Hydrological Data Reanalysis via Uncertainty Quantification

A UQ Pipeline for Synthetic Ensemble Generation Supporting Red Tide (Karenia brevis) Bloom Prediction — Peace River, Florida

Student Authors: John Holik, Claiton Pinto, Christopher Foster, Jesse Prieto | Course Instructors: Dr. Fernando Gonzalez, Mr. Paul Allen, Dr. Anna Koufakou | Project Sponsor: Dr. Elshall

FGCU | U.A. Whitaker College of Engineering

Abstract

This study focuses on improving predictive modeling for harmful algal blooms, specifically red tide events in the Peace River region of Florida. The approach integrates observed environmental data with simulated model outputs to generate a reanalysis dataset using data assimilation techniques. By aligning and correcting model predictions with real-world observations, the system produces a more accurate and continuous representation of environmental conditions. Additionally, synthetic time-series data is generated to extend the dataset and capture long-term seasonal patterns, enabling more robust model training and evaluation. The resulting framework enhances the reliability of red tide prediction models and provides a scalable approach for environmental monitoring and decision support systems.

Methodology

Data — Daily discharge (CMS), total nitrogen, and total phosphorus from USGS Station 02296750 (Peace River at Arcadia, FL), 1999–2023.

Forward Model — The Watershed Assessment Model (WAM), calibrated 2014–2023 and validated 2000–2023, provides the physics-based hydrological simulation. An LSTM trained on WAM output serves as a surrogate forecast operator for ensemble propagation, as WAM produces a static trajectory rather than a callable dynamical model.

Uncertainty Quantification — Four methods generate 200-member ensembles (5 selected for visualization):

- Bootstrap — Resamples surrogate model residuals to produce empirical prediction intervals.
- GLUE — Retains behavioral parameter sets exceeding a likelihood threshold; spread reflects parametric uncertainty.
- EnKF — Sequentially assimilates USGS observations, propagating each member through a separate LSTM forward pass with per-member process noise.
- LPU — First-order Taylor expansion analytically propagates input/parameter uncertainties; assumes linearity and Gaussian error structure.

Bloom Prediction — Reanalysis products are passed individually (200 runs per method) to a Karenia brevis ML model, preserving the full uncertainty distribution.
Evaluation — Deterministic: NSE, KGE. Probabilistic: coverage probability and spread-skill ratio (target ≈ 1.0).

Technology Stack

Core Data	ML / UQ	Viz & Explore	Dev & Test
• pandas ≥ 2.2	• scikit-learn ≥ 1.7	• Matplotlib 3.9	• Jupyter NB 00–05
• numpy ≥ 2.0	• spotpy (GLUE)	• seaborn	• pytest + coverage
• scipy ≥ 1.13	• HydroErr metrics	• Plotly	• pyproject.toml

~2,400	6	4	6
Lines of Code	Notebooks	UQ Methods	Obs Stations

Package: red-tide-reanalysis (src-layout) | Serialization: joblib | pytest overall coverage: 87% ✓
Domain target: Total Nitrogen (TN) & Total Phosphorus (TP) at Peace River Arcadia station (02296750)

Project Phases

1 Phase 1 — Data Preparation & Cleaning

Standardize all 6 observation CSVs: uniform schema, column names, datetime format, units (mg/L). Remove duplicates; flag/handle non-detects and negatives; remove implausible outliers with documented thresholds.

Load Arcadia WAM (TN & TP) and USGS daily discharge; align to observation dates. Merge all six datasets into a single long-format DataFrame indexed by date and station ID. Produce station timeline & overlap window catalog — for each station pair, record exact date ranges and concurrent observation counts.

2 Phase 2 — Normalize & Merge Observations

Extract paired concurrent obs during overlap windows. Run scatter/correlation analysis (flow-dependence + seasonality) to select normalization per station. 5 methods implemented:
① Linear Regression Bias Correction ② Quantile Mapping (CDF matching) ③ Z-Score (variance-mean matching) ④ Flow-Dependent Bias Correction ⑤ Hierarchical Priority Selection (naive baseline)

Validate on held-out overlap data; select best method per station. Resolve temporal overlaps → single TN & TP series, one value per timestep, spanning 2000–2023.

3 Phase 3 — Ensemble Reanalysis

Use Arcadia WAM as model backbone; merged Phase 2 series as assimilation target. Generate ensemble perturbations of the WAM trajectory. Apply Bootstrap, GLUE, and EnKF to produce TN & TP ensemble reanalysis products.

Inflate obs. error variance for non-Arcadia timesteps (transfer uncertainty). Evaluate with NSE, KGE, coverage probability, and spread-skill ratio. Feed each ensemble member individually into the Red Tide prediction model.

TN Ensemble — Bootstrap CI Band

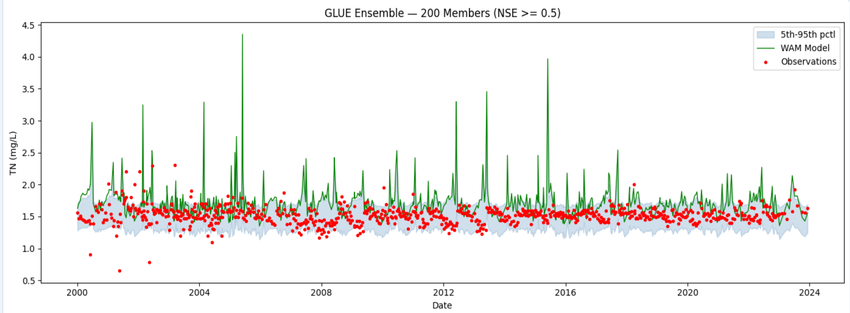

Fig 1. Bootstrap 5th–95th percentile CI (blue band) for TN at Arcadia (02296750), 1999–2024. WAM model (green) = continuous baseline; red dots = sparse observations.

TN Ensemble — EnKF CI Band

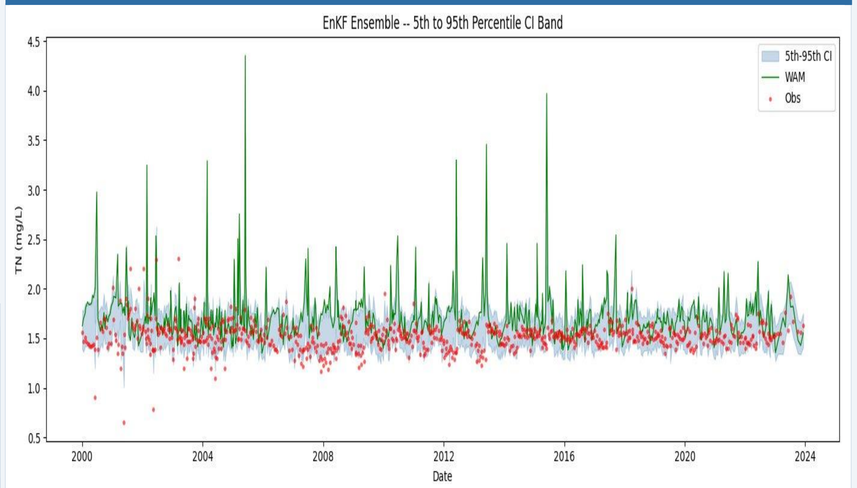

Fig 2. EnKF 5th–95th percentile CI — tighter band reflects sequential state updates. Narrower vs. Bootstrap; better tracks regime shifts post-2014.

Results: ML Bloom Classification — TN Ensembles

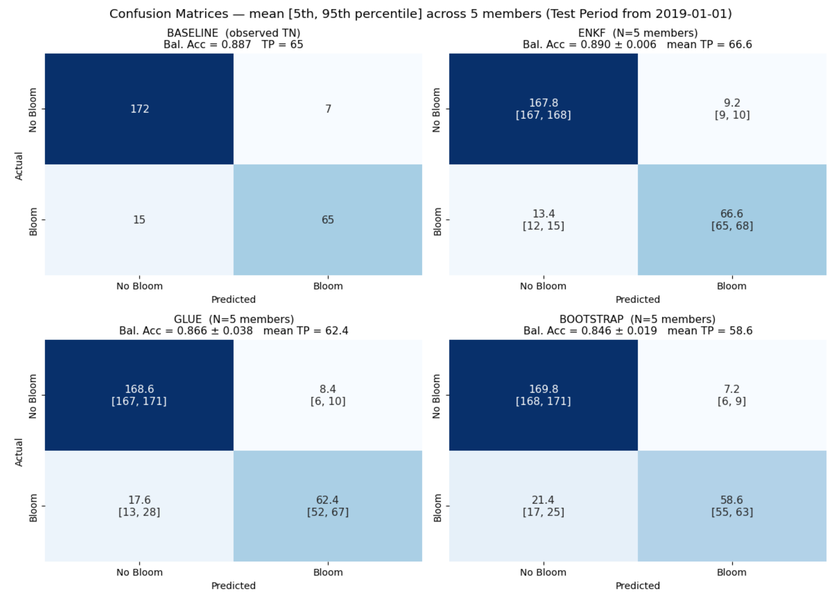

Fig. 3. Confusion matrices comparing ML bloom classification across all UQ methods against the observed TN baseline (Bal. Acc = 0.887, test period 2019–2024). EnKF best matches the baseline (Bal. Acc = 0.890 ± 0.006, mean TP = 66.6), with GLUE and Bootstrap showing progressively more false negatives.

Results: Discharge Ensembles (All 4 UQ Methods)

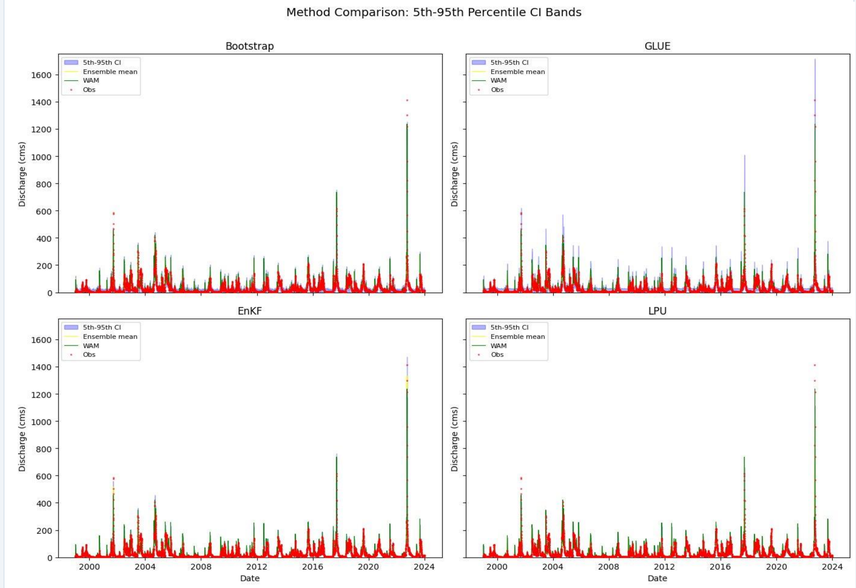

Fig 4. 5th–95th percentile CI bands for Discharge (cms) at Arcadia, 1999–2024. Bootstrap & EnKF closely track observations; GLUE shows widest spread; LPU narrowest analytical CI. All methods capture peak flood events.

Results: ML Bloom Classification (Test Period 2019–2024)

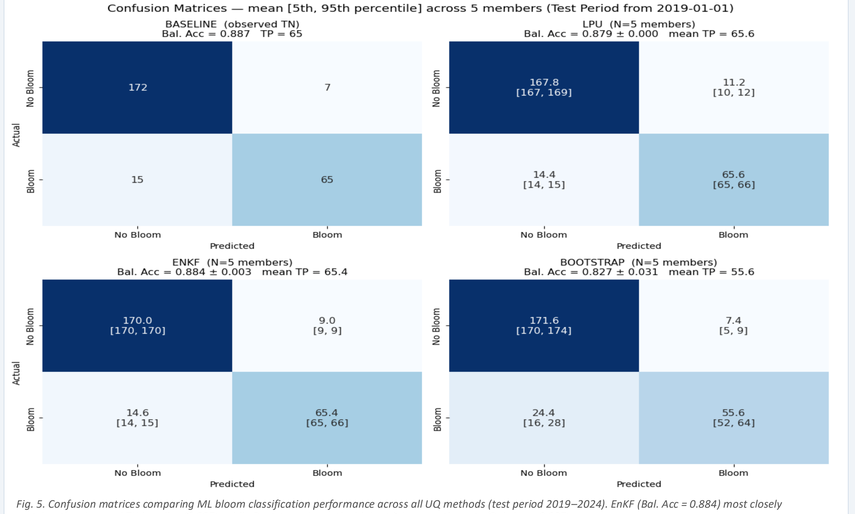

Fig. 5. Confusion matrices comparing ML bloom classification performance across all UQ methods (test period 2019–2024). EnKF (Bal. Acc = 0.884) most closely matches the observed TN baseline (0.887), while Bootstrap shows the highest false-negative rate (Bal. Acc = 0.827).

Results: ML Bloom Classificatio (Joint Ensembles)