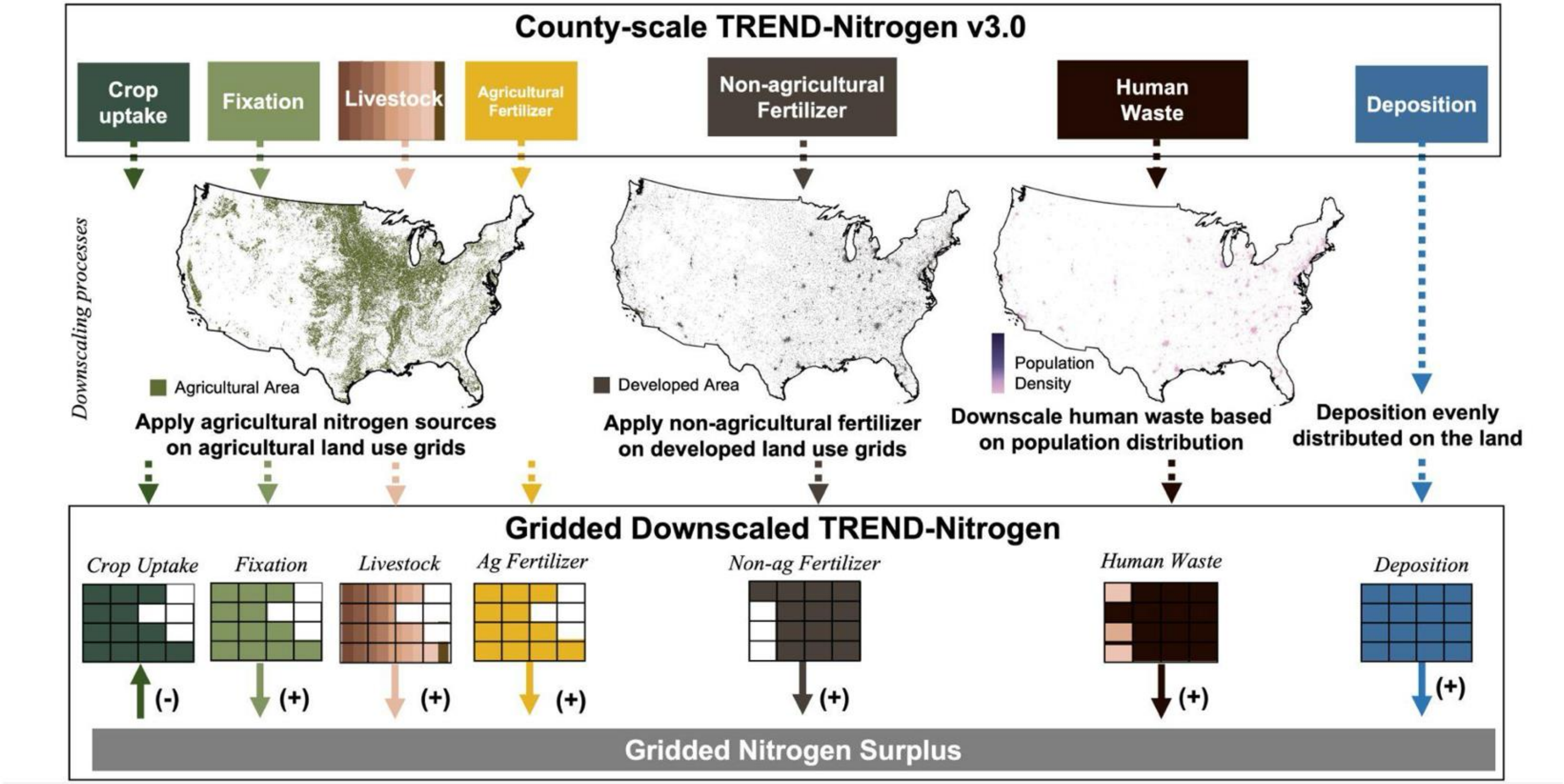




IWAND Dataset

Chang et al., submitted to *Nature Scientific Data*

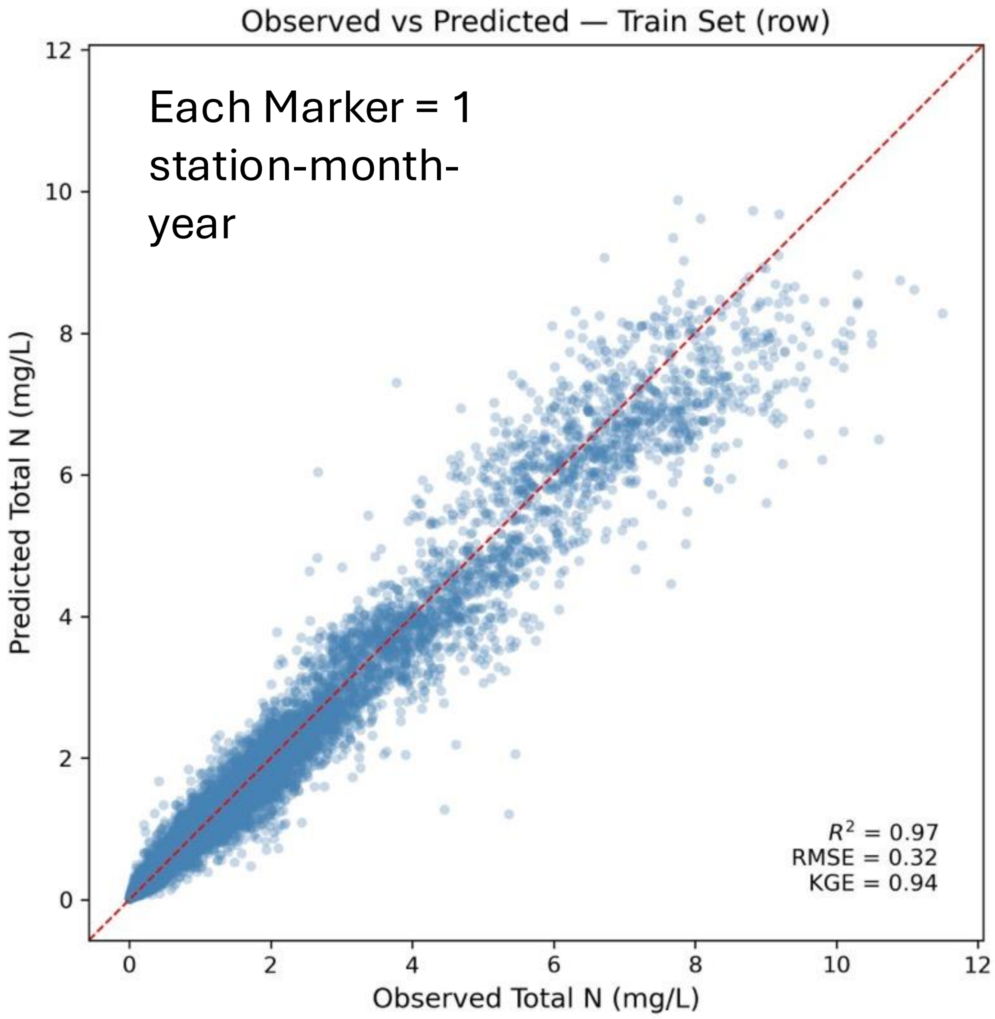
gTREND Dataset

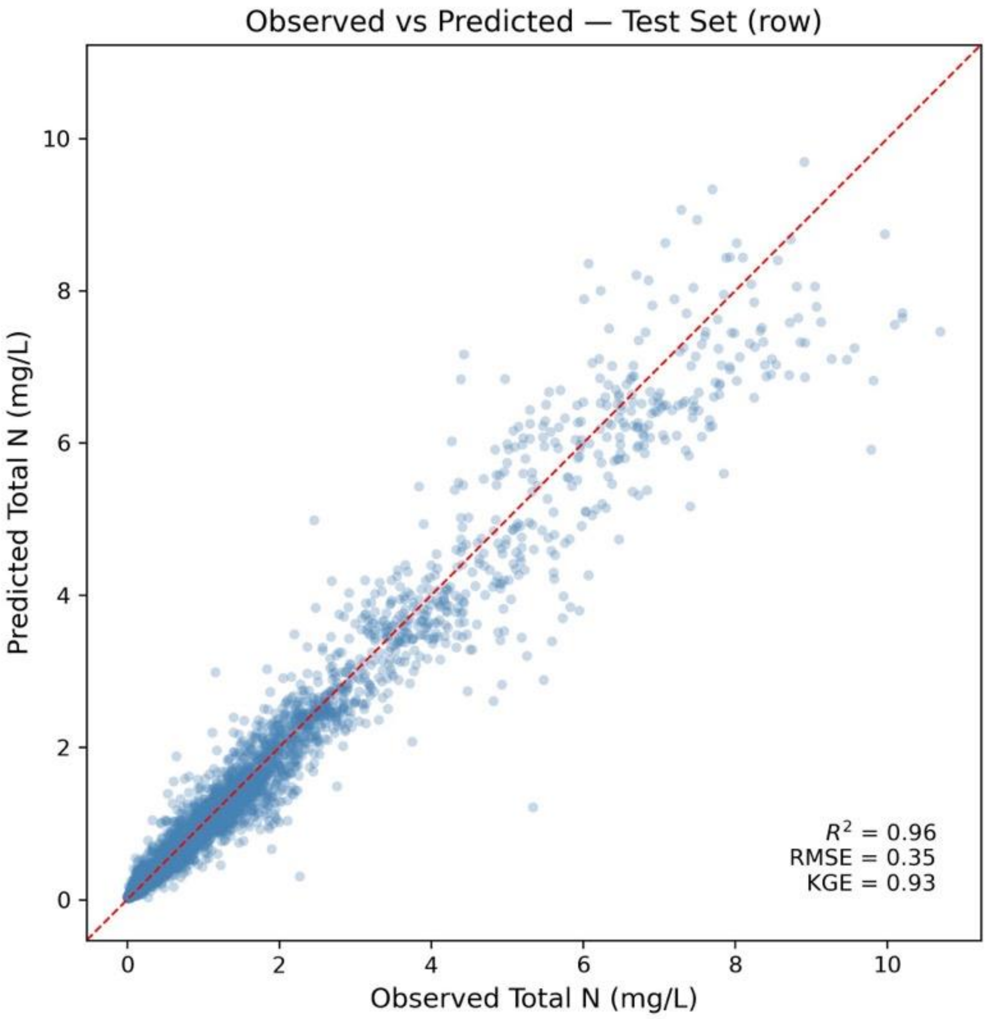


Chang et al., under revision, *Nature Scientific Data*

Random Forest Results

- Develop Random Forest Models to Predict Nutrient Concentrations and Loads at a Monthly Time Scale



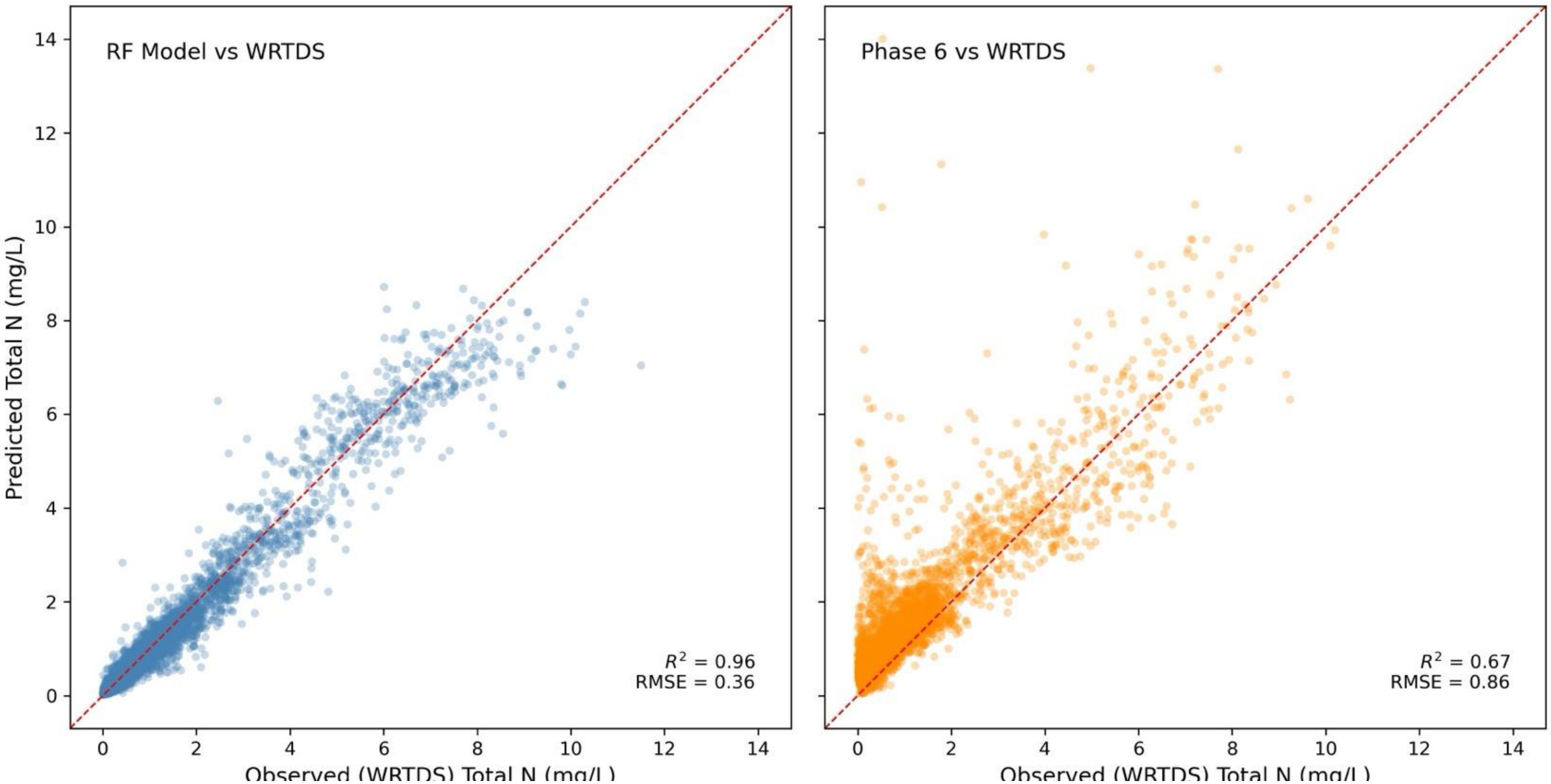


81 NTN Stations
1985-2020

80%/20% Train-Test Split

Current Results

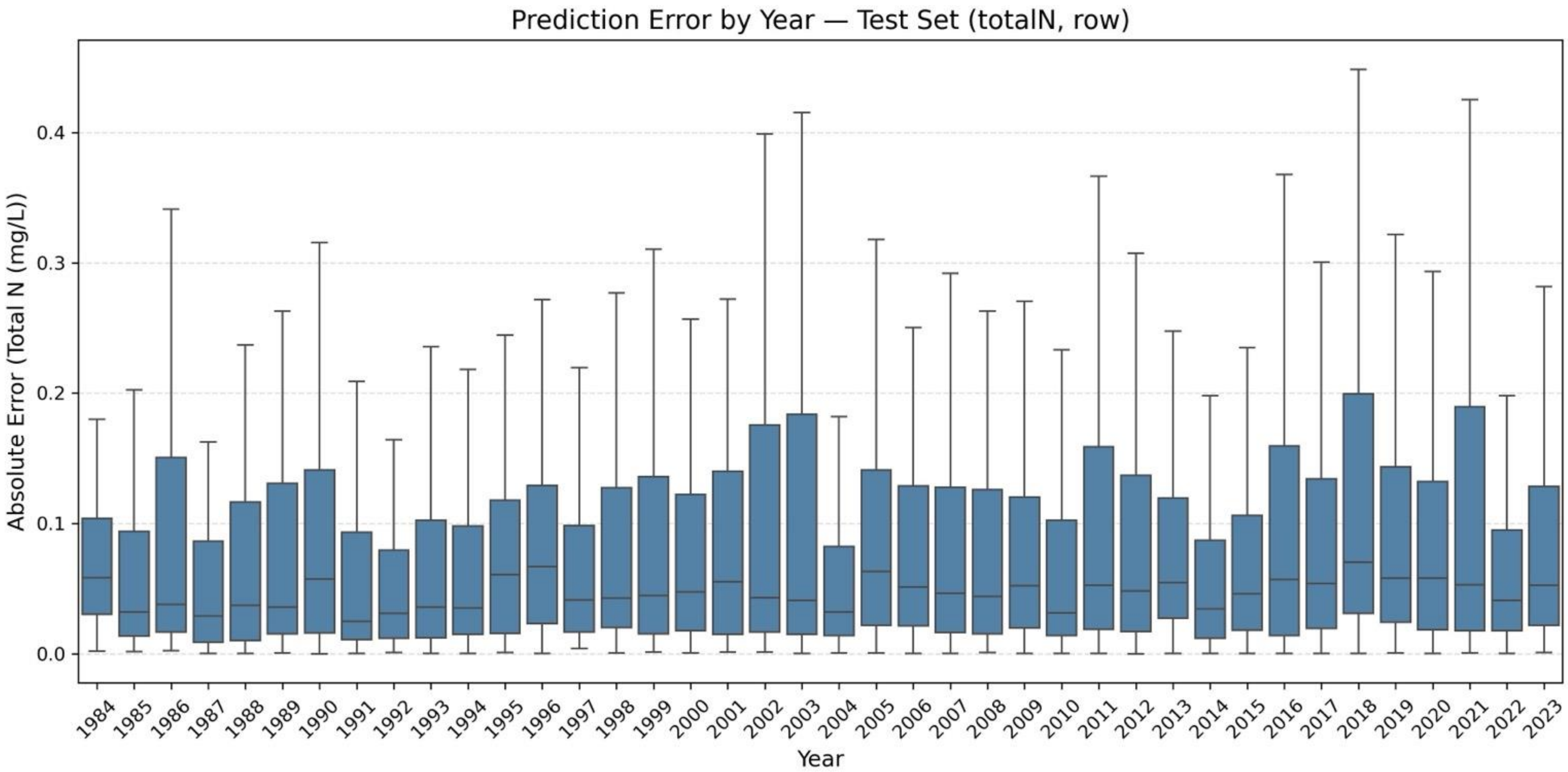
- Develop Random Forest Models to Predict Nutrient Concentrations and Loads at a Monthly Time Scale



RF Model Comparison with Phase 6 Model Predictions

Current Results

- Develop Random Forest Models to Predict Nutrient Concentrations and Loads at a Monthly Time Scale



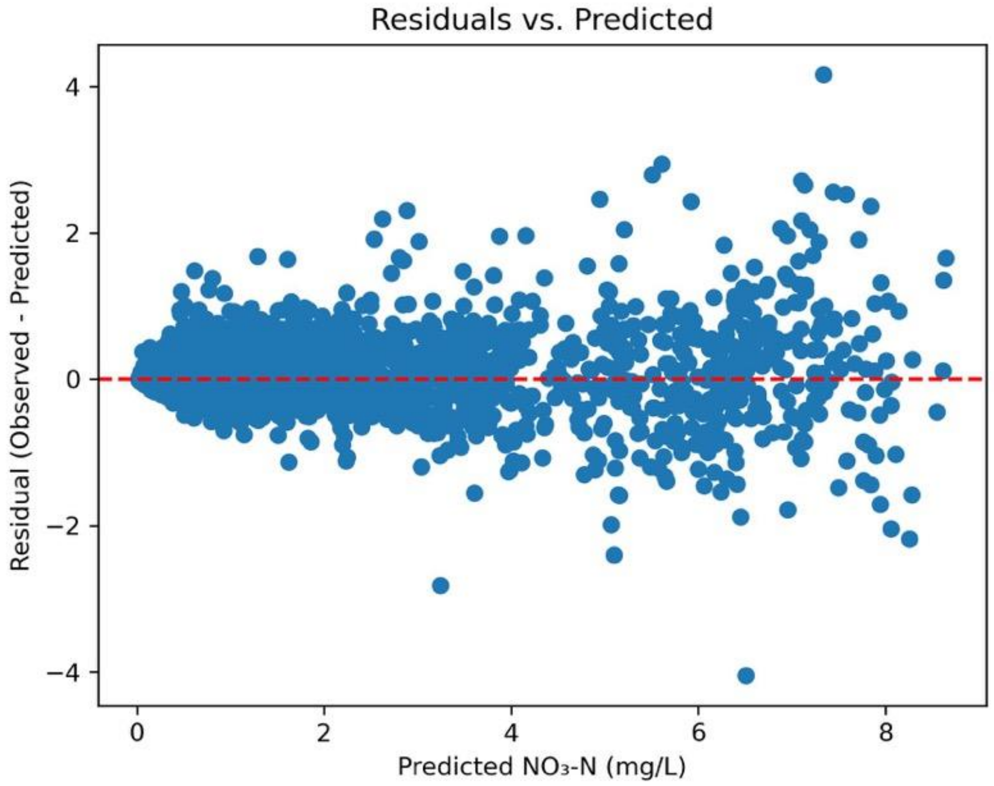
81 NTN Stations
1985-2020

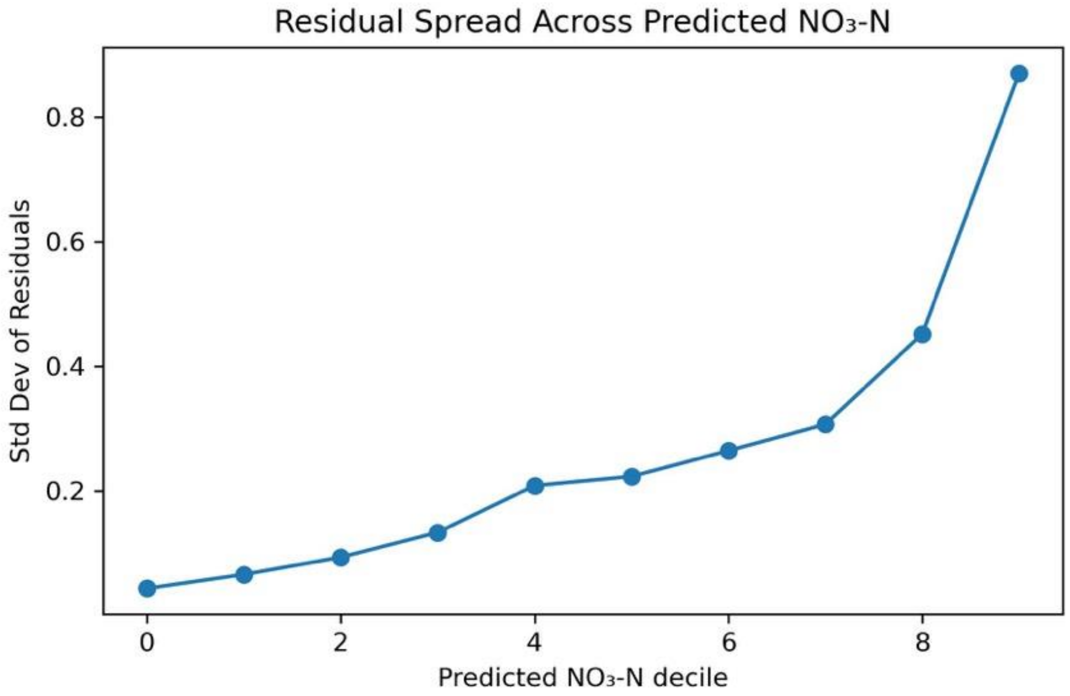
80%/20% Train-Test Split

Chesapeake HydroML

Current Results

- Develop Random Forest Models to Predict Nutrient Concentrations and Loads at a Monthly Time Scale

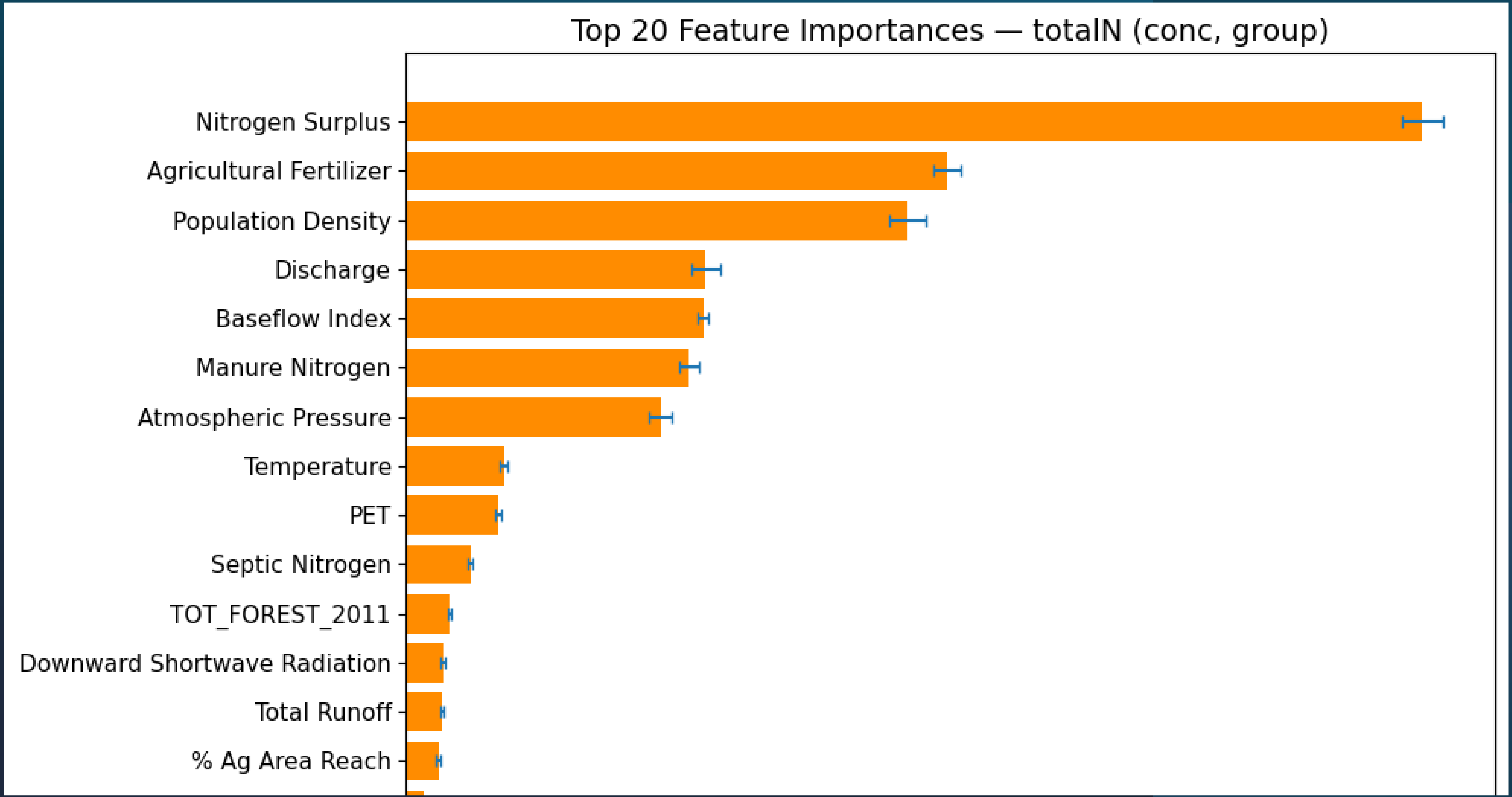




Higher levels of error in high-concentration months/at high-concentration stations

81 NTN Stations
1985-2020

Chesapeake HydroML



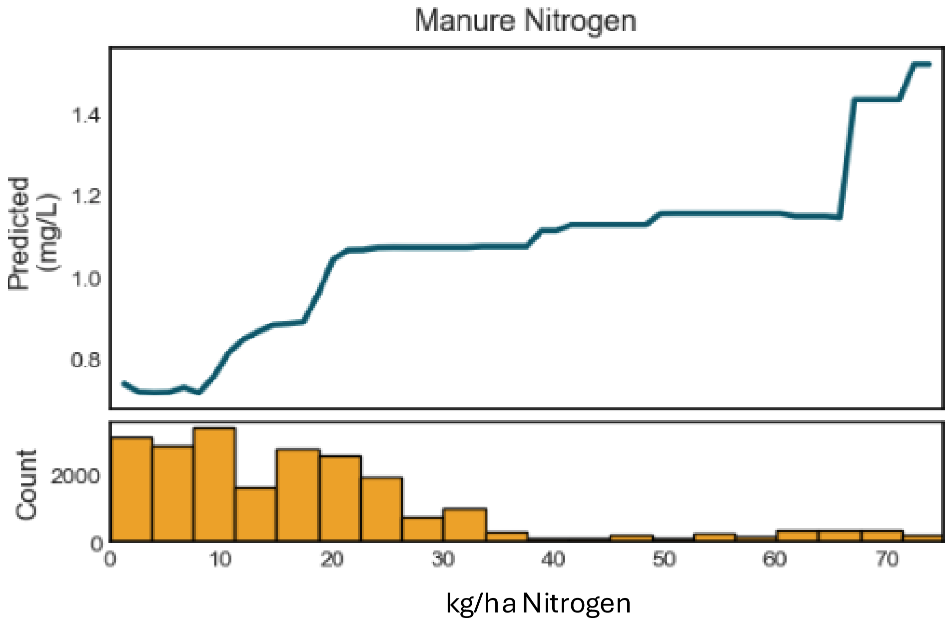
Top 20 Feature Importances — totalN (conc, group)
Nitrogen Surplus
Agricultural Fertilizer
Population Density
Discharge
Baseflow Index
Manure Nitrogen
Atmospheric Pressure
Temperature
PET
Septic Nitrogen
TOT_FOREST_2011
Downward Shortwave Radiation
Total Runoff
% Ag Area Reach

From Machine Learning to Management

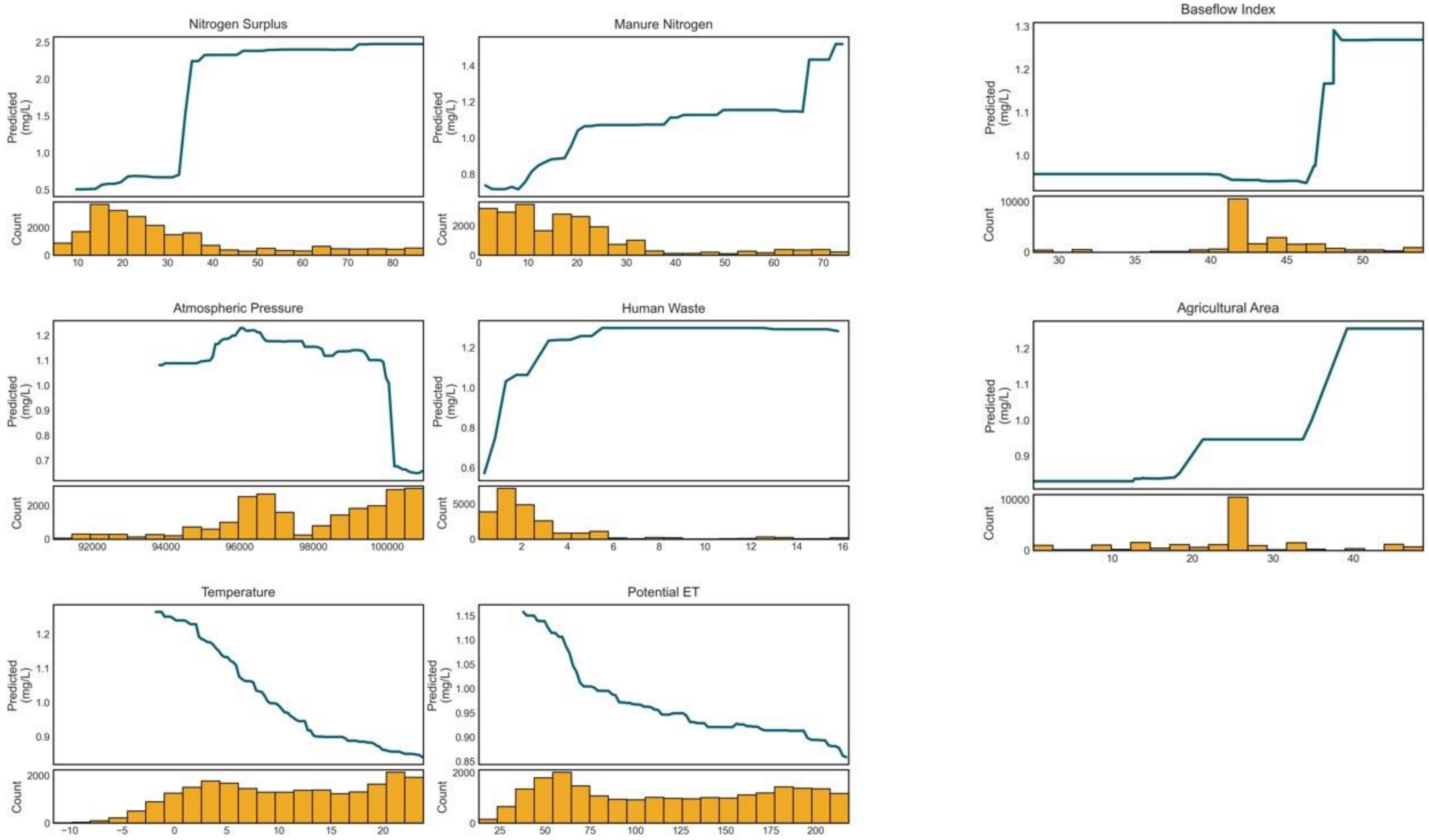
Deriving land-to-water factors

Our **random forest models** predict stream nutrient concentrations (e.g., TN, TP) from watershed-scale predictors such as:

- % agriculture, % forest, impervious surface
- fertilizer or manure inputs
- tile drainage density
- precipitation, soil carbon, slope, etc.



Partial dependence plots (PDPs) show how predicted concentration responds to changes in each predictor, *holding all others constant*


Nitrogen Surplus
Manure Nitrogen
Baseflow Index
Atmospheric Pressure
Human Waste
Agricultural Area
Temperature
Potential ET
Predicted (mg/L)
Count

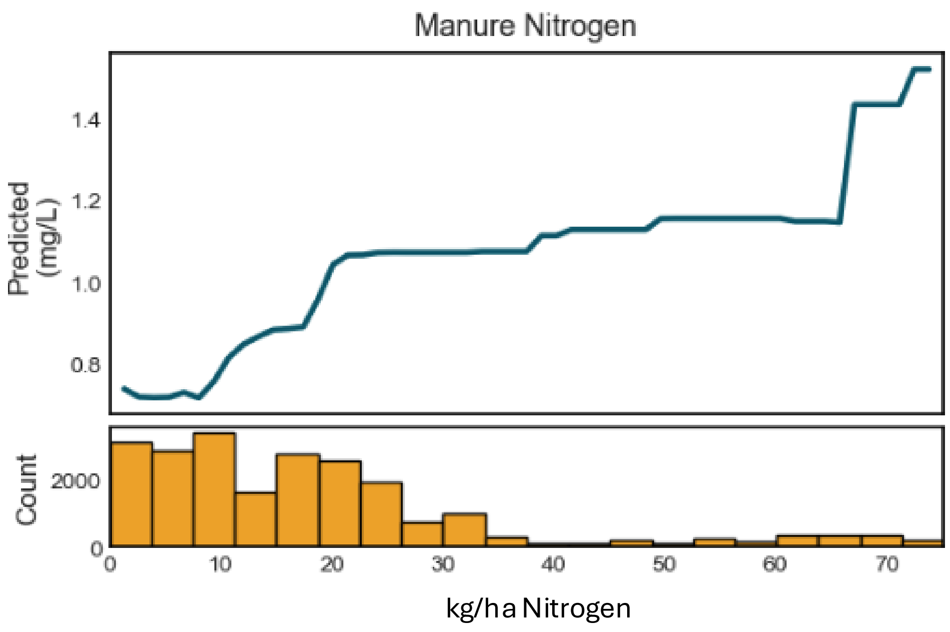
From Machine Learning to Management

Deriving land-to-water factors

Partial dependence plots (PDPs) show how predicted concentration responds to changes in each predictor, *holding all others constant*

From each PDP, Can we calculate a land-to-water response factor?

For example, a PDP might show that decreasing manure N inputs from 70 to 50 kg/ha would decrease total N concentrations by ~ 0.2 mg/L



Work in Progress with Nutrient Modeling

1. Add additional stations with WRTDS data across the CBW
2. Work with fine-scale land-use data and integrate into model
    1. Aggregate stream-reach data across the upstream stream reaches
3. Explicitly add wastewater inputs to the model

From each PDP, Can we calculate land-to-water response factors?