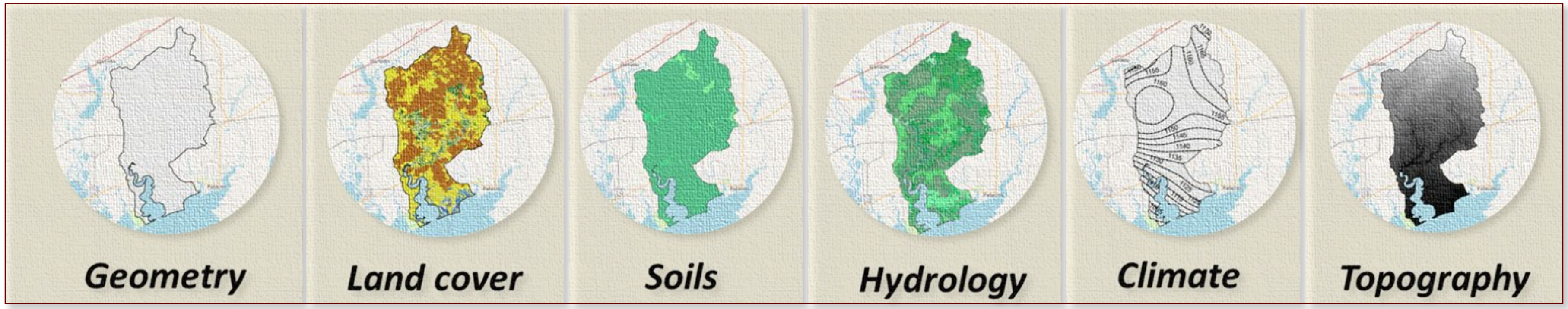



# Prediction and Inference of Instream Nutrient and Sediment Concentrations using Extreme Gradient Boosting (XGB)

**Shubham Jain, Arun Bawa,** Katie Mendoza, Raghavan Srinivasan,
Rajbir Parmar, Deron Smith, Kurt Wolfe, John M Johnston, Joel Corona

TEXAS A&M AGRILIFE RESEARCH

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

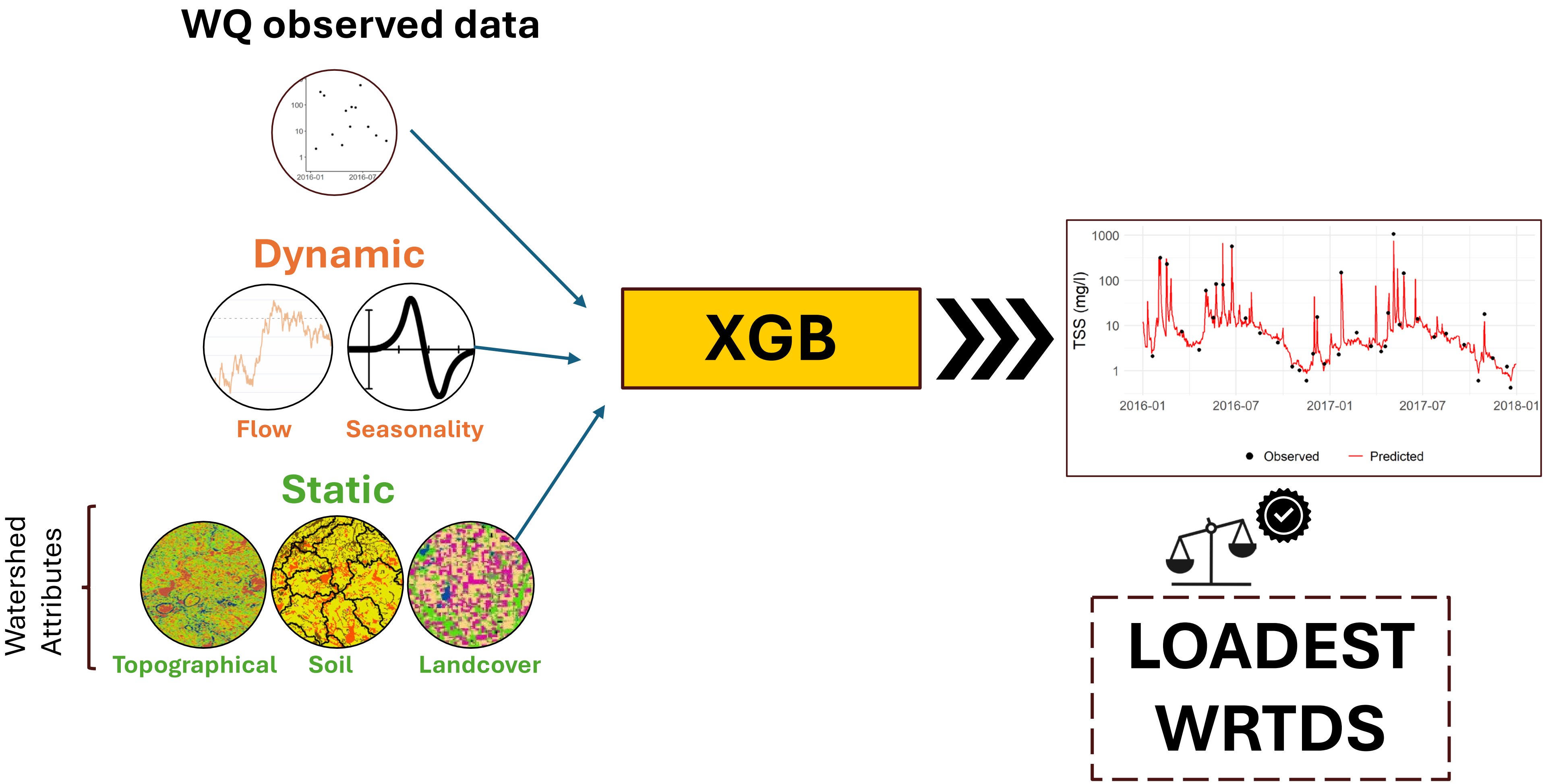
Soil & Water Assessment Tool SWAT

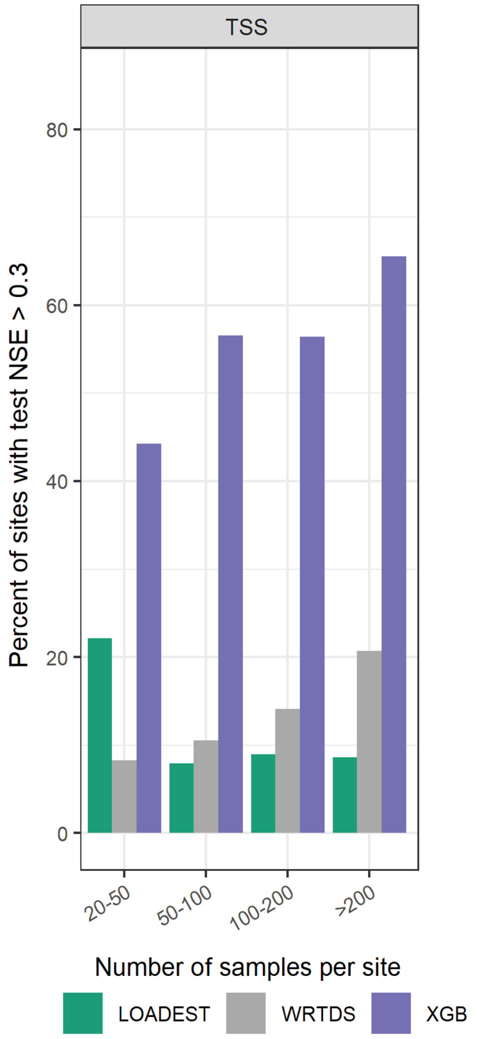
*Shubham Jain*
*PhD candidate*
*Biological & Ag Engineering*
*Texas A&M University*

# Summary

Methodology

Results



# Lack of sufficient monitoring data

## - elevate uncertainty in water quality modeling and decision-making

**Sampling Frequency**

Biweekly

Monthly

Seasonal

Global data counts

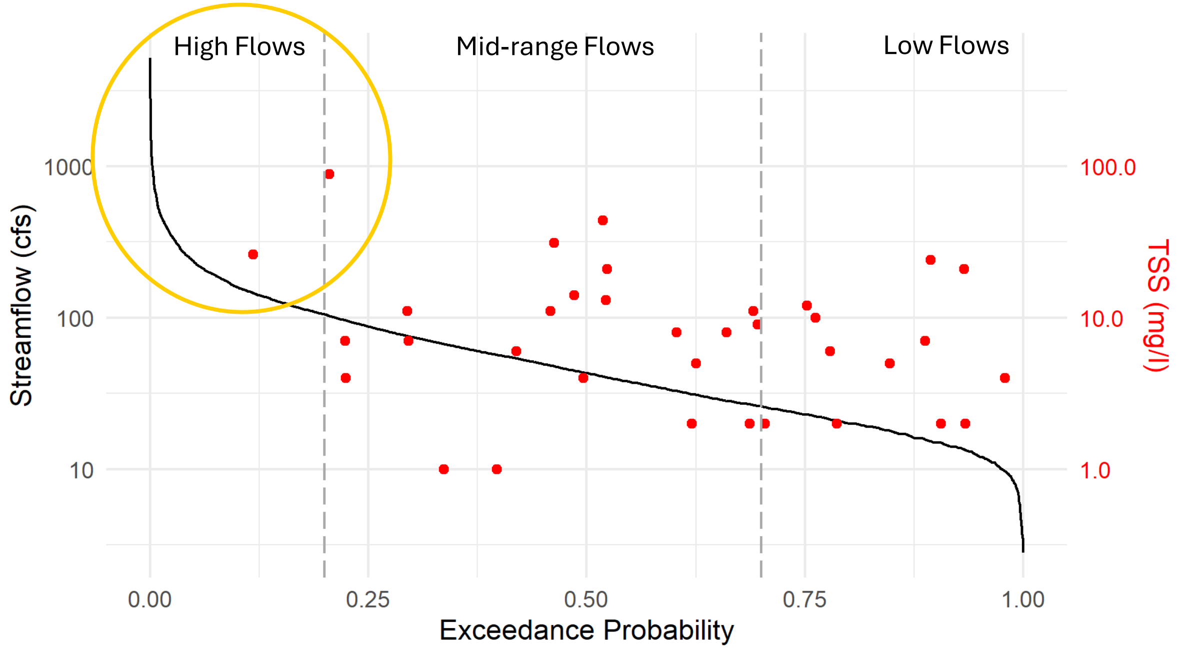
**29** over 4.2 years

**This study**

499 sites, 1996-2020 (25 years)*

- TSS- 71
- TN- 89
- TP- 95

Insufficient sampling values in high flow regime



# LOADEST often results in high biases!

Regression based approach

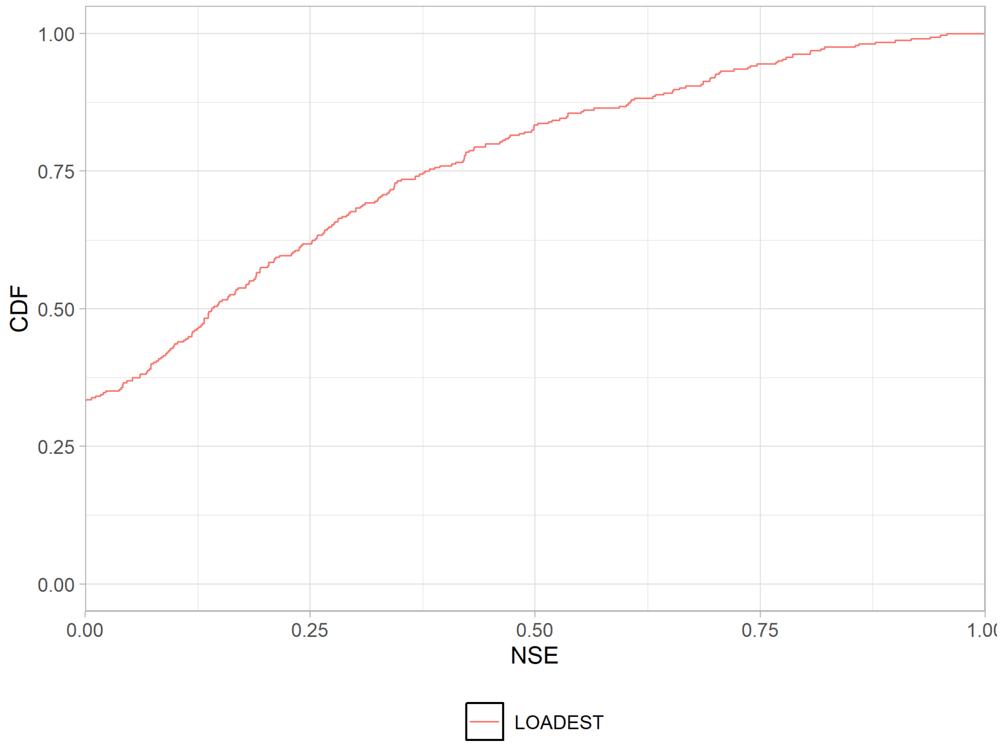
Predictors- Time and Discharge

9 predefined equation | AIC based Selection

Individual sites | > 12 samples

Median sample size TSS- 71
- Training- 57 (80%)
- Test- 14 (20%)

Only training stats



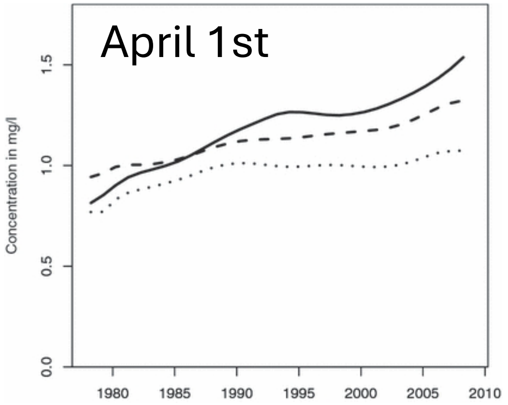
# WRTDS recommended >100 samples over 10 years

$$\ln(C_i) = \beta_{0,i} + \beta_{1,i} t_i + \underbrace{\beta_{2,i} \ln(Q_i)}_{\text{Flow dynamics}} + \underbrace{\beta_{3,i} \sin(2\pi t_i) + \beta_{4,i} \cos(2\pi t_i)}_{\text{Seasonality}} + \underbrace{\varepsilon_i}_{\text{Unexplained variation}}$$

Regression based approach

Predictors- Time, Discharge, **Season**

Weighted Regressions on Time, Discharge, and Season (Hirsch et al., 2010)

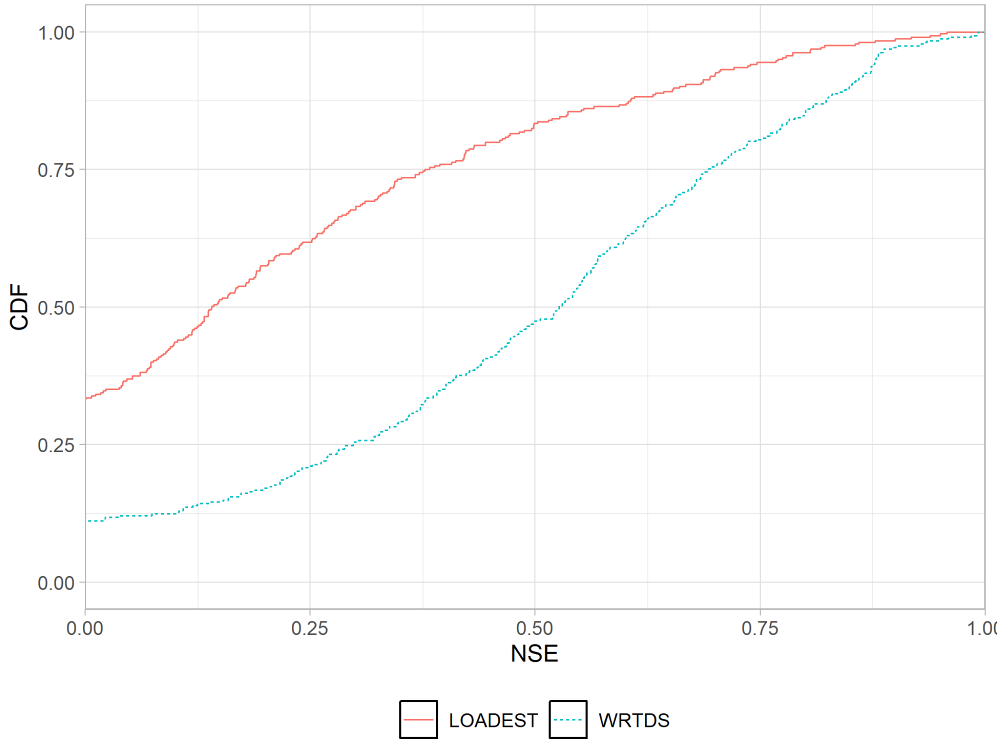



Individual sites

> 100 samples

Median sample size TSS- 71

> 10 years



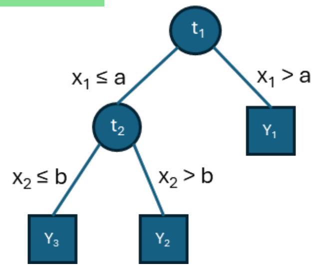

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs. Journal of the American Water Resources Association, 46(5), 857-880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

# XGB trained on combined WQ data improved predictions!

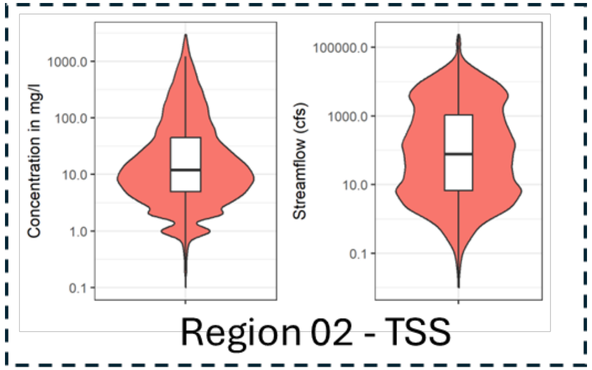
Tree based regression approach

Optimized handling of sparse and missing data

Ability to incorporate regularization to prevent overfitting

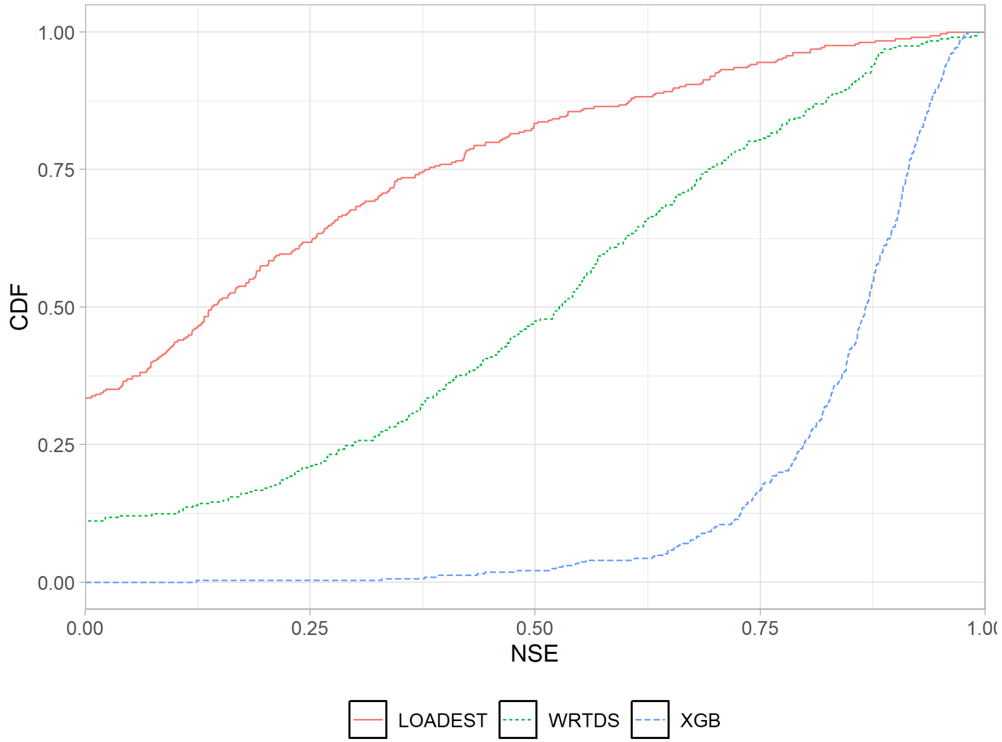


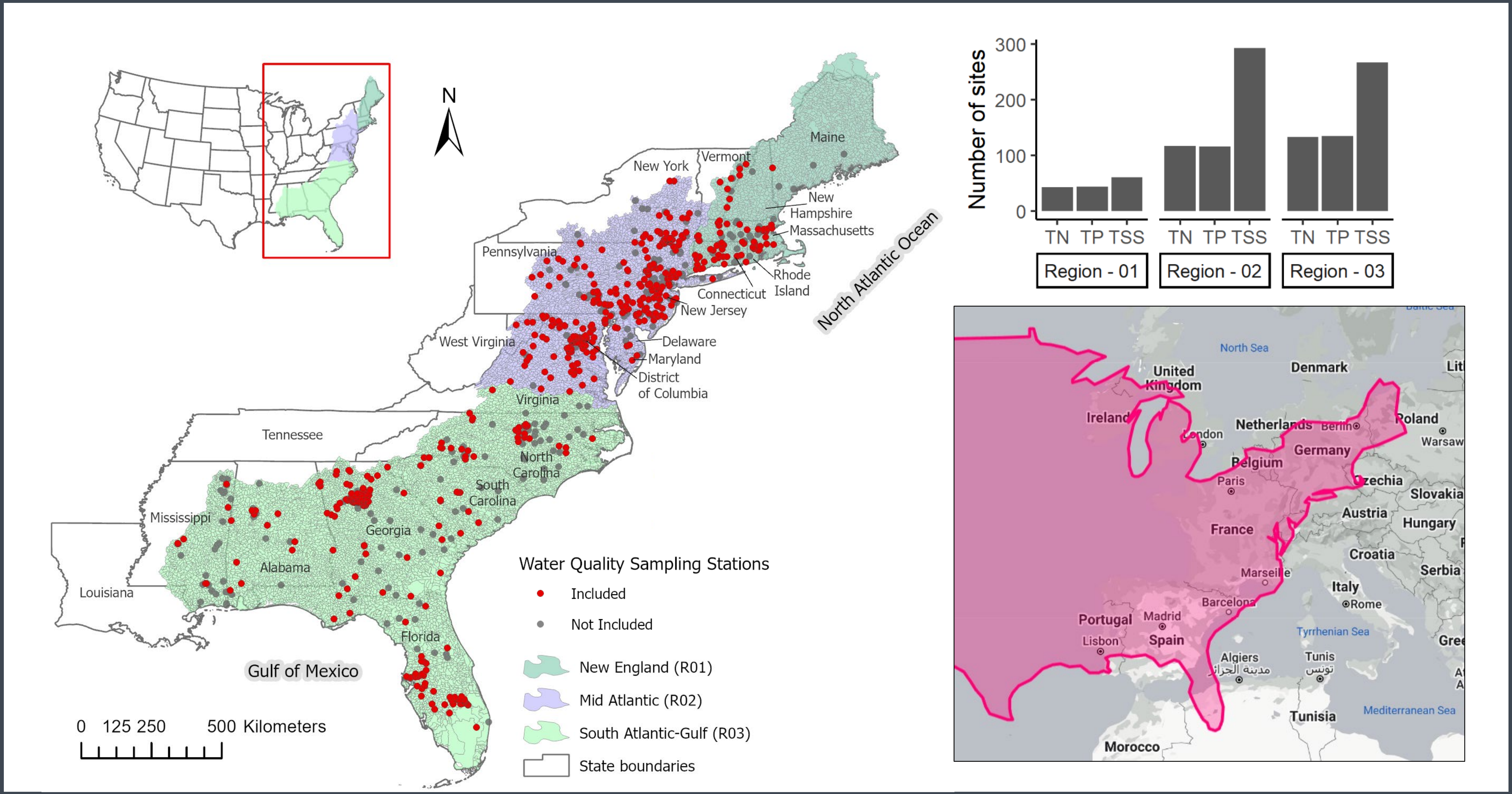

Combining WQ data across sites overcomes limitation of insufficient data at individual sites




**One region - One model**

Daily WQ prediction at individual sites




N
New York
Vermont
Maine
New
Hampshire
Massachusetts
Pennsylvania
Rhode
Island
Connecticut
New Jersey
North Atlantic Ocean
West Virginia
Delaware
Maryland
District
of Columbia
Virginia
Tennessee
North
Carolina
South
Carolina
Mississippi
Georgia
Alabama
Louisiana
Florida
Gulf of Mexico
0 125 250 500 Kilometers
Water Quality Sampling Stations
Included
Not Included
New England (R01)
Mid Atlantic (R02)
South Atlantic-Gulf (R03)
State boundaries
Number of sites
300
200
100
0
TN TP TSS
Region - 01
Region - 02
Region - 03
North Sea
United
Kingdom
Denmark
Ireland
London
Netherlands
Berlin
Poland
Warsaw
Germany
Belgium
Paris
Czechia
Slovakia
Austria
Hungary
France
Croatia
Serbia
Marseille
Italy
Rome
Barcelona
Portugal
Madrid
Spain
Lisbon
Tyrrhenian Sea
Algiers
Tunis
Tunisia
Mediterranean Sea
Morocco

# HAWQS

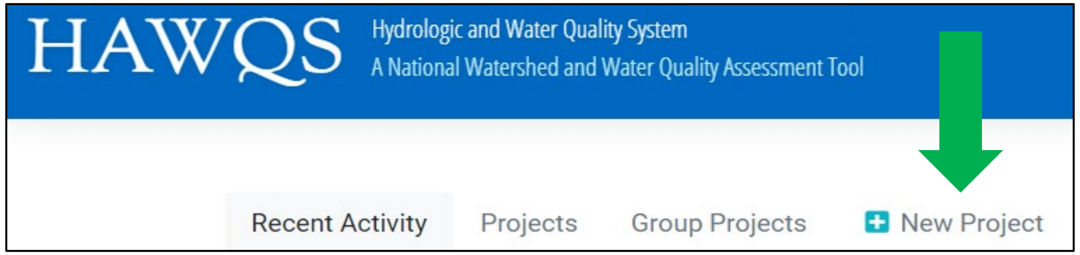
## Hydrologic and Water Quality System

TEXAS A&M AGRILIFE RESEARCH

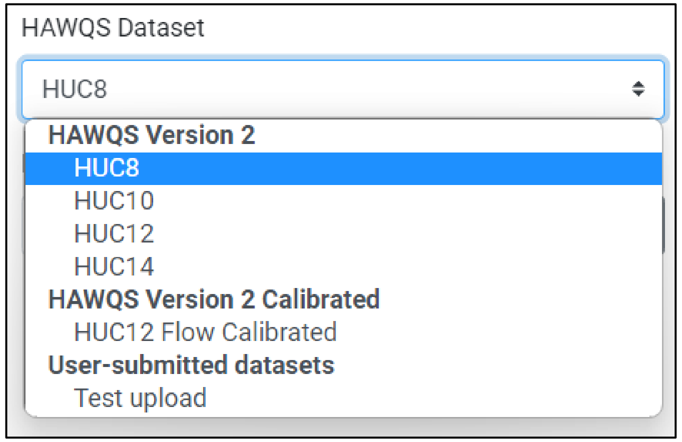
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

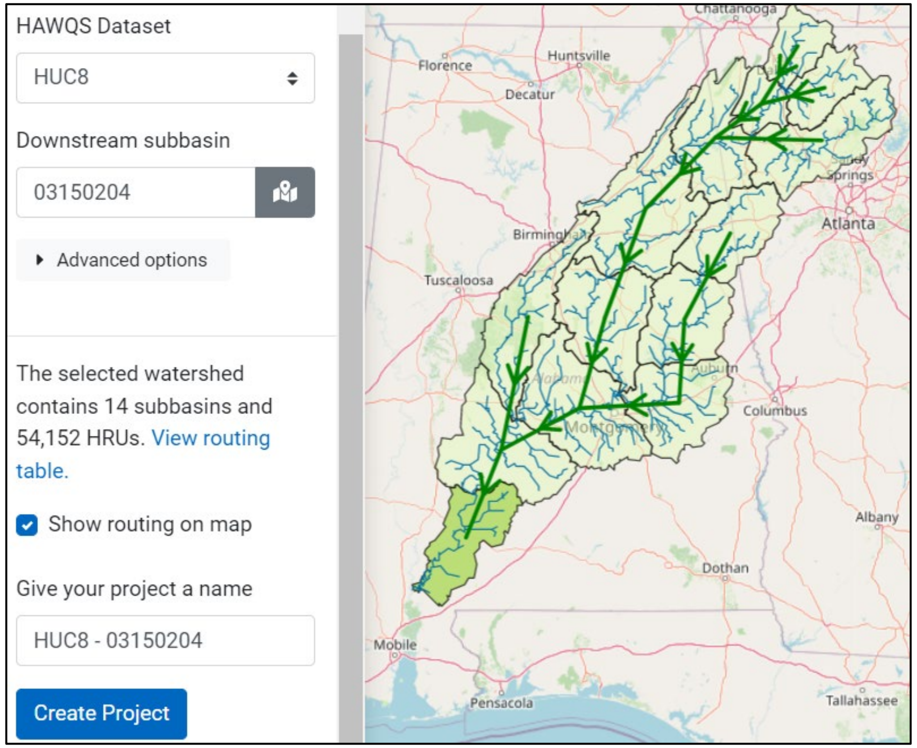
USDA ars Agricultural Research Service

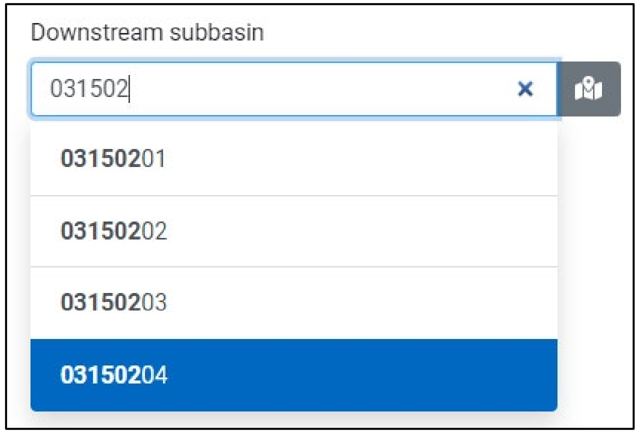
# SWAT Model setup- just a few clicks away!!




hawqs.tamu.edu



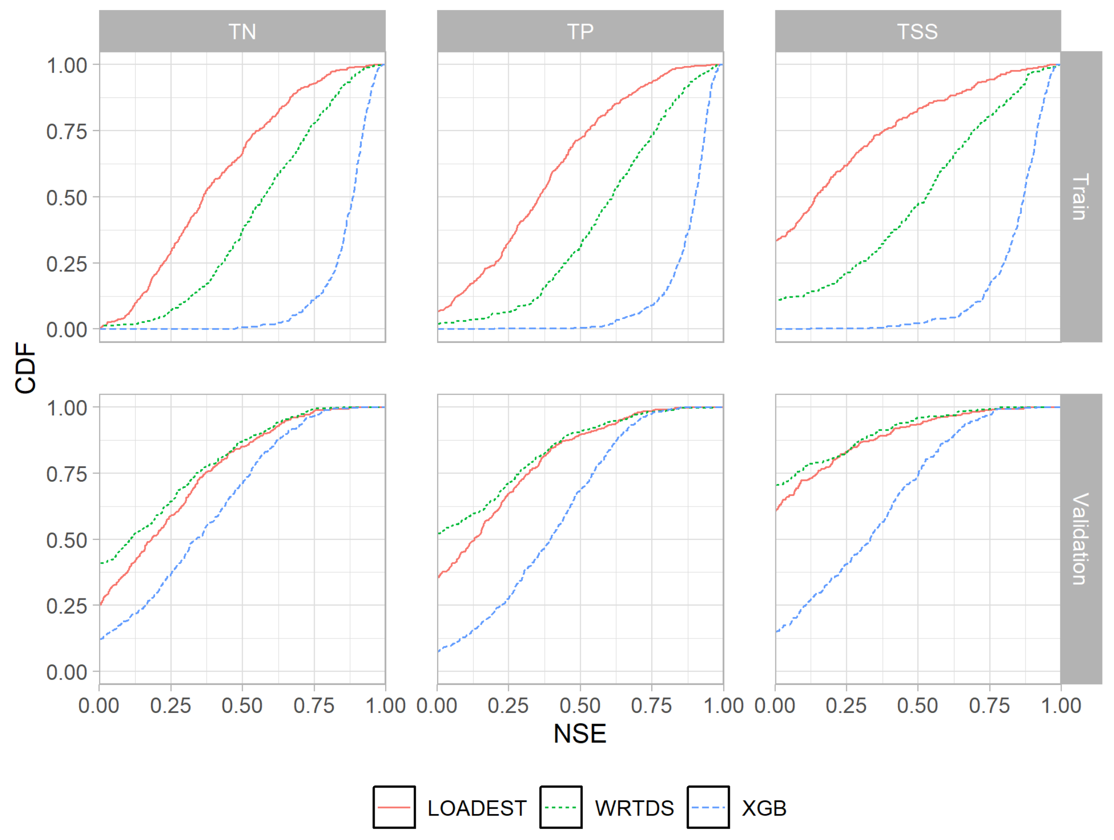



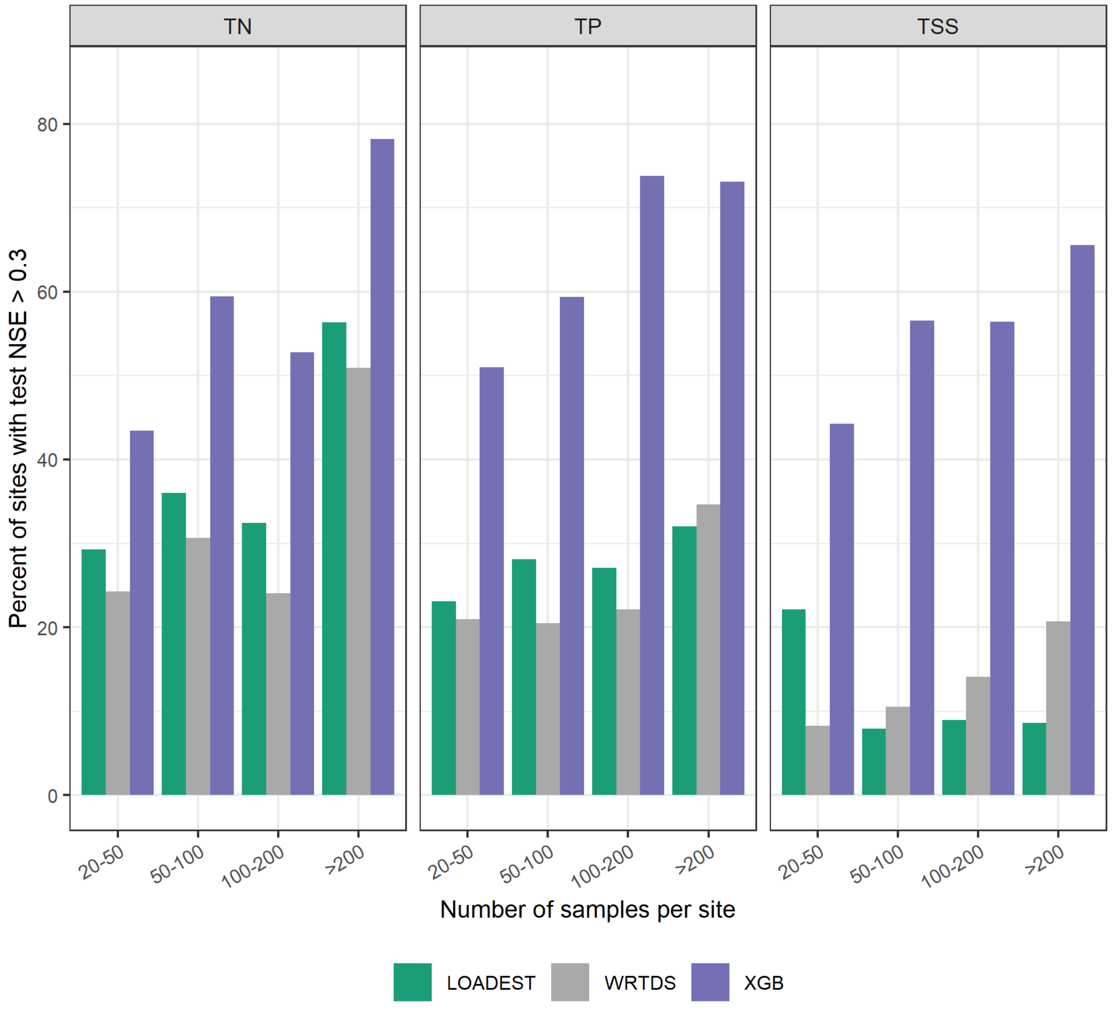

**HAWQS API**

# INTERNATIONAL HAWQS PLATFORMS

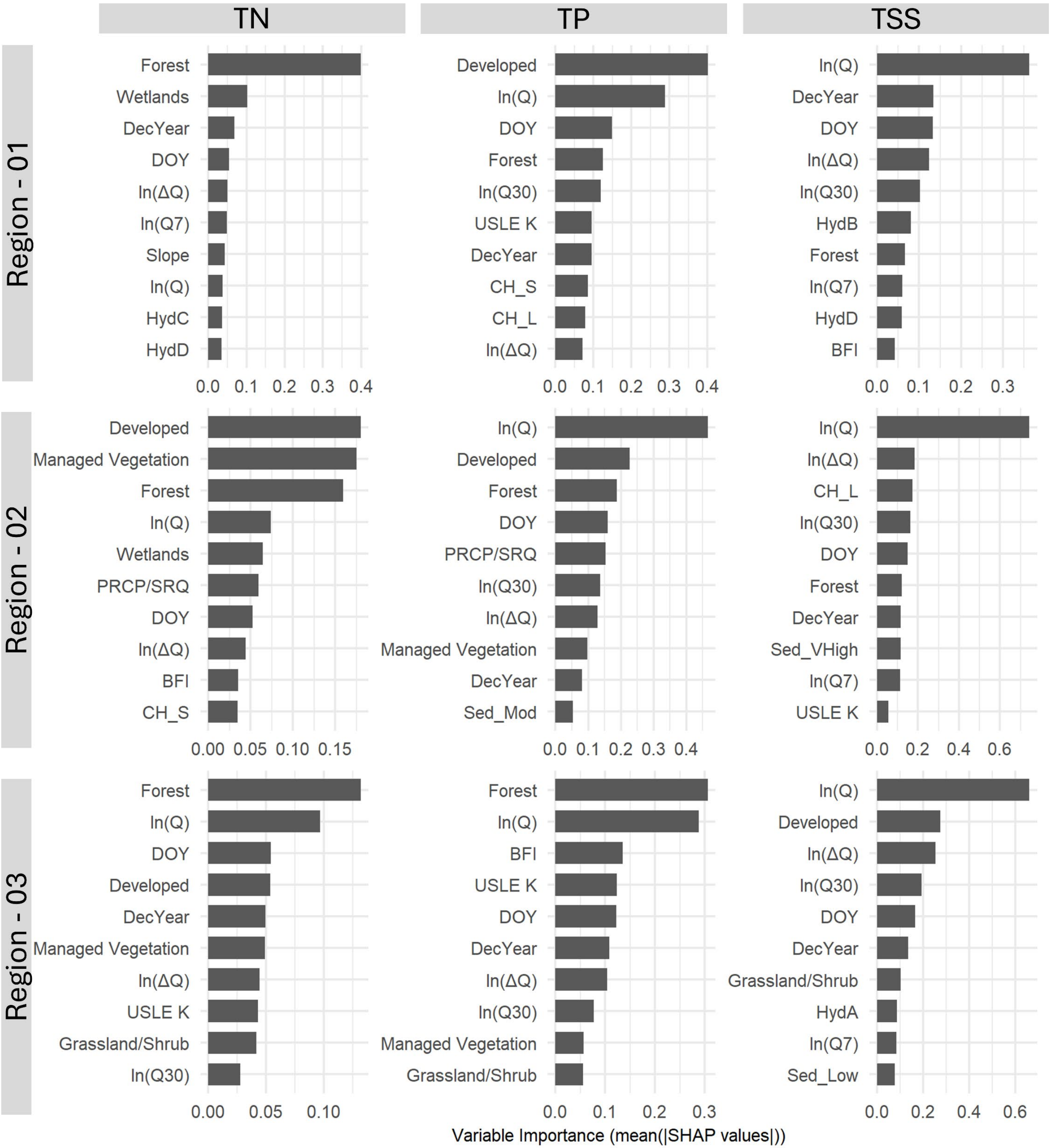
- South Africa (HAMSA)
  [hamsa.hawqs.tamu.edu]
- Pernambuco Brazil (SUPer)
  [super.hawqs.tamu.edu]
- Hydrologic Unit Model for InDia (HUMID)
  [bhuvan.nrcs.gov.in]
- Global HAWQS
  [global.hawqs.tamu.edu]
- Coming Soon: Ukraine, Nepal

# XGB outperformed LOADEST and WRTDS



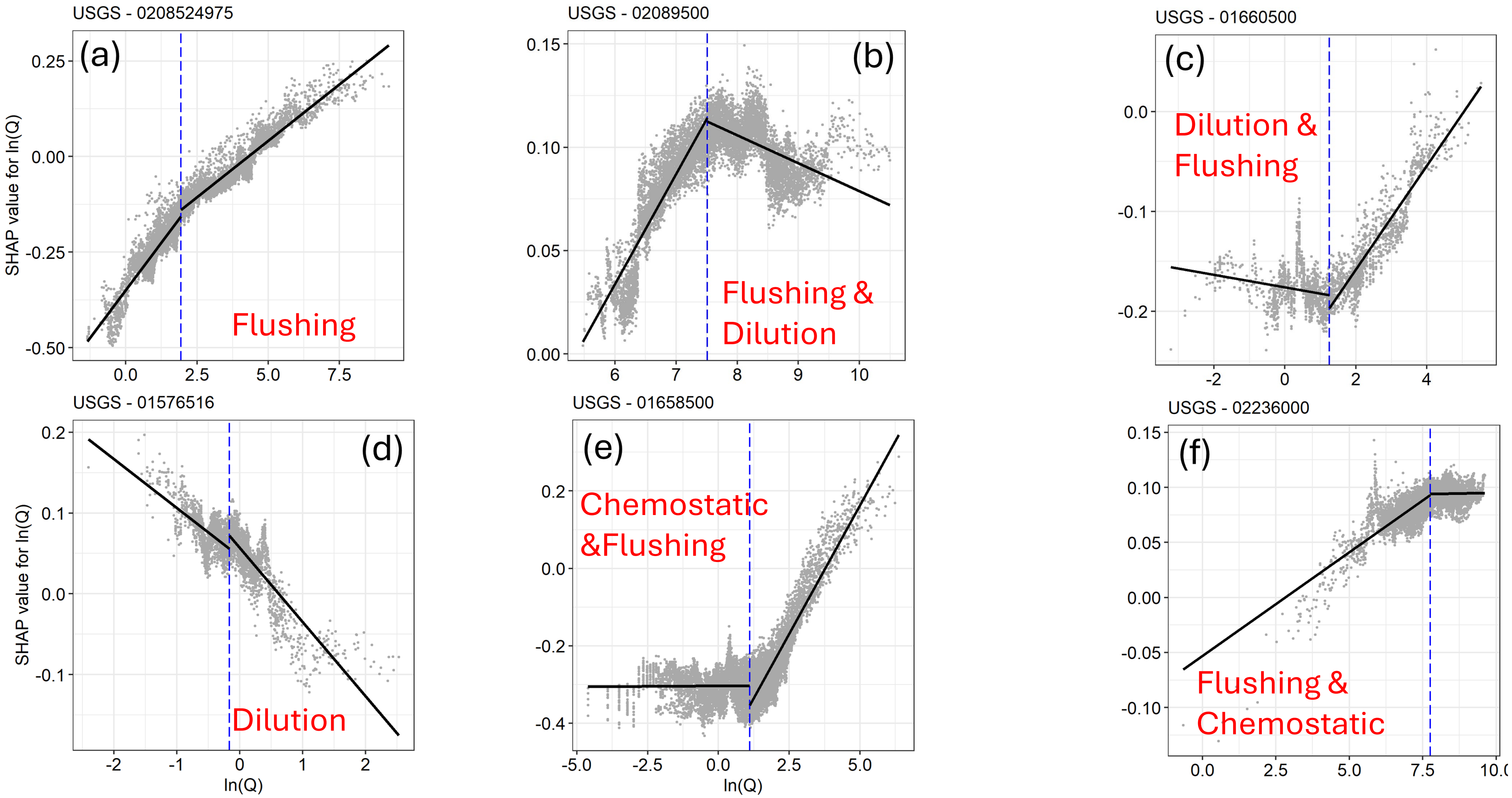



# Watershed attributes played key role in WQ predictions



# Six C-Q pattern: TN

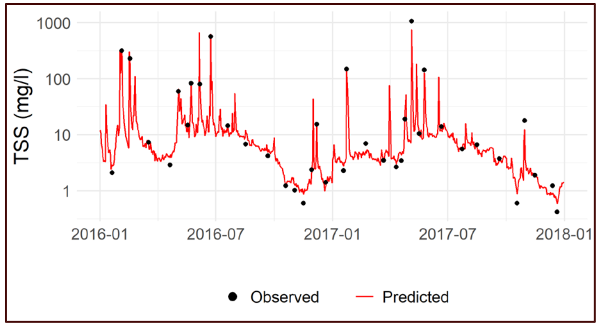
Flushing:
> 95% TSS & TP



## Key Takeaways

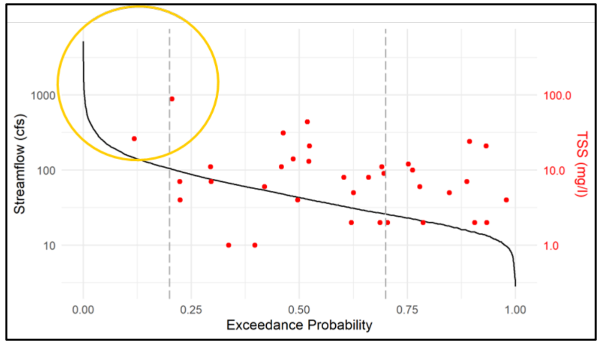
Water Research
under review

- ✓ New ML based WQ interpolation/extrapolation tool

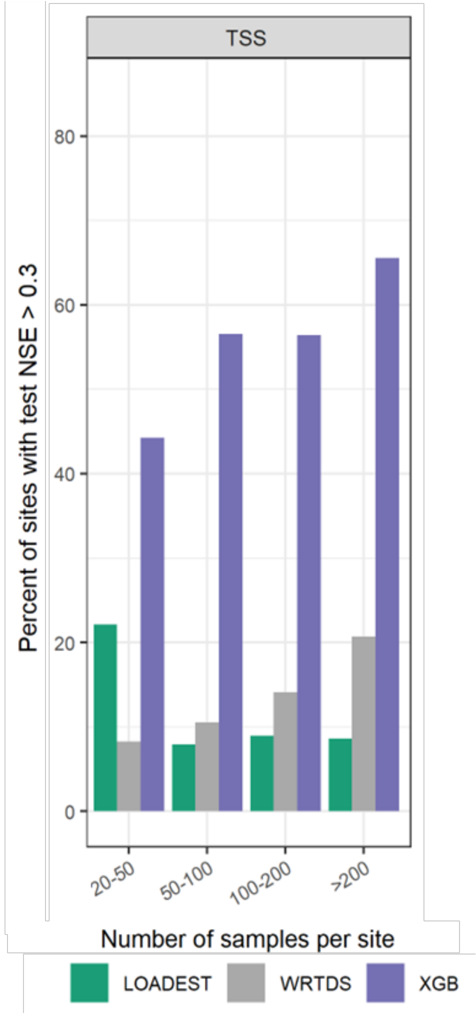



- ✓ Daily WQ estimates for US - **HAWQS**

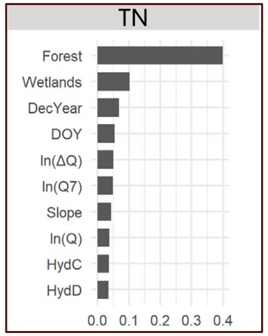
- ✓ **XGB** model outperforms LOADEST and WRTDS

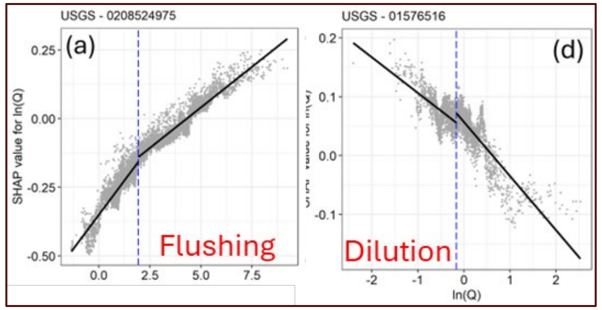



- ✓ Combining WQ data across sites overcomes limitation of insufficient data at individual sites




- ✓ ML-WQ inferences using **Explainable AI**
    - aid in model interpretation
    - increasing trust in Black-Box model



***XGB-WQ-Prediction***
***GITHUB***

***Predicted WQ data for USA***
***Texas Data Repository***

***[HAWQS](hawqs.tamu.edu)***
***hawqs.tamu.edu***

# ***Thank you!***


***Arun Bawa***
*(arun.bawa@ag.tamu.edu)*
*Texas A&M AgriLife Research*
*Blackland Research Ext. Center*


TEXAS A&M AGRILIFE RESEARCH