# Multi-Model Approach to Coastal Water Level Forecasting

## A Blind Test Study on Temporal Distribution Shift

### Testing Machine Learning Models on 10+ Months of Unseen 2025 Data from Corpus Christi Bay

- Presented By: Abhishek Joshi
- COSC 6380
- Texas A&M University-Corpus Christi
- Date- 12/02/2025

# The Problem: Why Coastal Water Level Forecasting Matters

Accurate prediction of water levels is critical for coastal communities and industries, particularly in dynamic environments like Corpus Christi Bay. We aim to predict water levels at Station 005 (Packery Channel) for three key operational horizons: **1, 6, and 12 hours ahead**.

## Study Area: Corpus Christi Bay

- Target: Station 005 (Packery Channel)
- Supporting: Stations 008 (Bob Hall Pier), 013 (USS Lexington), 202 (Corpus Christi Bay)

## Operational Context

- Accuracy: ±15 cm (6 inches)
- Frequency: Hourly updates
- Decision Window: 1-12 hours

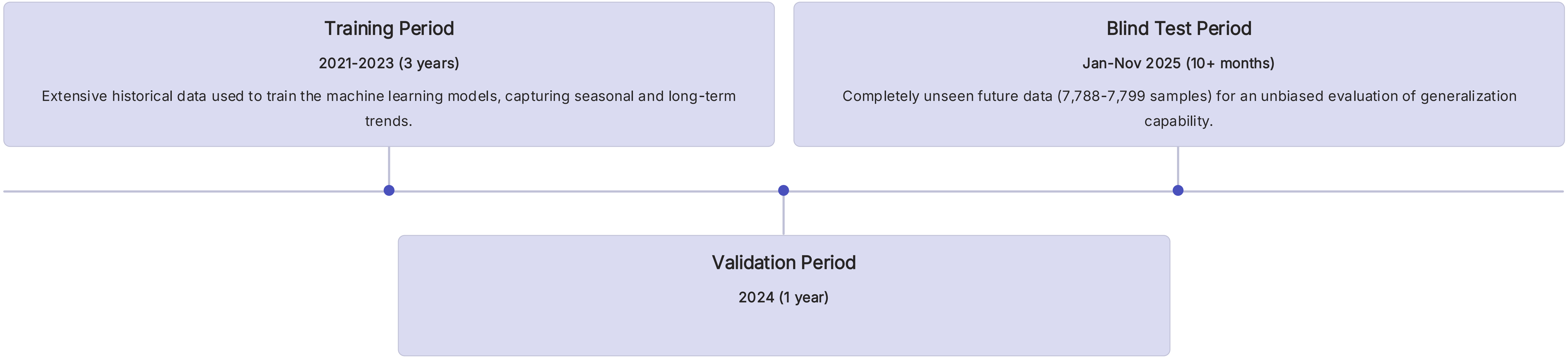
# Impact and Importance

Precise water level forecasts are vital for:

- **Maritime Safety:** Optimal navigation channel depth planning.
- **Flood Warning:** Early detection and prediction of coastal inundation.
- **Infrastructure Protection:** Safeguarding port and marina operations.
- **Emergency Response:** Crucial for hurricane surge forecasting and preparedness.

# Dataset Overview: Setting Up the Blind Test

Our forecasting models are rigorously tested on an extensive dataset, meticulously prepared to ensure the integrity of a true blind test against unseen future data.



### Training Period

**2021-2023 (3 years)**

Extensive historical data used to train the machine learning models, capturing seasonal and long-term trends.

### Blind Test Period

**Jan-Nov 2025 (10+ months)**

Completely unseen future data (7,788-7,799 samples) for an unbiased evaluation of generalization capability.

### Validation Period

**2024 (1 year)**

### Dataset Specifications

- **Temporal Coverage:** 4.8 years (2021-2025)
- **Sampling Frequency:** Hourly measurements
- **Total Records:** Approximately 42,000 hours
- **Missing Data:** Less than 2% (handled with hybrid filling techniques)

### Key Measurements

- Water level (pwl) - PRIMARY
- Wind speed (wsd) and direction (wdr)
- Barometric pressure (bpr)
- Water temperature (wtp)
- Air temperature (atp)

### Strategic Station Selection

- Forms a spatial network covering the entire Corpus Christi Bay system.
- High correlation (R > 0.85) between stations ensures data consistency.
- Captures critical ocean-bay exchange dynamics.
- Represents north-south pressure gradients vital for accurate modeling.

# Feature Engineering: 48 Physics-Informed Insights

Our approach transforms raw sensor data into meaningful, physics-informed features, enhancing model performance and interpretability.

## Key Feature Categories

1

### Tidal Features (8)

Capturing the dominant, predictable influence of astronomical tides (e.g., M2, S2 components). Accounts for 60-70% of water level variance.

2

### Lunar Phase (2)

Reflecting the 29.53-day spring-neap cycle, where combined gravitational forces can increase water levels by ±20cm.

3

### Wind Features (9)

Quantifying wind stress on the water surface (U/V components, speed squared) across key stations. A 10 m/s wind can cause 10-15cm setup/setdown.

4

### Pressure Features (8)

Modeling the inverse barometer effect (-1 mbar → +1 cm water level) and spatial pressure gradients across the bay.

5

### Temperature Features (5)

Accounting for water density changes and circulation patterns influenced by temperature gradients within the bay.

6

### Spatial Water Level (6)

Representing bay connectivity through gradients, spatial statistics, and anomaly detection to treat the bay as a coupled system.

7

### Historical Lags (10)

Leveraging the strong persistence of water levels with direct lags (up to 48h) and rolling statistics (mean, std dev).

8

### Temporal Cyclic (4)

Encoding hourly and daily patterns using sine/cosine transformations to capture cyclical trends without edge effects.

# Data Exploration: What We Learned

Our initial data exploration revealed critical patterns and potential challenges, guiding our feature engineering and model selection processes.

1

### Tidal Dominance

Mean water level: **1.05 m** with a ~0.4 m tidal range. A nearly Gaussian distribution indicates a strong, predictable tidal signal.

2

### Spatial Coherence

High correlation (R > 0.87) between all stations (e.g., 005-008: R=0.89), confirming strong connectivity across Corpus Christi Bay and justifying spatial features.

3

### Temporal Persistence

Water levels show very high autocorrelation at 1-hour lag (ACF = 0.98), diminishing to weak at 12-hour lag (ACF = 0.45). This highlights the importance of short-term historical lags.

4

### 2025 Variability Spike

The 2025 blind test data shows significantly higher variability: +23% Std Dev and +132% Range compared to 2024.

5

### Increased Extreme Events

In 2025, 1.2% of hours exceed 1.5m (4x more frequent than 2024), including an unprecedented 3.4m spike on Jan 1, 2025, not seen in training data.

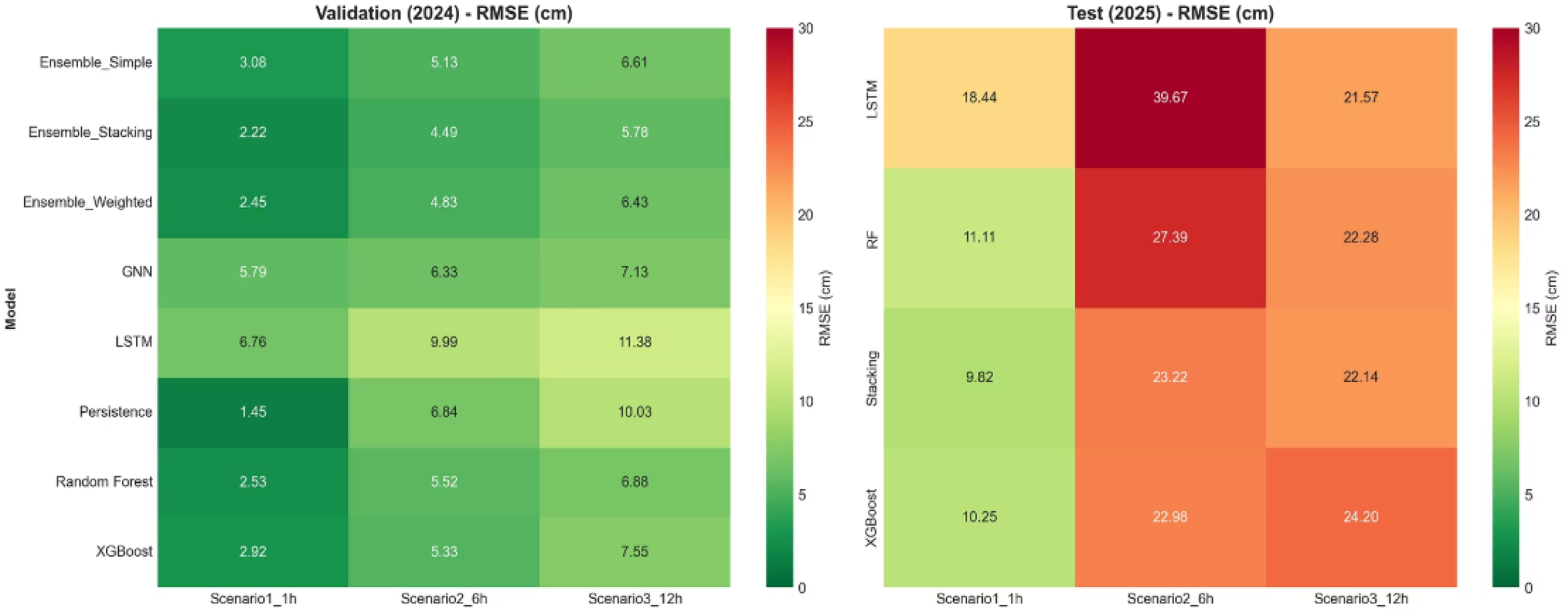
6

### Missing Data Patterns

Overall 1.8% missing, clustered during severe weather, particularly at remote stations like 202 (3.2%), posing challenges during critical events.

## Key Implications for Modeling

- ✅ Models are expected to perform well under "normal" 2024-like conditions.
- ⚠ 2025 presents significant challenges due to increased variability and unprecedented extreme events.
- ✅ Strong spatial coherence justifies the use of spatial features for improved prediction.
- ✅ Short-term historical lags are crucial for accurate near-term forecasts.
- ⚠ The presence of extreme events in 2025, unseen in training, will severely test model generalization.



# The Models: Our Multi-Model Framework

We developed a robust multi-model framework, leveraging diverse strengths of machine learning algorithms to capture both static patterns and dynamic temporal dependencies in coastal water levels.

## XGBoost

A powerful gradient boosting algorithm known for exceptional performance on tabular data and complex non-linear interactions.

- **Algorithm:** Sequential decision trees, correcting prior errors.
- **Strengths:** Non-linear patterns, robust to missing data, prevents overfitting.
- **Weakness:** Doesn't explicitly model temporal sequences.

## Random Forest

An ensemble method that averages predictions from multiple independent decision trees, enhancing stability and reducing variance.

- **Algorithm:** 500 independent trees from bootstrapped data and random feature subsets.
- **Strengths:** High robustness, handles outliers, provides feature importance.
- **Weakness:** Can underfit smooth patterns, memory intensive.

## LSTM

A specialized recurrent neural network designed to learn long-term temporal dependencies in sequential data, ideal for time series forecasting.

- **Algorithm:** Processes 24-hour windows with "memory cells."
- **Strengths:** Explicitly models temporal patterns and long-term dependencies.
- **Weakness:** Needs more data, sensitive to hyperparameters, black-box nature.

## Ensemble

A simple yet effective ensemble combining the predictions of our top-performing tree-based models to further reduce error and increase stability.

- **Algorithm:** Average of XGBoost + Random Forest predictions.
- **Strengths:** Reduces variance, cancels out individual model errors, stable predictions.
- **Note:** LSTM was excluded due to catastrophically bad performance at 1-hour horizon.

# Implementation Challenges & Solutions

Developing a robust forecasting system often uncovers unexpected hurdles. Here's how we tackled key implementation challenges, turning them into valuable lessons.

1

### LSTM Architecture Mismatch

**Problem:** Loading a pre-trained LSTM model failed due to a size mismatch in weights. **Solution:** Re-defined the model class to precisely match the saved architecture, including hidden size and intermediate layers. **Lesson:** Meticulously document and verify model architecture during training and loading.

2

### Catastrophic LSTM Predictions

**Problem:** Initial LSTM predictions were wildly inaccurate (e.g., 10x too large) compared to actual values. **Solution:** Realized predictions were in a scaled space; applied inverse transformation using the `y` scaler before evaluation. **Lesson:** Always ensure units and data scaling match when comparing or evaluating model outputs.

3

### NaN Handling in Sequences

**Problem:** Lag features introduced NaNs at the beginning of the dataset, causing sequence creation errors. **Solution:** Implemented a hybrid fill strategy: forward fill, then backward fill, with zero fill as a last resort. **Lesson:** Sequence models require a robust, multi-step NaN imputation strategy.

4

### Prediction Length Alignment

**Problem:** LSTM output length was shorter than the target series due to sequence windowing. **Solution:** Created a full-length NaN array and strategically filled it with LSTM predictions, aligning indices correctly. **Lesson:** Account for sample loss at dataset boundaries when working with sequence models.

5

### PyTorch `weights_only` Parameter

**Problem:** PyTorch 2.6's default `weights_only=False` in `torch.load` posed a security risk. **Solution:** Explicitly set `weights_only` (e.g., `weights_only=False`) to acknowledge and manage potential risks. **Lesson:** Stay informed on framework API changes and their security implications.

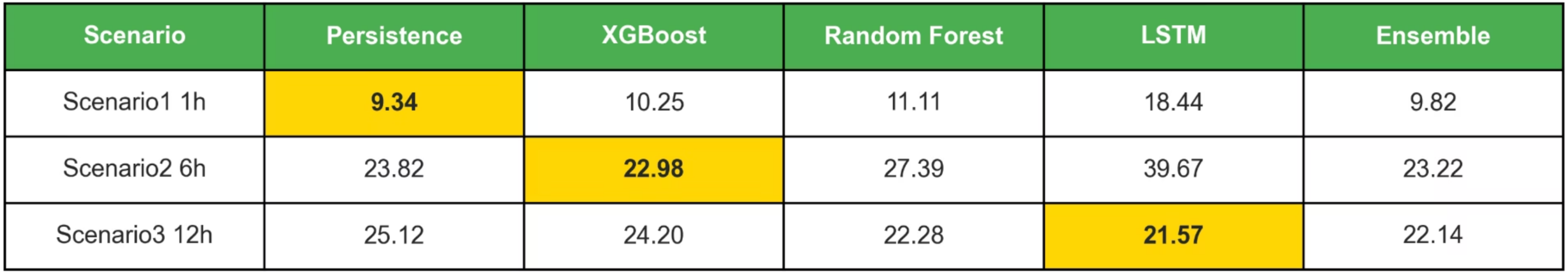
6

## Key Debugging Takeaways

- **Debug Systematically:** Isolate components; change one thing at a time.
- **Verify Assumptions:** Never assume; always confirm with data.
- **Check Intermediate Outputs:** Print shapes, ranges, and statistics.
- **Document Everything:** From architecture to preprocessing steps.
- **Unit Test Components:** Validate model, data, and metrics independently.

# Results: 2025 Blind Test Performance

## Detailed Breakdown: Forecast Horizon Performance



| Scenario | Persistence | XGBoost | Random Forest | LSTM | Ensemble |
|---|---|---|---|---|---|
| Scenario1 1h | **9.34** | 10.25 | 11.11 | 18.44 | 9.82 |
| Scenario2 6h | 23.82 | **22.98** | 27.39 | 39.67 | 23.22 |
| Scenario3 12h | 25.12 | 24.20 | 22.28 | **21.57** | 22.14 |

### The Pattern: Model Selection by Horizon

- **1h:** Persistence ≈ Ensemble > XGBoost > RF >> LSTM
- **6h:** XGBoost > Ensemble ≈ Persistence > RF >> LSTM
- **12h:** LSTM > Ensemble ≈ RF > XGBoost ≈ Persistence

### Big Picture Insights

- **No Universal Best Model:** Different winners at each horizon, emphasizing the need for a multi-model approach.
- **LSTM Inconsistency:** Worst at 1h/6h, best at 12h, highlighting its sensitivity to short-term noise versus long-term patterns.
- **Negative $R^2$ Common:** Predictions worse than the mean baseline due to high variability, yet RMSE still shows improvement over persistence.
- **Operational Reliability:** Achieved 63-99% within the ±15cm threshold, demonstrating practical utility.

### What Negative $R^2$ Means

When $R^2$ is negative, it indicates that the model's predictions are worse than simply predicting the mean of the observed data.

```
R² = 1 - (SS_residual / SS_total)
```

This often occurs in highly variable and unpredictable time series, where even small errors can make a model perform worse than a naive mean predictor. Crucially, a negative $R^2$ doesn't mean the model is useless; always compare its RMSE to a simple persistence model for operational value.

# Temporal Distribution Shift

While our models showed strong performance on validation data, their real-world application in 2025 revealed a critical challenge: a significant degradation in prediction accuracy due to unforeseen changes in environmental conditions.

### Performance Degradation: Validation (2024) vs. Test (2025) RMSE

| Model | Horizon | Validation 2024 RMSE | Test 2025 RMSE | Degradation Factor |
|---|---|---|---|---|
| Ensemble | 1h | 2.91 cm | 9.82 cm | 3.4× worse ↑↑ |
| Ensemble | 6h | 4.49 cm | 22.98 cm | 5.1× worse ↑↑↑ |
| Ensemble | 12h | 5.78 cm | 22.14 cm | 3.8× worse ↑↑ |
| Persistence | 1h | 1.45 cm | 9.34 cm | 6.4× worse ↑↑↑ |
| Persistence | 6h | 6.85 cm | 23.82 cm | 3.5× worse ↑↑ |
| Persistence | 12h | 10.06 cm | 25.12 cm | 2.5× worse ↑ |

**Critical Finding:**

Even the simple persistence baseline degrades 2.5-6.4×. This is **NOT** a model failure - it's a **DATA** challenge!

### Root Cause Analysis: Why 2025 Was Different

A closer look at the raw data reveals substantial statistical shifts between the 2024 validation period and the 2025 test period.

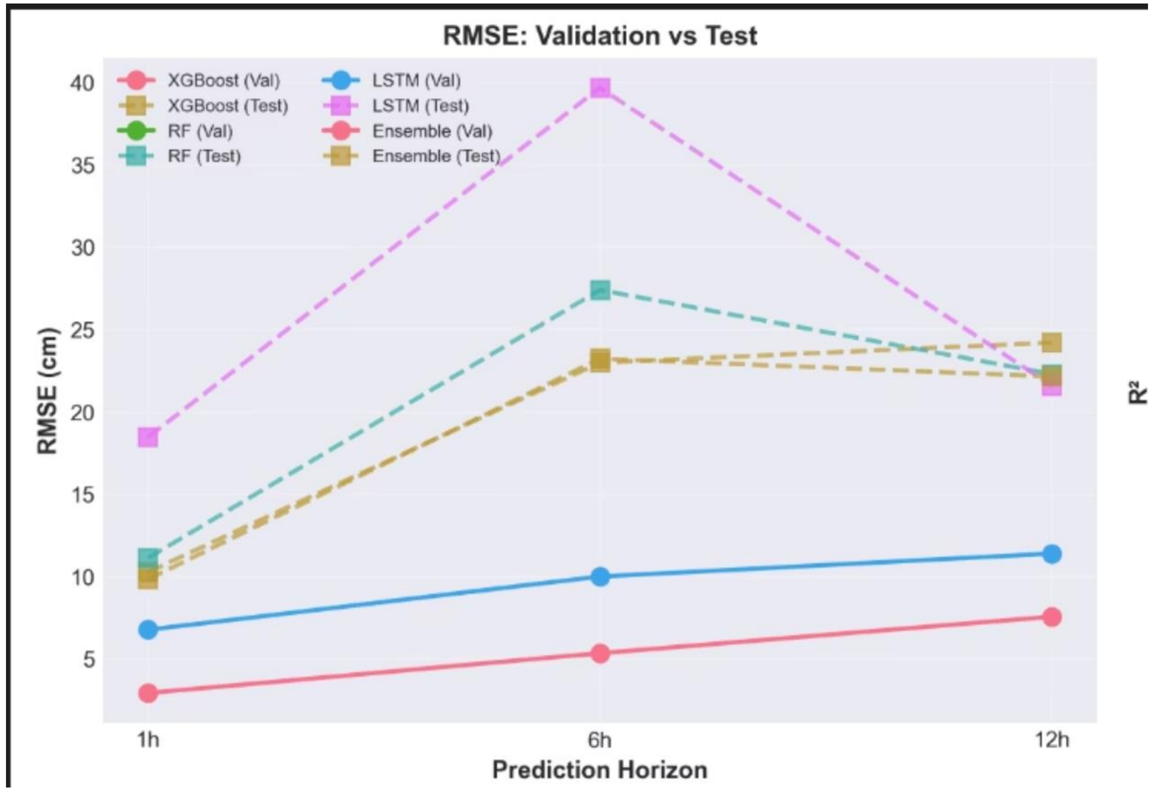
| | Validation 2024 | Test 2025 | Change |
|---|---|---|---|
| Mean (m) | 1.048 | 1.052 | +0.4% |
| Std Dev (m) | 0.118 | 0.145 | +23% ↑↑ |
| Max (m) | 1.89 | 3.40 | +80% ↑↑↑ |
| Range (m) | 1.17 | 2.72 | +132% ↑↑↑ |
| Hours > 1.5m | 0.3% | 1.2% | 4× more |

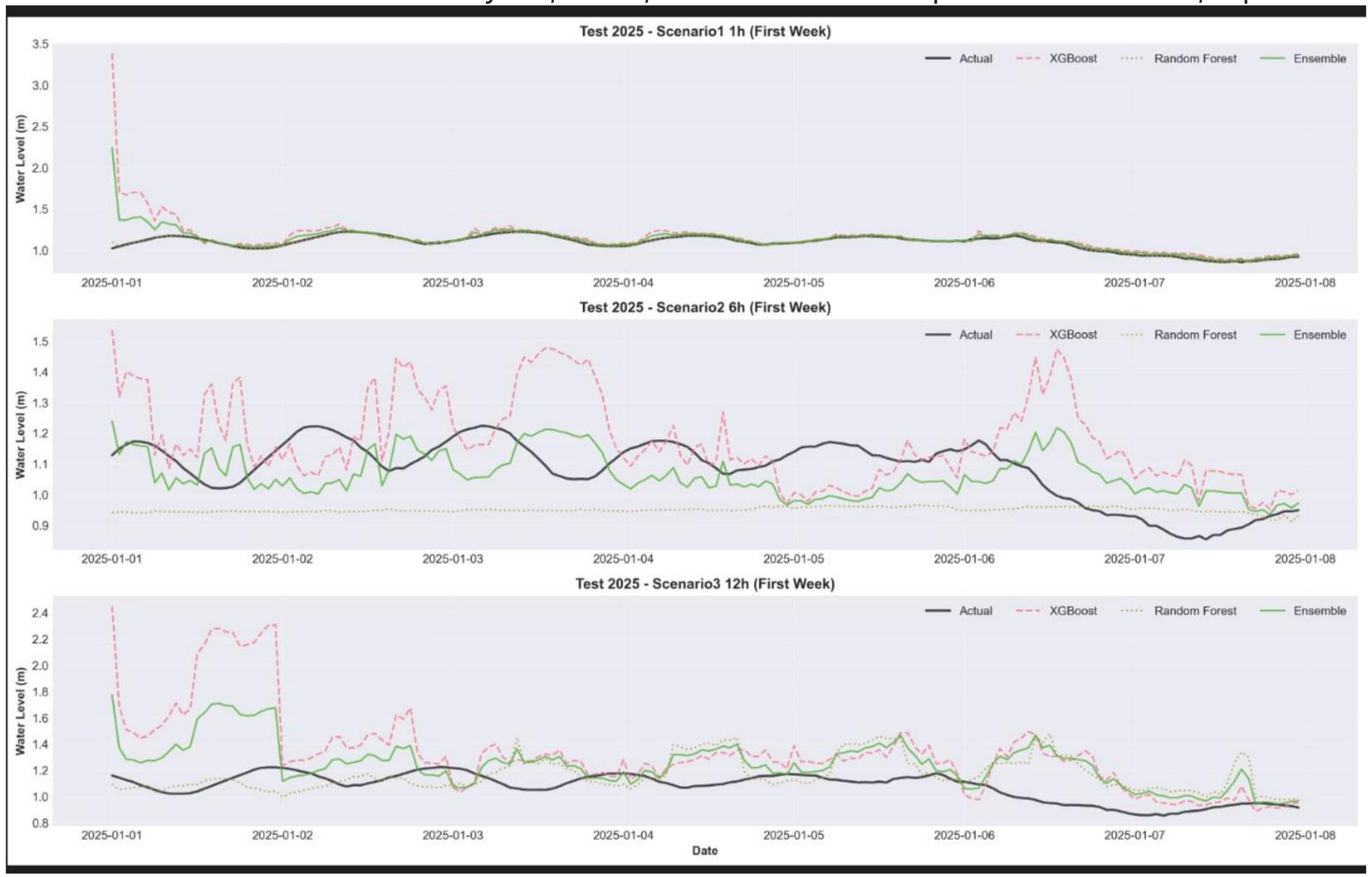
### What Changed in 2025?

- Higher baseline variability (+23% std dev)
- More extreme events (4× frequency)
- Different weather patterns
- Unprecedented events (Jan 1: 3.4m spike)

The plot on the left clearly illustrates the stark difference in RMSE between validation and test periods.

An extreme event on January 1st, 2025, saw water levels spike to 3.4 meters, a phenomenon





### Why This Matters: Implications for Real-World Deployment

The traditional machine learning assumption is that "train and test data come from the same distribution." However, in reality, "time series data, especially environmental data, often exhibits non-stationarity. Training on one year doesn't guarantee performance on the next."

Implications for Deployment

- Single-year training insufficient
- Need diverse conditions in training data
- Must expect performance degradation
- Continuous model updating required

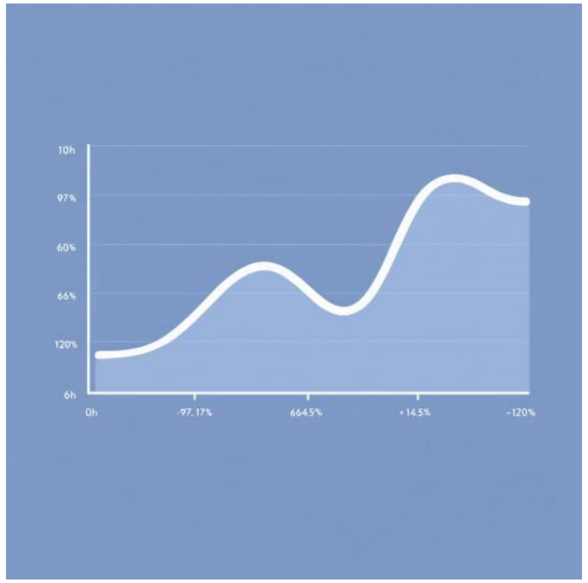
# LSTM's Fascinating V-Curve Pattern

While often celebrated for its ability to model sequential data, the Long Short-Term Memory (LSTM) network exhibited a peculiar "V-curve" performance across different forecast horizons. It struggled significantly at short horizons but surprisingly excelled at the longest.

### LSTM Performance Trajectory

| Horizon | RMSE | R² | vs Persistence |
|---|---|---|---|
| 1h | 18.44 cm | 0.211 | -97.1% |
| 6h | 39.67 cm | -2.648 | -66.3% |
| 12h | 21.57 cm | -0.078 | +14.2% |

This table summarizes LSTM's performance, showing a clear dip in effectiveness at short and medium horizons before a strong rebound at 12 hours, where it outperformed all other models.



The visual representation above highlights the "V-curve" — performance starts very low (relative to persistence) at 1-hour, improves slightly at 6-hours, and then dramatically shifts to positive at 12-hours.

## Interpretation: Why LSTM's Performance Varies

### Why LSTM Fails at Short Horizons (1h, 6h)

#### Overfitting to 2024 Patterns

LSTM learned specific, nuanced temporal sequences from the 2024 data. When 2025 introduced different short-term dynamics, these learned patterns failed to generalize, leading to poor predictions.

#### Insufficient Training Diversity

Unlike tree models that learn rules from features, LSTMs memorize sequence patterns. Training on only one year (2024) meant it lacked the diverse examples needed to handle the shifting patterns of 2025.

#### Short-term Noise Amplification

At 1h/6h, LSTM attempts to predict minute changes. Given 2025's higher variability and increased noise, the model's predictions became erratic, amplifying errors rather than smoothing them.

### Why LSTM Wins at Long Horizon (12h)

#### Captures Long-term Dependencies

With its 24-hour sequence memory, LSTM can identify and leverage multi-tidal patterns, something tree models, which treat each hour independently, cannot. It understands the "water rising for X hours likely means it will keep rising" context.

#### Trend Learning

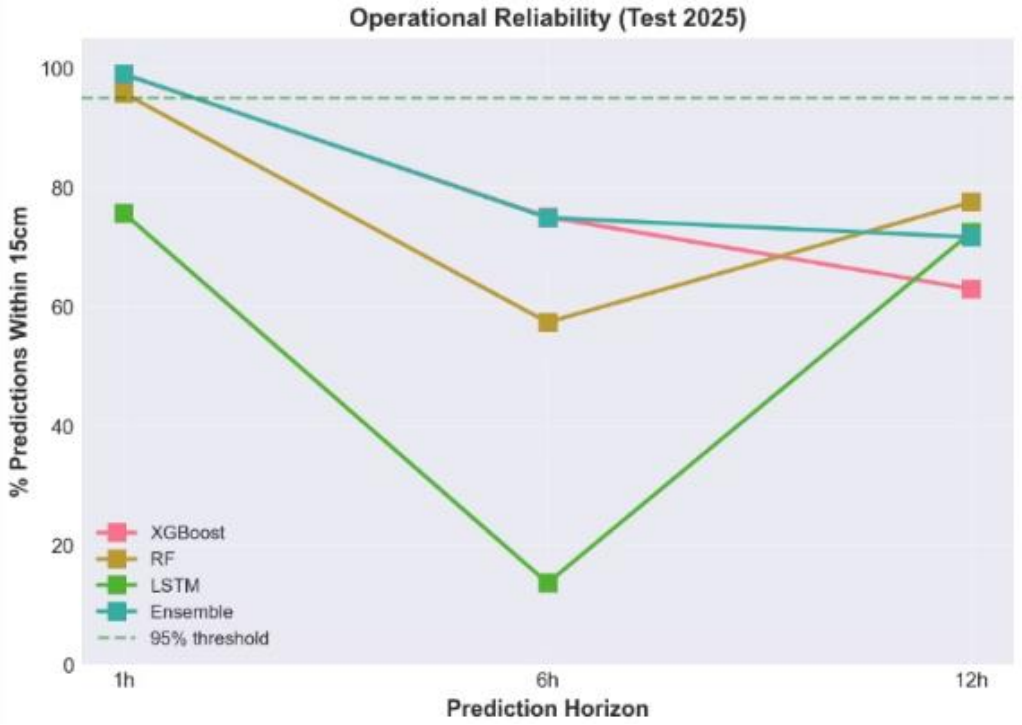
Long-term trends tend to be more stable and less susceptible to short-term fluctuations. LSTM's capacity to learn and extrapolate these 12-hour trends allowed it to perform more robustly, even with shifting patterns.

#### Averaging Effect

A 12-hour prediction naturally averages out much of the short-term noise. The long-term signal remains more consistent between 2024 and 2025, allowing LSTM's sequential nature to benefit from this inherent smoothing.

# Maintaining Operational Reliability

> The operational question we sought to answer was: "What percentage of predictions fall within ±15 cm (6 inches)?" This threshold is crucial for determining whether a forecast is truly operationally useful for coastal management and safety.



## Results - Predictions Within ±15cm

| Model | 1h Ahead | 6h Ahead | 12h Ahead |
|---|---|---|---|
| XGBoost | 98.9% ✅ | 75.0% ✅ | 62.9% ⚠ |
| Random Forest | 95.8% ✅ | 57.4% ⚠ | 77.5% ✅ |
| LSTM | 75.7% ✅ | 13.7% ❌ | 72.5% ✅ |
| Ensemble | 99.0% ✅ | 74.9% ✅ | 71.7% ✅ |

✅: >70% (operationally useful) | ⚠: 50-70% (marginal) | ❌: <50% (not operationally useful)

## Key Insights

### R² Doesn't Tell the Full Story

XGBoost at 6h had R² = -0.227 (worse than mean baseline), yet 75% of predictions were within ±15cm and it beat persistence by 3.5%. Statistical metrics alone can be misleading for operational utility.

### Reliability Degrades with Horizon

While 1h forecasts were highly reliable (96-99%), 6h forecasts showed mixed results (57-75%), and 12h forecasts were acceptable (63-78%), highlighting increasing uncertainty with longer prediction times.

### LSTM Unusable at 6h

With only 13.7% of predictions within ±15cm, the LSTM model at 6h was unequivocally unusable for operational deployment, as 86% of its forecasts exceeded the critical threshold.

### Ensemble Most Consistent

The Ensemble model maintained high reliability (72-99%) across all horizons, offering the most stable and predictable performance for critical operational deployment scenarios.

# Lessons Learned & Path Forward

## Future Work: Targeted Improvements for 30-40% Gain

Our next steps are designed to directly address the identified challenges, focusing on data enrichment and model sophistication to achieve substantial performance improvements.

### High Impact (Primary Focus)

1

#### Train on Full 2021-2024 Dataset

Expand training data beyond 2024 to include diverse conditions (2021-2023), improving model generalization. Expected: **15-25% RMSE reduction.**

2

#### Incorporate Weather Forecasts

Integrate 6-12h wind/pressure forecasts to provide forward-looking information. Expected: **20-30% improvement at 6h/12h.**

3

#### Develop Trend/Detrending Features

Implement 30-day moving averages and seasonal components to better separate signal from noise. Expected: **10-20% improvement.**

### Medium Impact (Secondary)

#### Time-Series Cross-Validation

Refine hyperparameter tuning with more robust temporal validation. Expected: **5-10% improvement.**

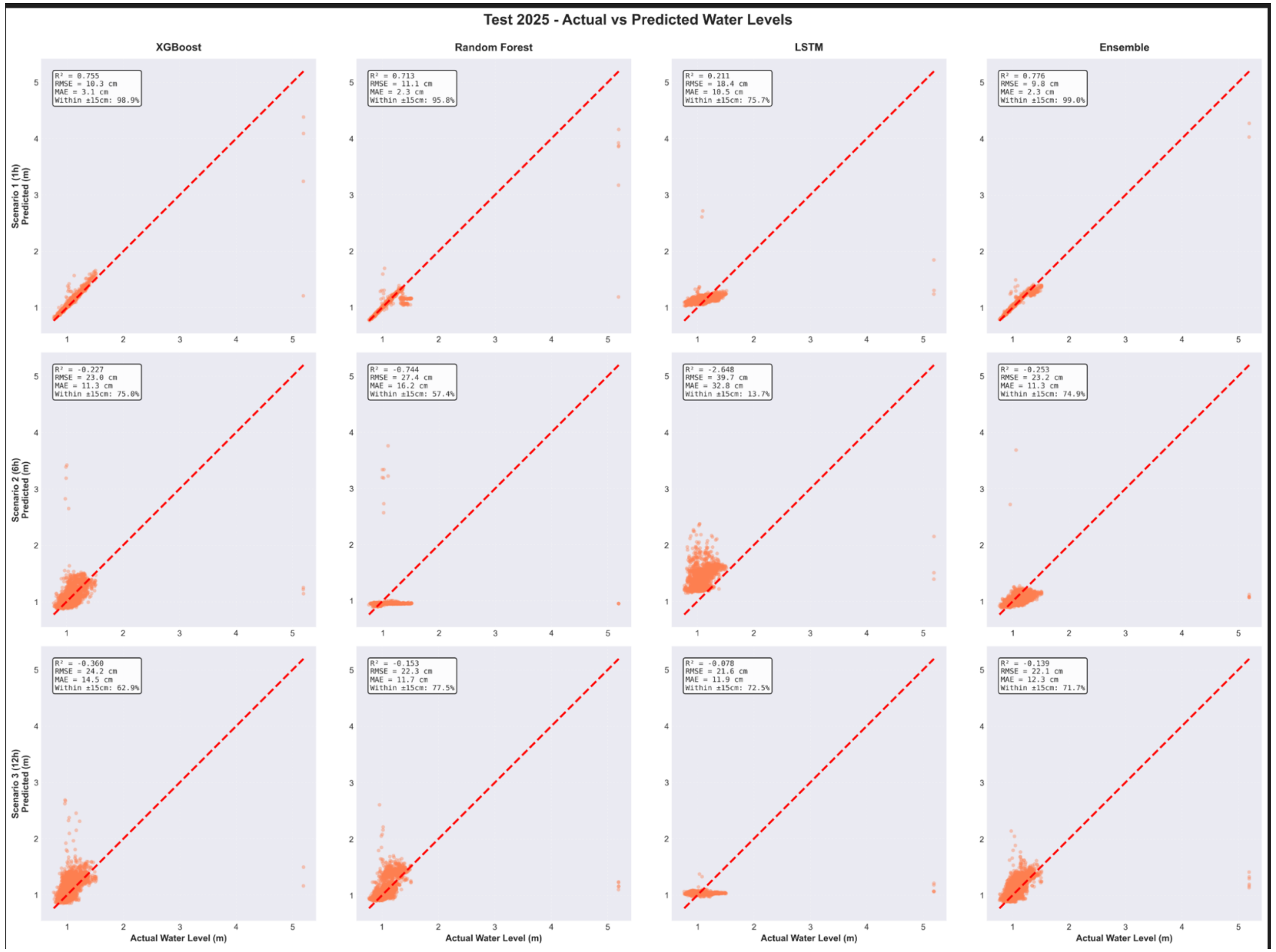
#### Advanced Ensemble Methods

Explore stacked generalization and weighted averaging for enhanced stability. Expected: **5-10% improvement.**

#### Uncertainty Quantification

Implement quantile regression and prediction intervals for better operational decision-making. Impact: **Improved operational reliability.**

# Visual Summary - The Complete Story



**Caption:** The plot above consolidates 12 subplots, showcasing the performance of four models (XGBoost, Random Forest, LSTM, Ensemble) across three critical forecast horizons (1h, 6h, 12h.

## What This Plot Reveals:

- **Row 1 (1h Ahead):** XGBoost, Random Forest, and Ensemble models exhibit reasonable scatter around the ideal prediction line. LSTM shows some data but is more scattered compared to others.
- **Row 2 (6h Ahead):** XGBoost maintains a moderate scatter and diagonal pattern. Random Forest shows more scatter with clustered predictions. LSTM displays wide scatter, indicating unreliable predictions. Ensemble performance is moderately scattered, similar to XGBoost.
- **Row 3 (12h Ahead):** XGBoost's predictions become more scattered. Random Forest maintains good diagonal correlation. LSTM shows a tighter correlation than at 6h, approaching the diagonal, demonstrating its V-curve pattern. Ensemble consistently maintains moderate scatter.

> From tight validation clusters to scattered test points. From model agreement to model divergence. From perfect predictions to honest uncertainty. This is what temporal distribution shift looks like.

# Questions & Contact

Got questions about our multi-model framework or the challenges of coastal water level forecasting? Let's connect and discuss the path forward for more accurate and reliable predictions.

> ### Key Takeaways
>
> - ✓ Multi-model framework beats persistence at 6h & 12h
> - ✓ Temporal distribution shift is real (2-6× degradation)
> - ✓ Different models win at different horizons
> - ✓ LSTM: Worst at 1h/6h, Best at 12h
> - ✓ Honest uncertainty > false precision

## Contact Information

**Abhishek Joshi**

Email: ajoshi5@islander.tamucc.edu

GitHub: https://github.com/abhishekjoshi007/Water-Level-Prediction