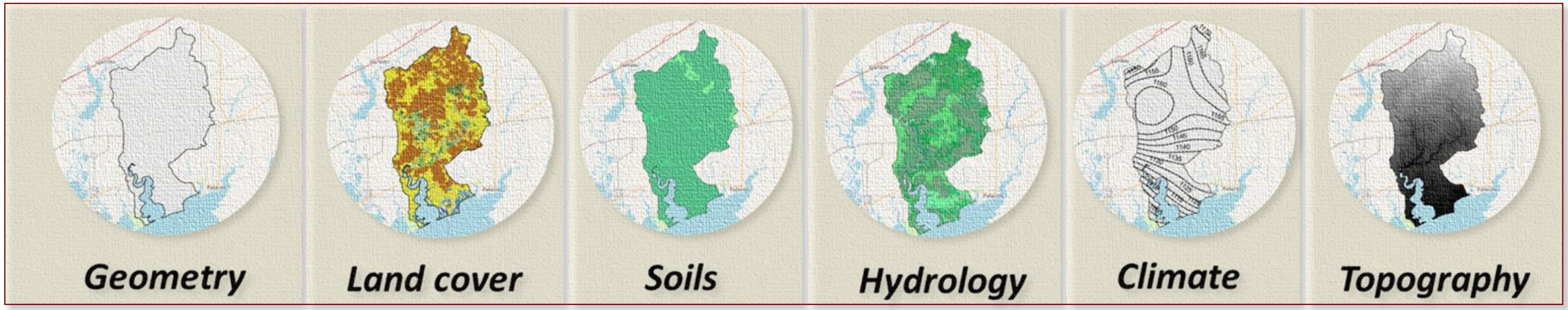


XGBest – An extreme gradient boosting-based tool for estimating daily in-stream nutrient and sediment concentrations

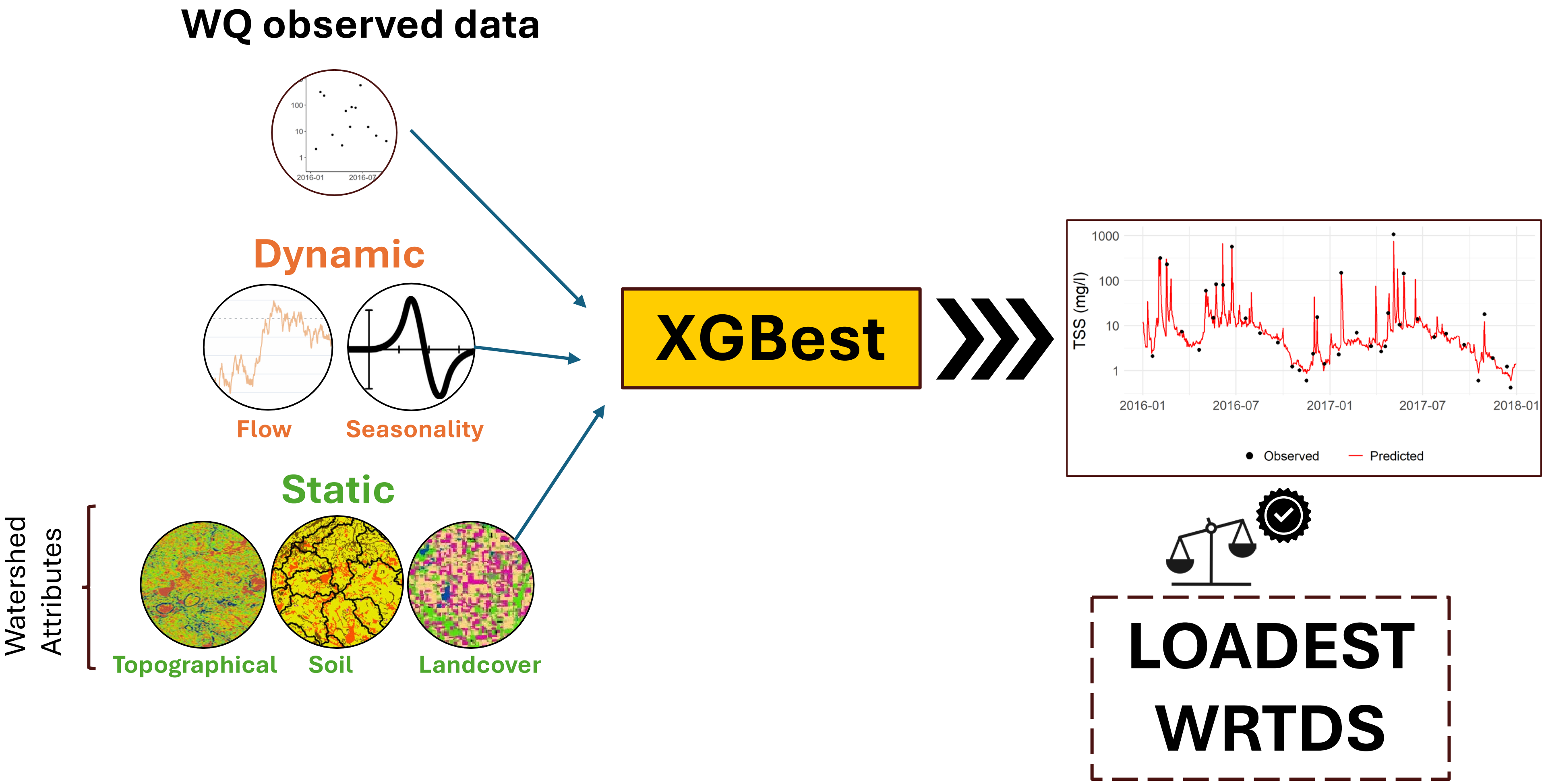
Shubham Jain, Arun Bawa, Katie Mendoza, Raghavan Srinivasan,
Rajbir Parmar, Deron Smith, Kurt Wolfe, John M Johnston, Joel Corona

Jain et al., 2025, Science of Total Environment

Summary

Methodology

Results



WQ observed data

Dynamic

Flow

Seasonality

Static

Watershed Attributes

Topographical

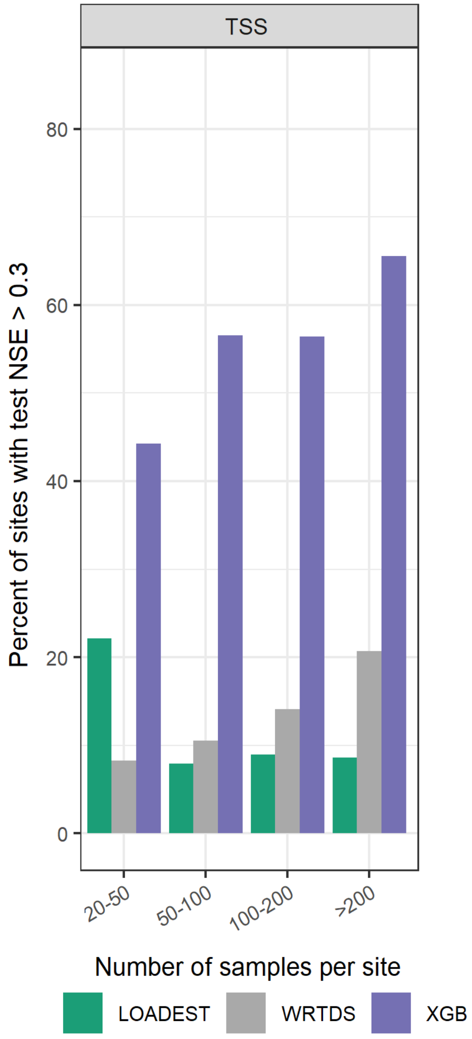
Soil

Landcover

XGBest

LOADEST

WRTDS



Lack of sufficient monitoring data

- elevate uncertainty in water quality modeling and decision-making

Sampling Frequency

Biweekly

Monthly

Seasonal

Global data counts

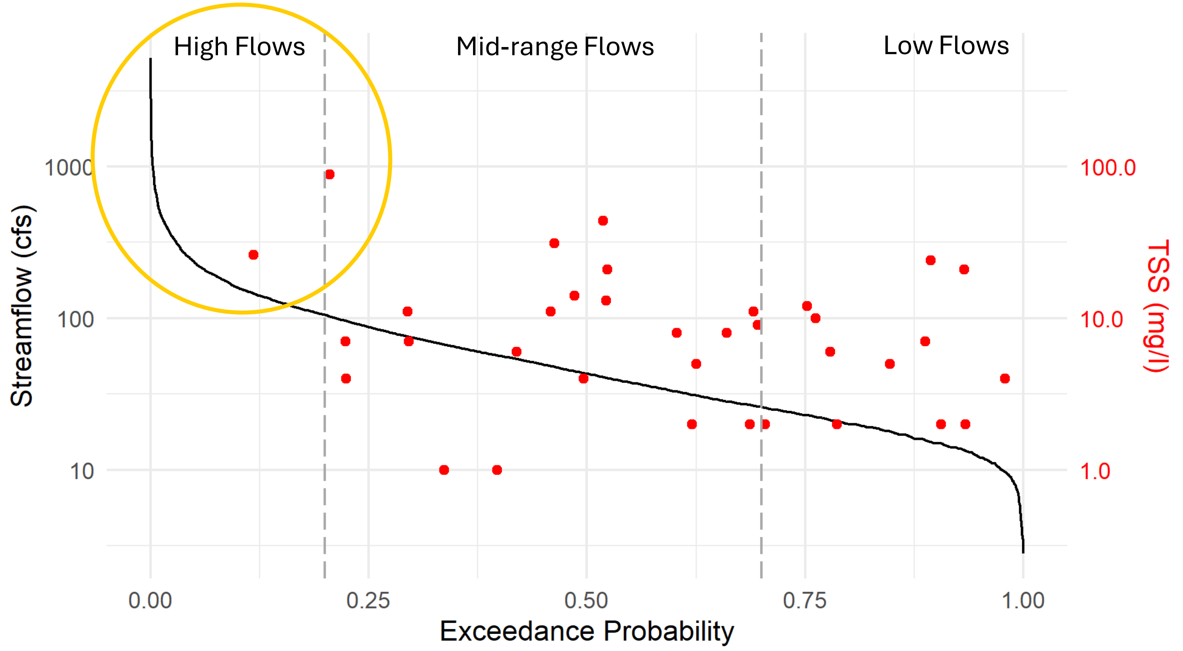
29 over 4.2 years

This study

499 sites, 1996-2020 (25 years)*

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

Insufficient sampling values in high flow regime



LOADEST a regression-based approach, often lead to over estimation

Regression based approach

Predictors- Time and Discharge

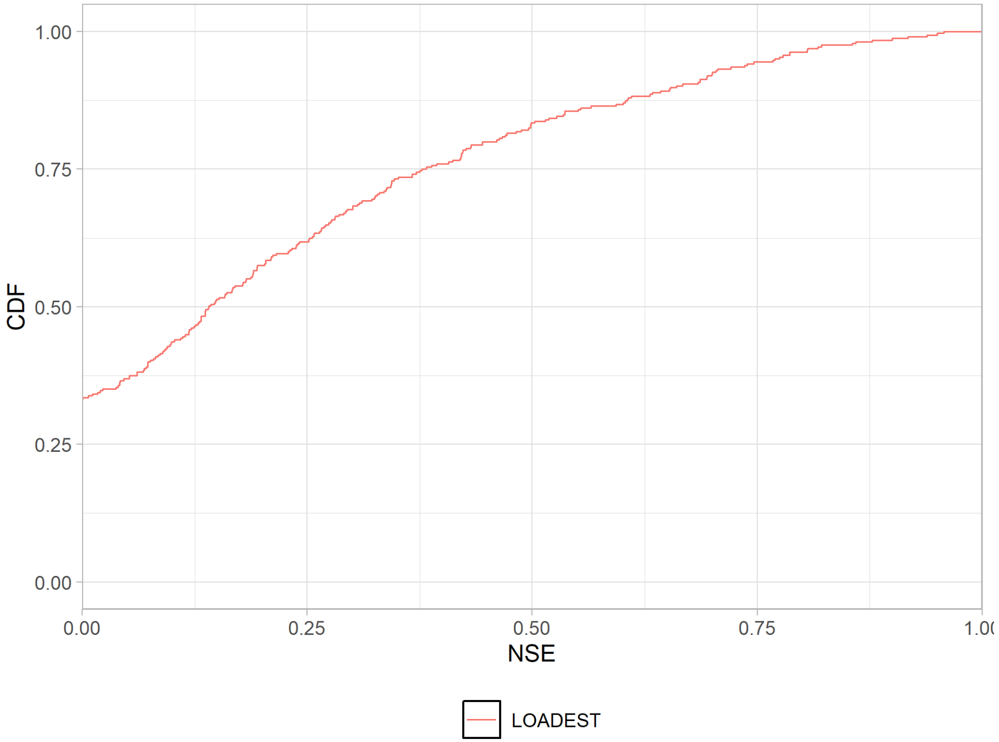
9 predefined equation | AIC based Selection

Individual sites | > 12 samples

Median sample size TSS- 71

- Training- 57 (80%)
- Test- 14 (20%)

Only training stats



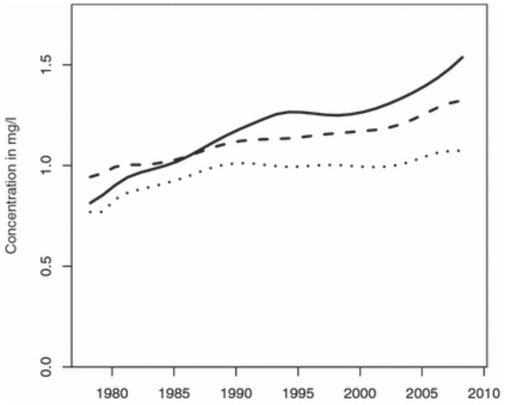
WRTDS recommended >200 samples over 20 years

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

Regression based approach

Predictors- Time, Discharge, **Season**

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

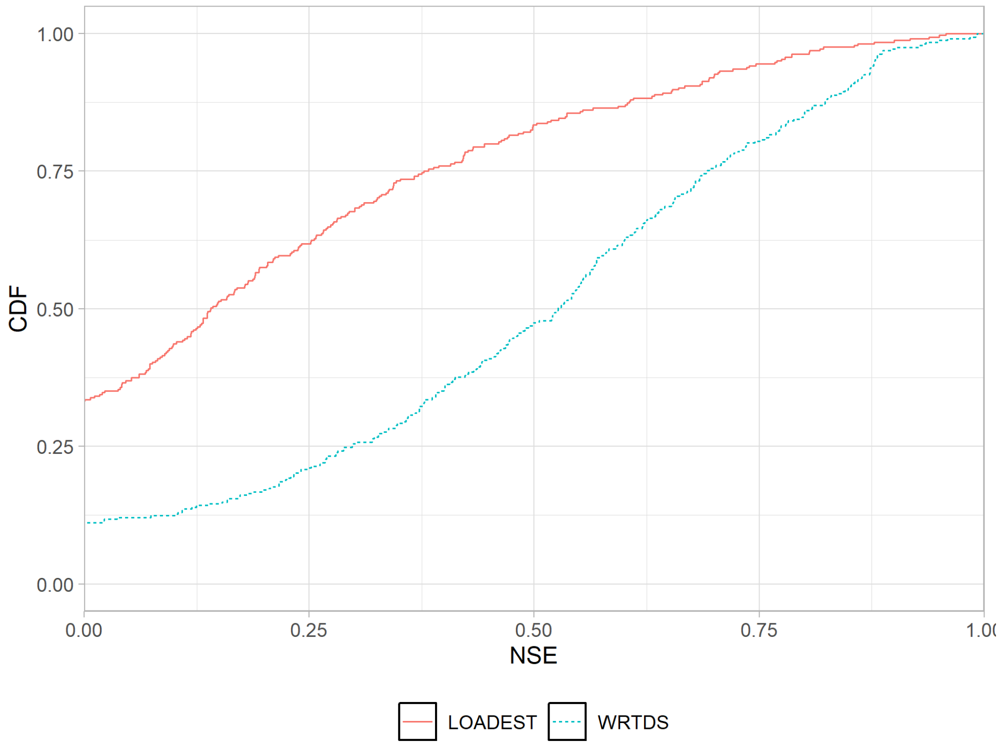


Individual sites

> 200 samples

Median sample size TSS- 71

> 20 years



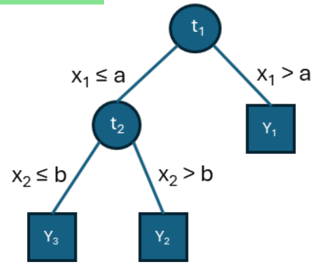
Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs. Journal of the American Water Resources Association, 46(5), 857-880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

XGB trained on combined WQ data improved predictions!

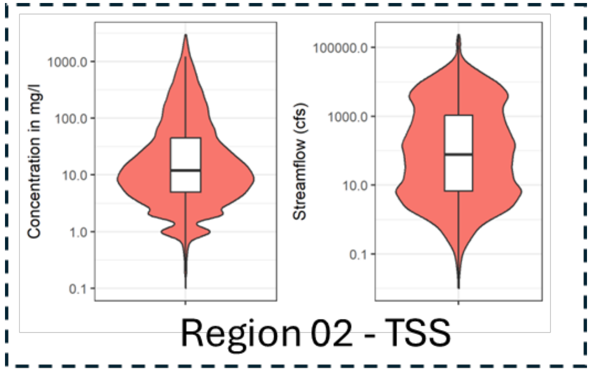
Tree based regression approach

Optimized handling of sparse and missing data

Ability to incorporate regularization to prevent overfitting



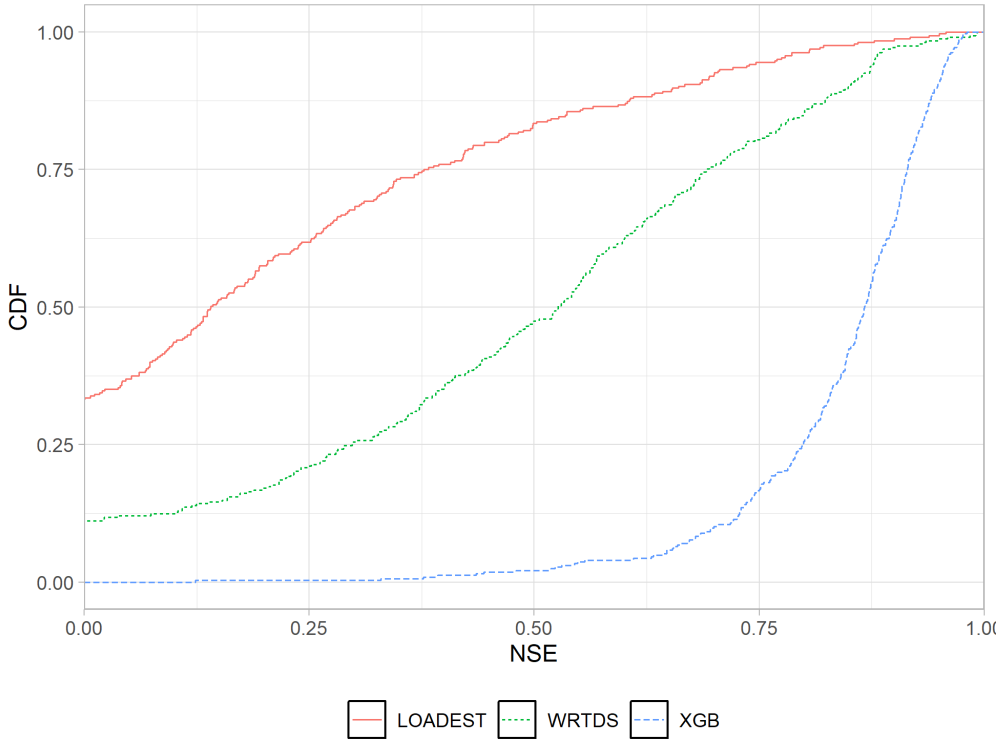
Combining WQ data across sites overcomes limitation of insufficient data at individual sites

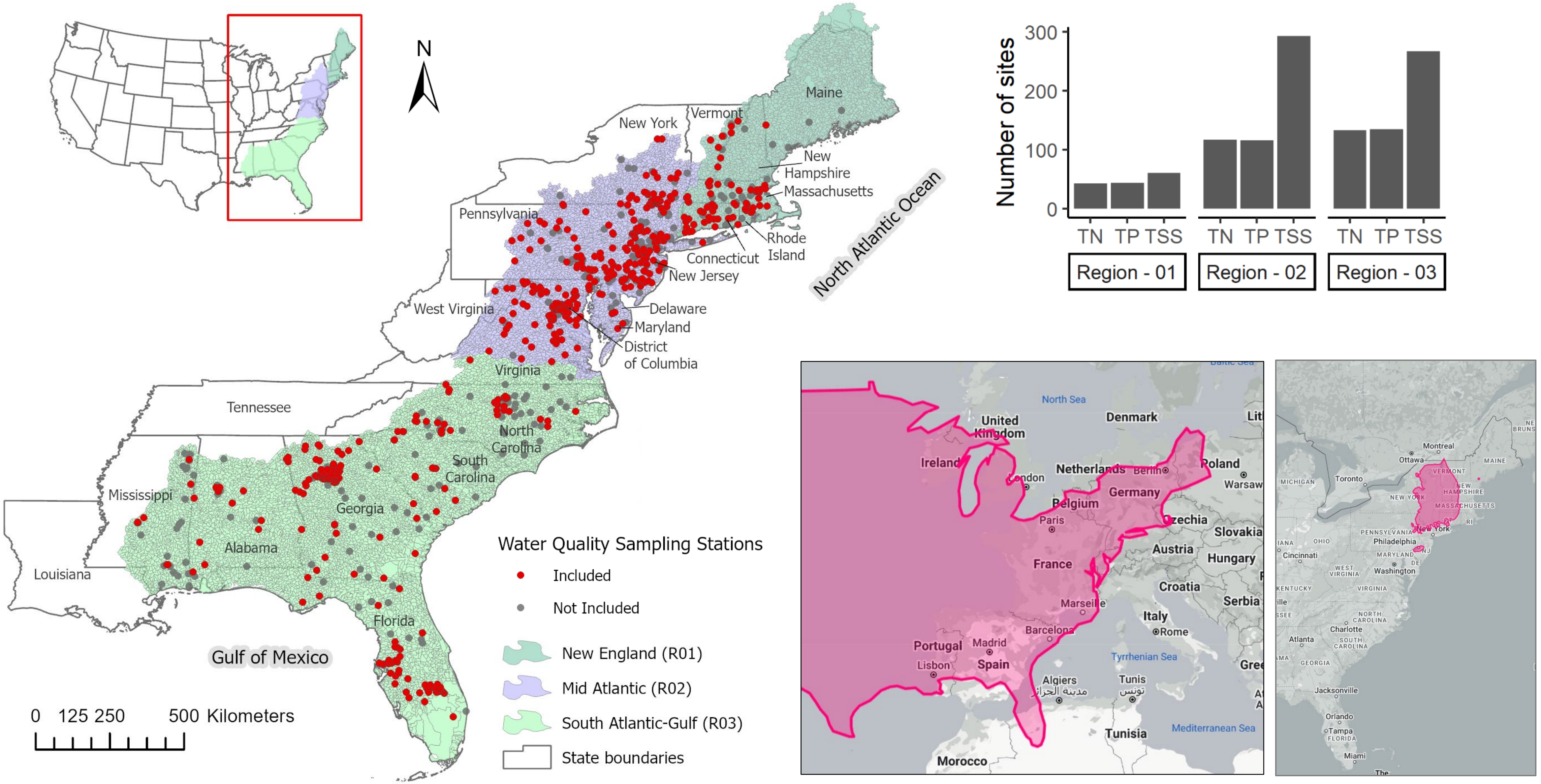


Region 02 - TSS

One region - One model

Daily WQ prediction at individual sites




N
Maine
Vermont
New York
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Pennsylvania
West Virginia
Delaware
Maryland
District of Columbia
Virginia
Tennessee
North Carolina
South Carolina
Georgia
Mississippi
Alabama
Louisiana
Florida
Gulf of Mexico
North Atlantic Ocean
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
0
100
200
300
TN
TP
TSS
Region - 01
Region - 02
Region - 03
North Sea
United Kingdom
Denmark
Ireland
Netherlands
Berlin
Poland
London
Warsaw
Belgium
Germany
Paris
Czechia
Slovakia
Austria
Hungary
France
Croatia
Serbia
Marseille
Italy
Barcelona
Rome
Portugal
Madrid
Spain
Lisbon
Tyrrhenian Sea
Algiers
مدينة الجزائر
Tunis
تونس
Tunisia
Mediterranean Sea
Morocco
Montreal
Ottawa
MAINE
MICHIGAN
Toronto
VERMONT
NEW HAMPSHIRE
NEW YORK
MASSACHUSETTS
New York
PENNSYLVANIA
Philadelphia
OHIO
Cincinnati
MARYLAND
Washington
WEST VIRGINIA
KENTUCKY
VIRGINIA
NORTH CAROLINA
Charlotte
SOUTH CAROLINA
Atlanta
GEORGIA
Jacksonville
Orlando
Tampa
FLORIDA
Miami

HAWQS

Hydrologic and Water Quality System

TEXAS A&M AGRILIFE RESEARCH

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

USDA Agricultural Research Service

27 predictors

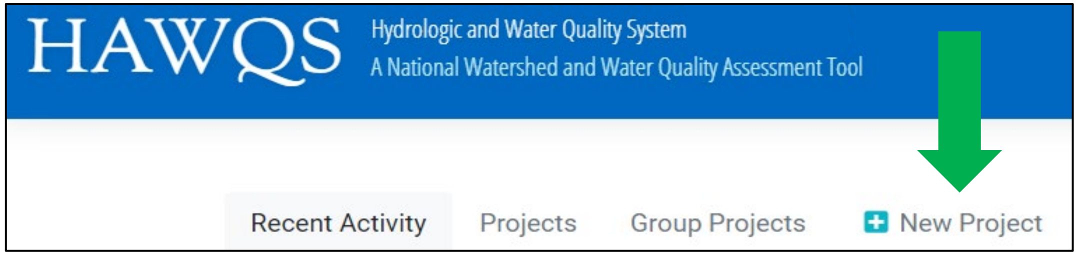
Dynamic- 5, Static- 22

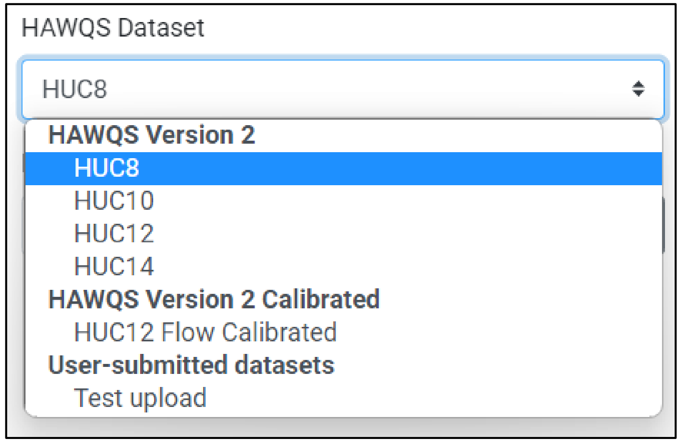
Response Variable: Concentration

Discharge	Seasonality	Water Cycle	Loading Potential	Channel	Soil	Landcover
ln(Q)	DOY	Mean PRCP	Slope	CH_L	HydA	Developed
ln(ΔQ)		PRCP/SRQ	USLE K	CH_W	HydB	Forest
ln(7dQ)		BFI	Sed_VHigh	CH_S	HydC	Wetland
ln(30dQ)			Sed_High	CH_D	HydD	Grassland/Shrub
			Sed_Mod			Managed veg.
			Sed_Low			
			Sed_VLow			
Dynamic		Static				

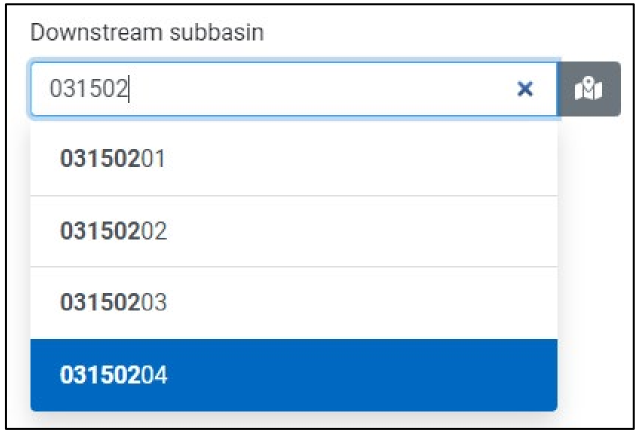
SWAT Model setup- just a few clicks away!!

hawqs.tamu.edu


HAWQS
Hydrologic and Water Quality System
A National Watershed and Water Quality Assessment Tool
Recent Activity
Projects
Group Projects
New Project

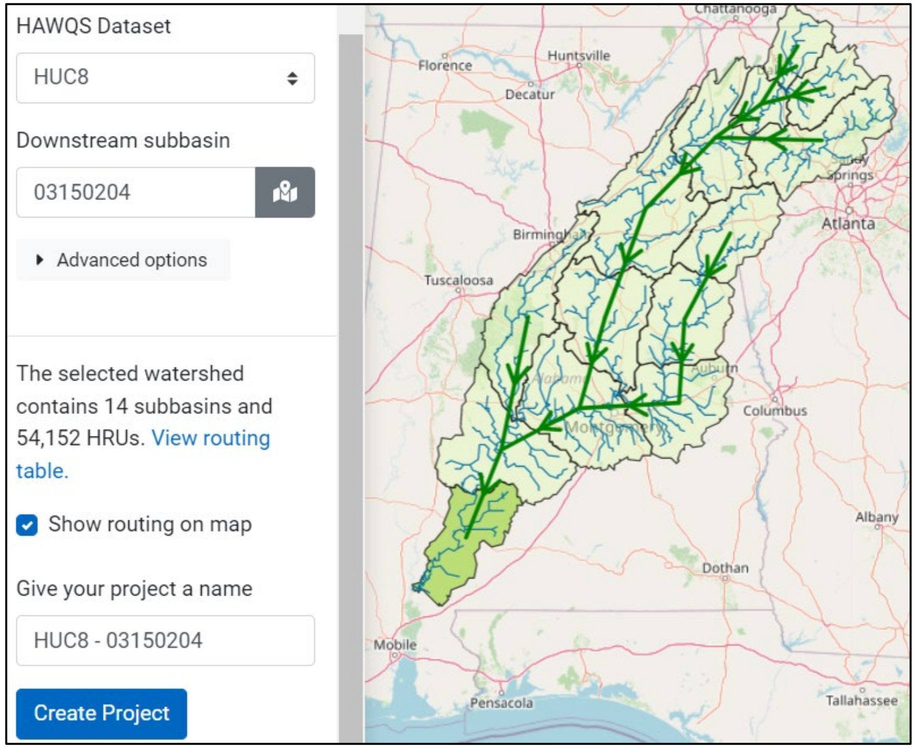

HAWQS Dataset

HUC8


Downstream subbasin

031502

HAWQS API


HAWQS Dataset

HUC8

Downstream subbasin

03150204

Advanced options

The selected watershed contains 14 subbasins and 54,152 HRUs. View routing table.

☑ Show routing on map

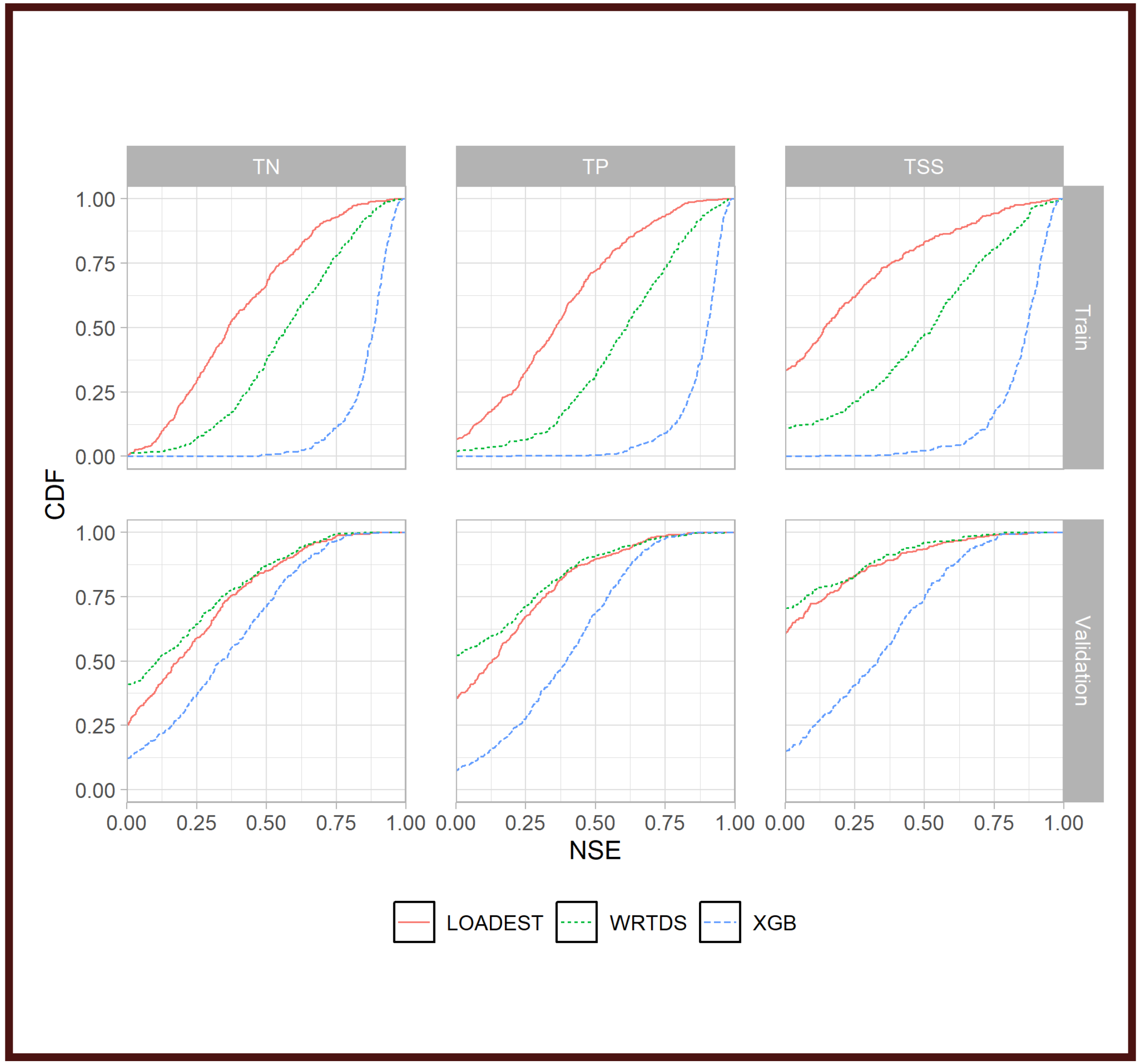
Give your project a name

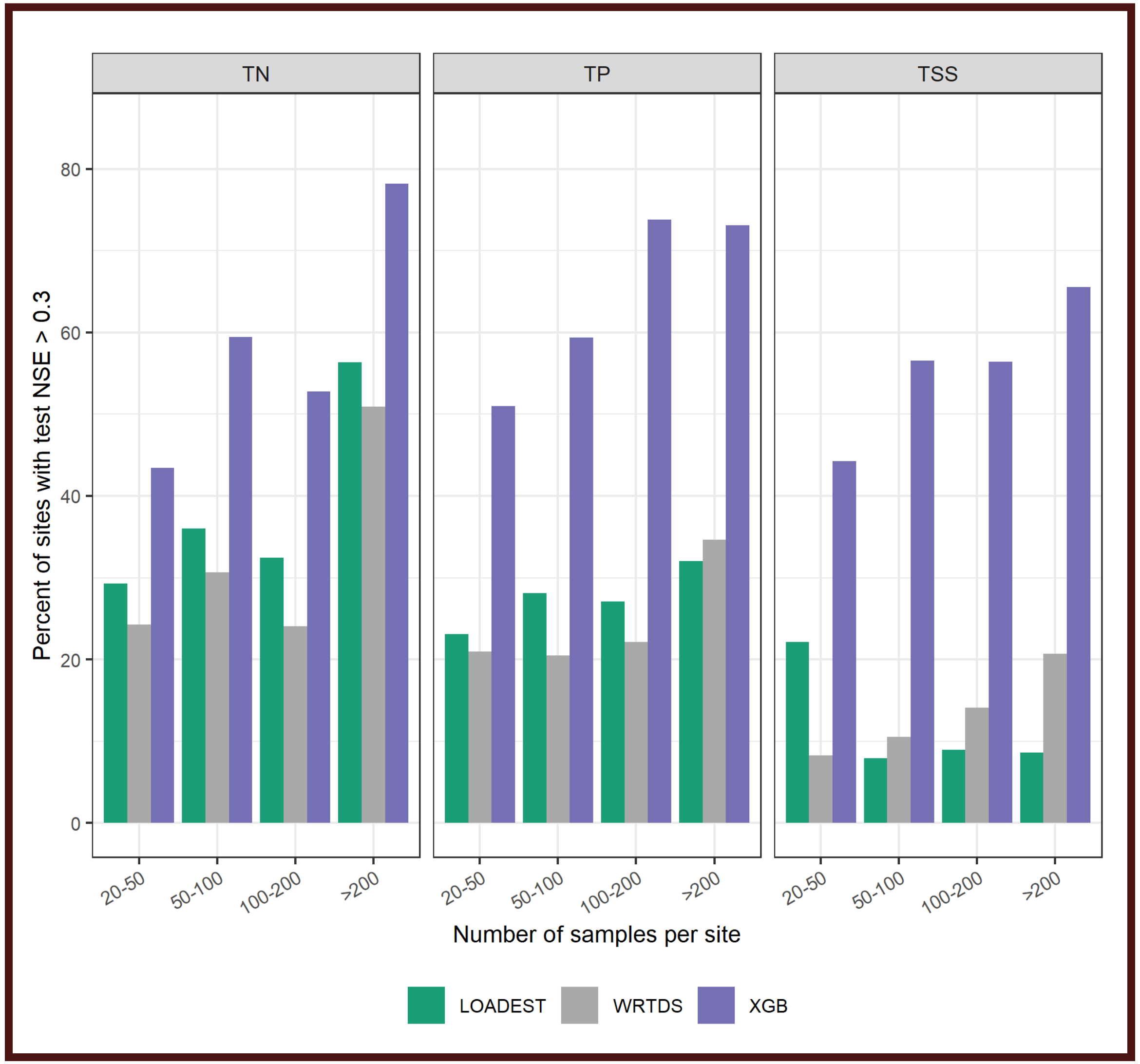
HUC8 - 03150204

INTERNATIONAL HAWQS PLATFORMS

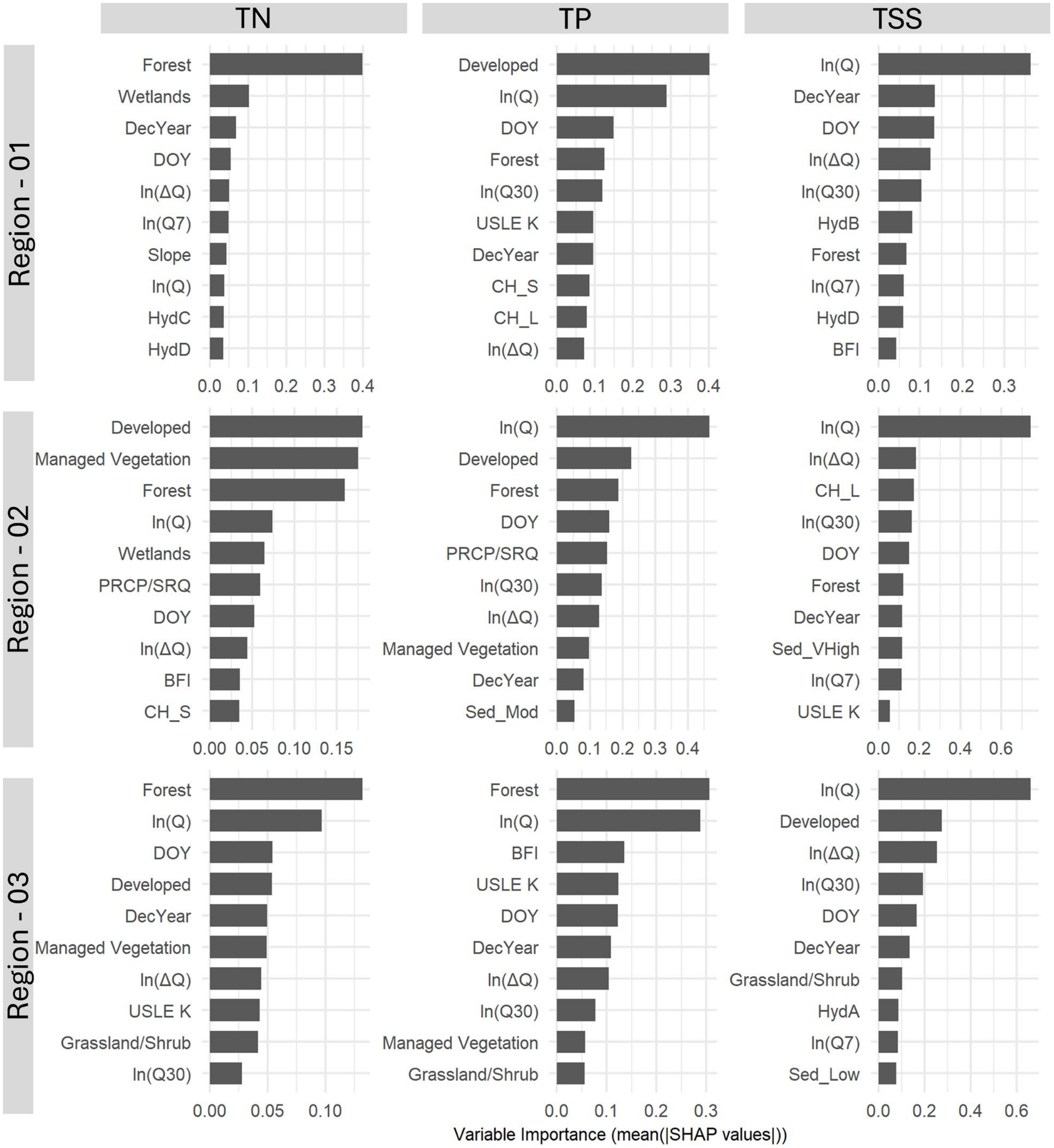
- South Africa (HAMSA)
  [hamsa.hawqs.tamu.edu]
- Pernambuco Brazil (SUPer)
  [super.hawqs.tamu.edu]
- Hydrologic Unit Model for InDia (HUMID)
  [bhuvan.nrcs.gov.in]
- Global HAWQS [India, Ukraine, Nepal]
  [global.hawqs.tamu.edu]
- Coming Soon: South Korea

XGB outperformed LOADEST and WRTDS



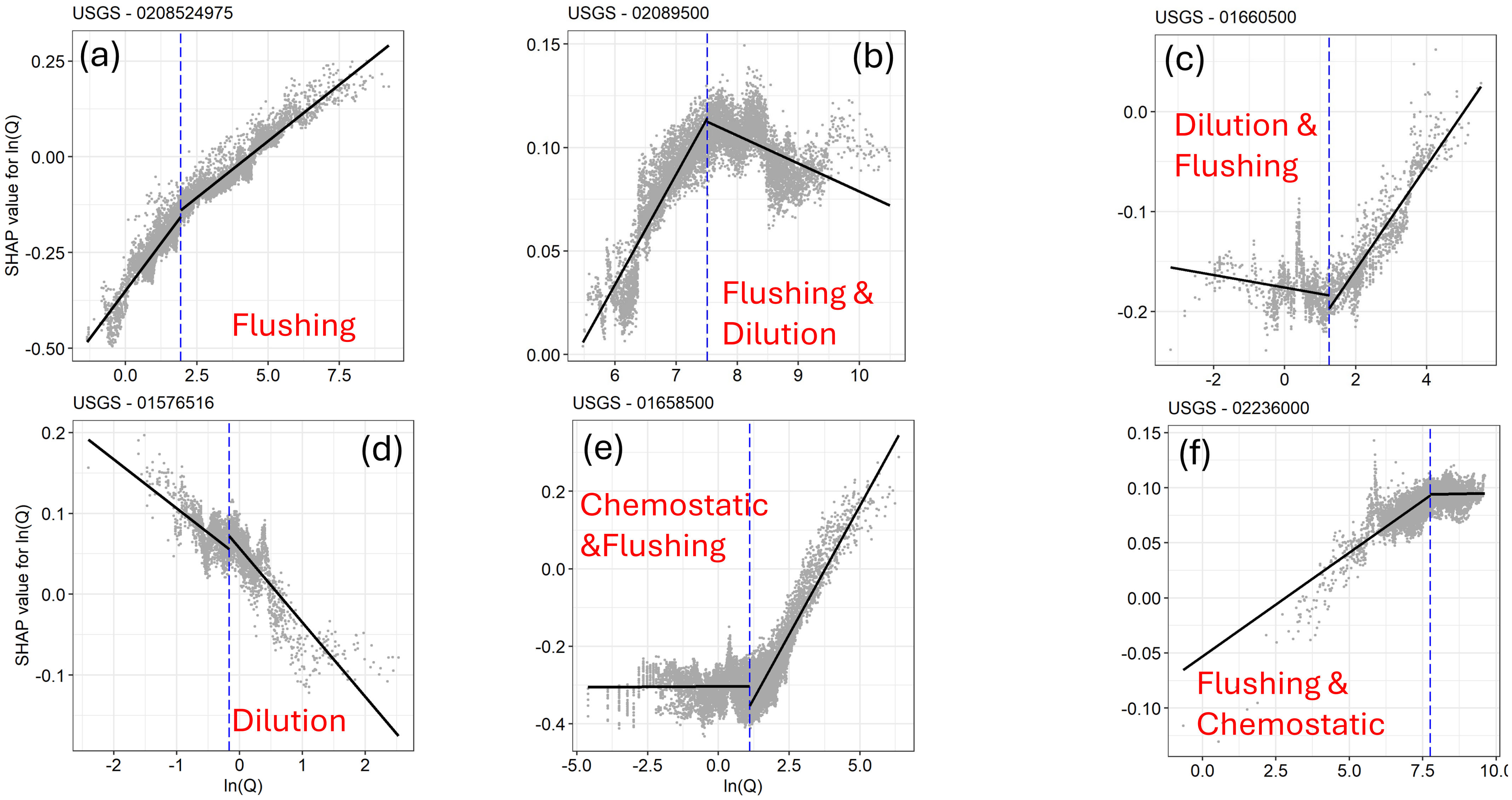


Watershed attributes played key role in WQ predictions



Six C-Q pattern: TN

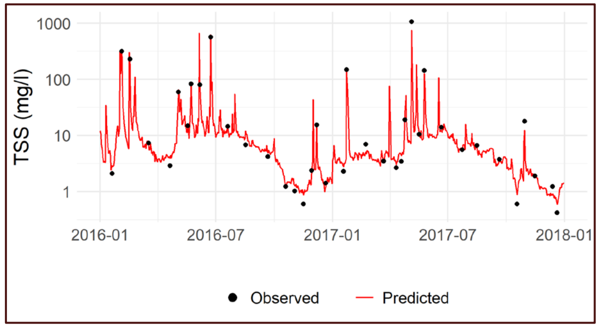
Flushing:
> 95% TSS & TP

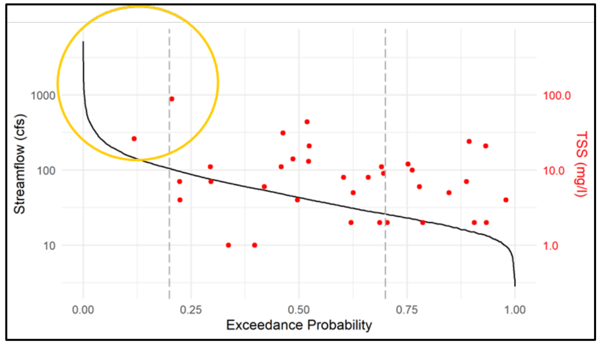


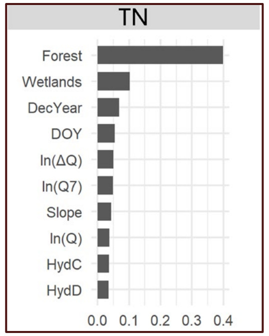
Key Takeaways

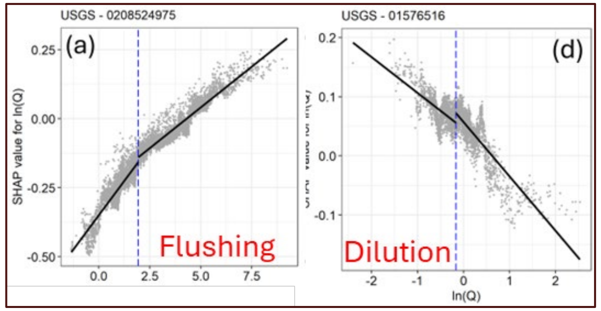
Water Research under review

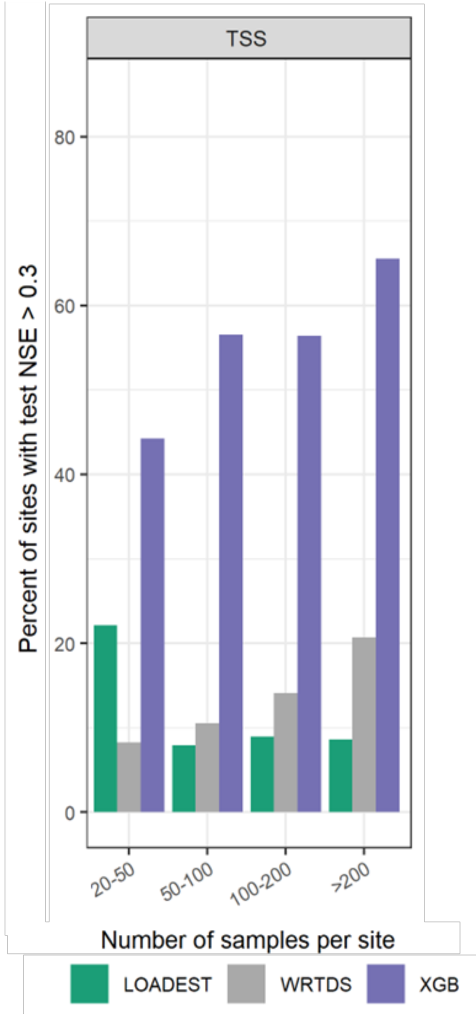
- ✓ New ML based WQ interpolation/extrapolation tool
- ✓ **XGB** model outperforms LOADEST and WRTDS
- ✓ Daily WQ estimates for US - **HAWQS**
- ✓ Combining WQ data across sites overcomes limitation of insufficient data at individual sites
- ✓ ML-WQ inferences using **Explainable AI**
  - aid in model interpretation
  - increasing trust in Black-Box model

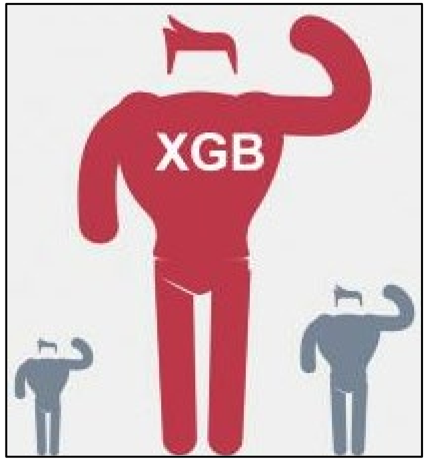












hawqs.tamu.edu

Thank you!

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

Texas A&M AgriLife Research