ENABLING A NOVEL LOW-COST SALINITY SENSOR THROUGH MACHINE LEARNING

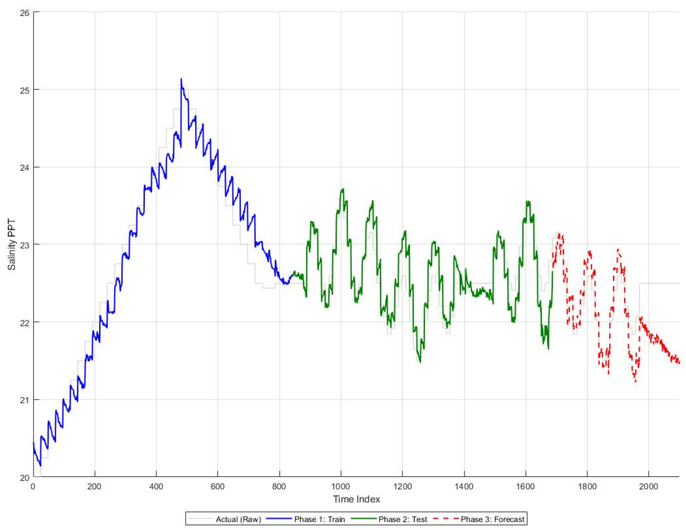
STUDENTS: WILLIAM BEI, AUGUST DUKICH, MANISH GATTI, RITHIKESH MUDDANA

Problem Statment

Exploration of machine-learning frameworks that correct for long-term drift in low-cost salinity sensors by analyzing time-series data, which includes features such as voltage, impedance, and temperature. Our models aim to automatically separate true environmental variations from apparent changes caused by sensor drift.

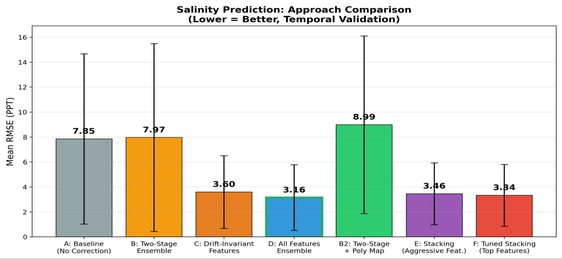
Framework 1: SSA-based Ridge Regression

- **Premise**: Assumes that drift is both time-dependent and not prone to sudden changes in trends
- Uses components derived from Single-Spectrum Analysis (SSA) for training model
- 40/40/20 Train/Test/Forecast Split; RMSE values of 0.263, 0.248, 0.504 respectively
- **Key Outcome**: Model shows initial promise but tends to overcompensate for the drift components, hybrid methods may appropriately correct overcompensation



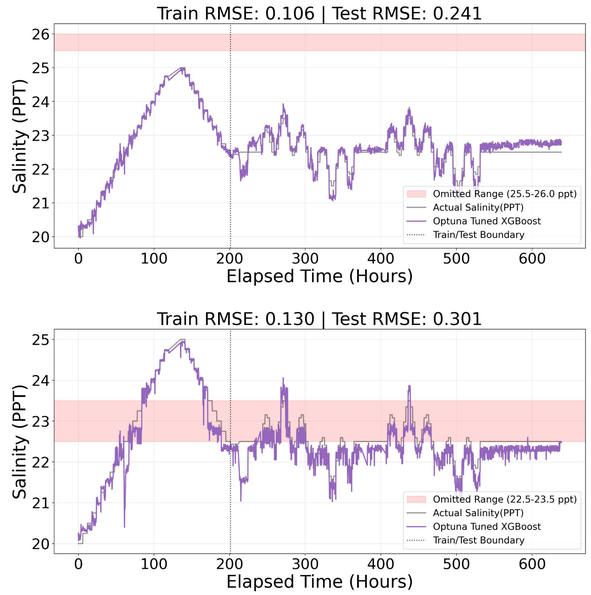
Framework 2: Ensemble

- **Premise**: Explores multi-model architectures, advancing from uncorrected linear baselines to two-stage drift prediction using impedance and temperature.
- **Feature Engineering**: Utilizes PCA-derived drift-invariant ratios (V0/R, V0/Ws, R/Ws) and delta features to capture physical signal degradation.
- **Key Outcome**: Peak performance achieved via a stacking ensemble (Gradient Boosting, Random Forest, Extra Trees, Ridge) that combined raw signals with the engineered features.



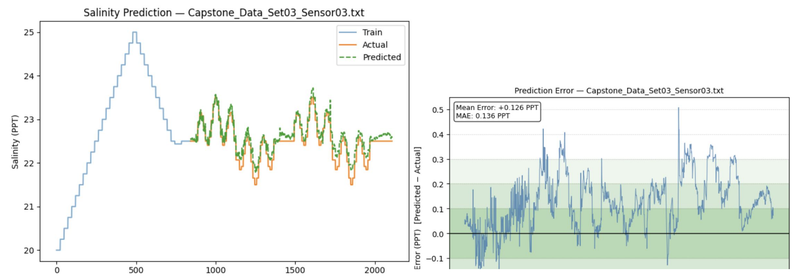
Framework 3: XGBoost

- **Premise**: XGBoost for non-linear mapping of membrane impedance to salinity
- **Testing & Optimization**: Out-of-Distribution omissions to map extrapolation boundaries, mitigated via Optuna Bayesian tuning for hardware-agnostic generalization
- **Key Outcome**: Minimized test RMSE and stabilized variance across noisy sensor profiles



Framework 4: Two-Stage Modeling

- **Premise**: Two-stage hybrid architecture designed to decouple the baseline environmental salinity pattern from complex hardware drift.
- **Methodology**: Applies a Ridge regression baseline on raw voltage and temperature, followed by a Gradient Boosting model trained on impedance features to predict and correct the residual errors.
- **Key Outcome**: The summed outputs successfully achieved a 0.14 PPT error, making it the highest performing framework against the strict 0.1 PPT target.



Summary and Future Work

- **Summary**: While all frameworks demonstrated merit, Two-Stage Modeling proved superior. By isolating baseline physical drift from residual errors, it achieved a 0.14 PPT error, closest to our 0.1 PPT target.
- **Future work**:
  - Exploration of how models fare in long-term deployment
    - Deep dive into training/testing schedules versus model robustness
    - Find correlation between length of model training vs. amount of time model can function before diverging too much
  - Explore cross-sensor training
    - Create a robust generalized base model that can be fine-tuned to specific sensors, as opposed to training entirely on each sensor
  - Combine multiple techniques in different hybrid approaches
    - Possibly utilize more frequency-analysis based methods in future frameworks

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

INDUSTRY ADVISERS: John Kucewicz, Anuscheh Nawaz FACULTY ADVISOR: M. P. Anantram

SPONSOR: UW Applied Physics Laboratory