AIMS Environmental Science

AIMS Environmental Science, 12(5): 916–935.
DOI: 10.3934/environsci.2025040
Received: 12 July 2025
Revised: 05 October 2025
Accepted: 14 October 2025
Published: 29 October 2025

https://www.aimspress.com/journal/environmental

Research article

Salinity forecasting in Chao Phraya River using XGBoost with missing data handling

Jiramate Changklom[1], Trang Prommana[1], Phakawat Lamchuan[2] and Adichai Pornprommin[1,*]

[1] Department of Water Resources Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
[2] Office of Engineering and Architectural Design, Royal Irrigation Department, Bangkok 10900, Thailand

* **Correspondence:** Email: fengacp@ku.ac.th.

Supplementary A. Observed time series

Figure S1 illustrates the observed time series of the predictor variables used in the forecasting model from 2014 to 2020. Panel (a) shows salinity measured at the Samlae station (S1), which served as the target variable. Panel (b) shows upstream water levels at CPY006, reflecting the effect of seawater flushing from the Chao Phraya Diversion Dam. Panel (c) presents water levels at PAS009, influenced by the Pasak River and reservoir operations. Panel (d) displays tidal levels at the Bangkok station (BB), capturing tidal fluctuations at the river mouth.

These time series highlight differences in variability and data availability across stations, with substantial gaps evident in CPY006 and PAS009, while BB exhibited continuous records throughout the study period. Such discrepancies underline the challenges of missing data and justify the need for methods capable of robust forecasting under incomplete input conditions.

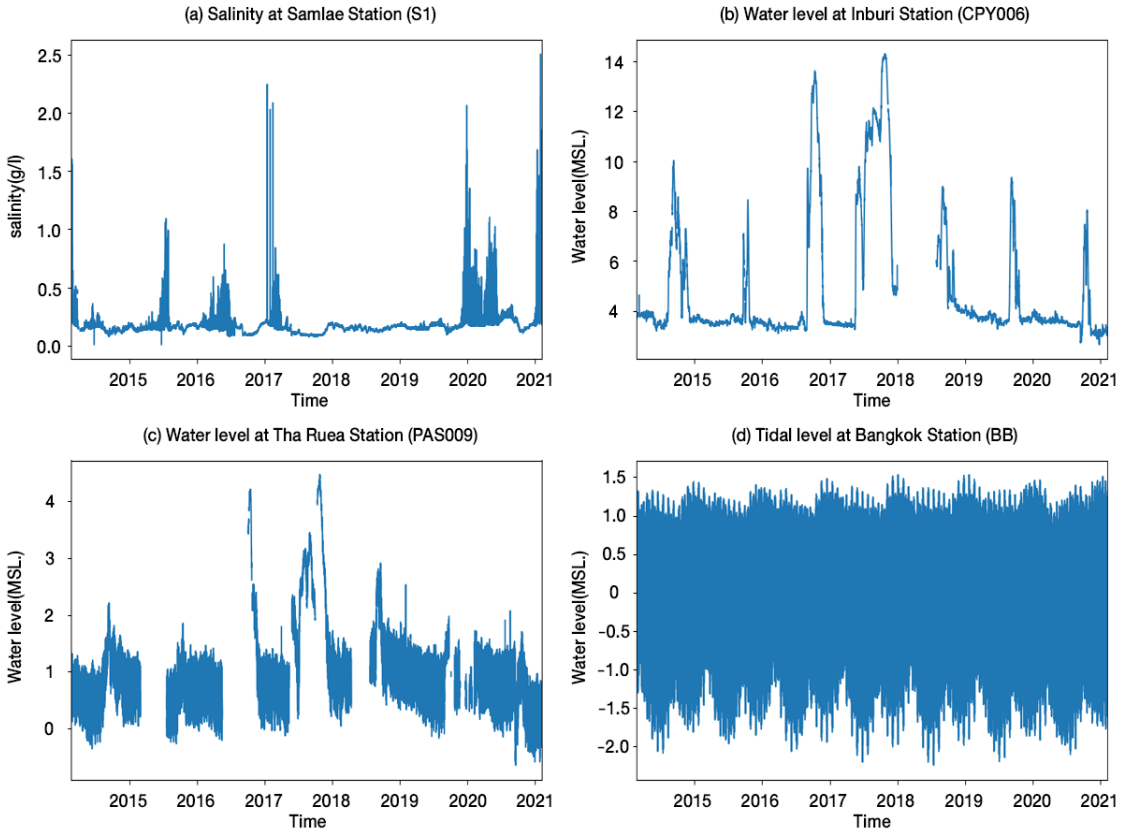

Figure S1. Time series of predictor variables: (a) salinity at S1, (b) water level at CPY006, (c) water level at PAS009, and (d) tidal level at BB during 2014–2020.

Supplementary B. Bootstrap significance testing

To assess the statistical significance of forecast performance differences between XGBoost and ANN models, a bootstrap resampling procedure was conducted. Performance metrics (RMSE and NSE) were resampled 10,000 times with replacement to estimate the distribution of model differences.

Table S1 presents the bootstrap results. For each comparison, the mean difference (Δ = XGB – ANN), the 95% confidence interval (CI) obtained from the bootstrap distribution, and the associated p-value are reported.

At the 24-hour horizon, XGBoost significantly outperformed the single-level ANN, yielding lower RMSE and higher NSE ($p < 0.001$). When compared with the multilevel ANN, the results showed slightly lower RMSE but reduced NSE; both differences were statistically significant.

At the 48-hour horizon, XGBoost again showed significantly better performance than the single-level ANN across both RMSE and NSE ($p < 0.001$). In contrast, differences between XGBoost and the multilevel ANN were not statistically significant, as indicated by wide 95% CIs and p-values > 0.05.

Table S1. Bootstrap comparison of forecast performance between XGBoost, single-level ANN, and multilevel ANN models at 24-hour and 48-hour horizons.

Forecast period	Metric	Comparison	Δ (XGB − ANN)	95% CI	p-value
24 hours	RMSE	XGB vs. Single-level ANN	-0.009	[−0.013, −0.004]	0.0002
	NSE	XGB vs. Single-level ANN	0.048	[0.024, 0.071]	0.0002
	RMSE	XGB vs. Multilevel ANN	0.006	[0.003, 0.009]	0.0000
	NSE	XGB vs. Multilevel ANN	-0.028	[−0.040, −0.016]	0.0000
48 hours	RMSE	XGB vs. Single-level ANN	-0.012	[−0.018, −0.005]	0.0008
	NSE	XGB vs. Single-level ANN	0.090	[0.041, 0.136]	0.0002
	RMSE	XGB vs. Multilevel ANN	-0.001	[−0.004, 0.003]	0.7444
	NSE	XGB vs. Multilevel ANN	0.004	[−0.024, 0.031]	0.7688

Supplementary C. Feature importance analysis

Figure S2 shows the feature importance results derived from XGBoost models under two forecast horizons: (a) 24 hours and (b) 48 hours. In both cases, the lagged salinity variable ($S1_{t-24}$ and $S1_{t-48}$) overwhelmingly dominated the model, with an important score far higher than the other predictors.

The remaining predictors—$CPY006_{t-125}$, $PAS009_{t-33}$ (24-hour model) or $PAS009_{t-57}$ (48-hour model), and BB_{t-6}—contributed marginally. Their relative values of importance were close to zero, indicating limited influence on forecast accuracy. These results are consistent across both forecast horizons and highlight that the model relied primarily on the lagged salinity input, with the other predictors playing supporting but less critical roles.

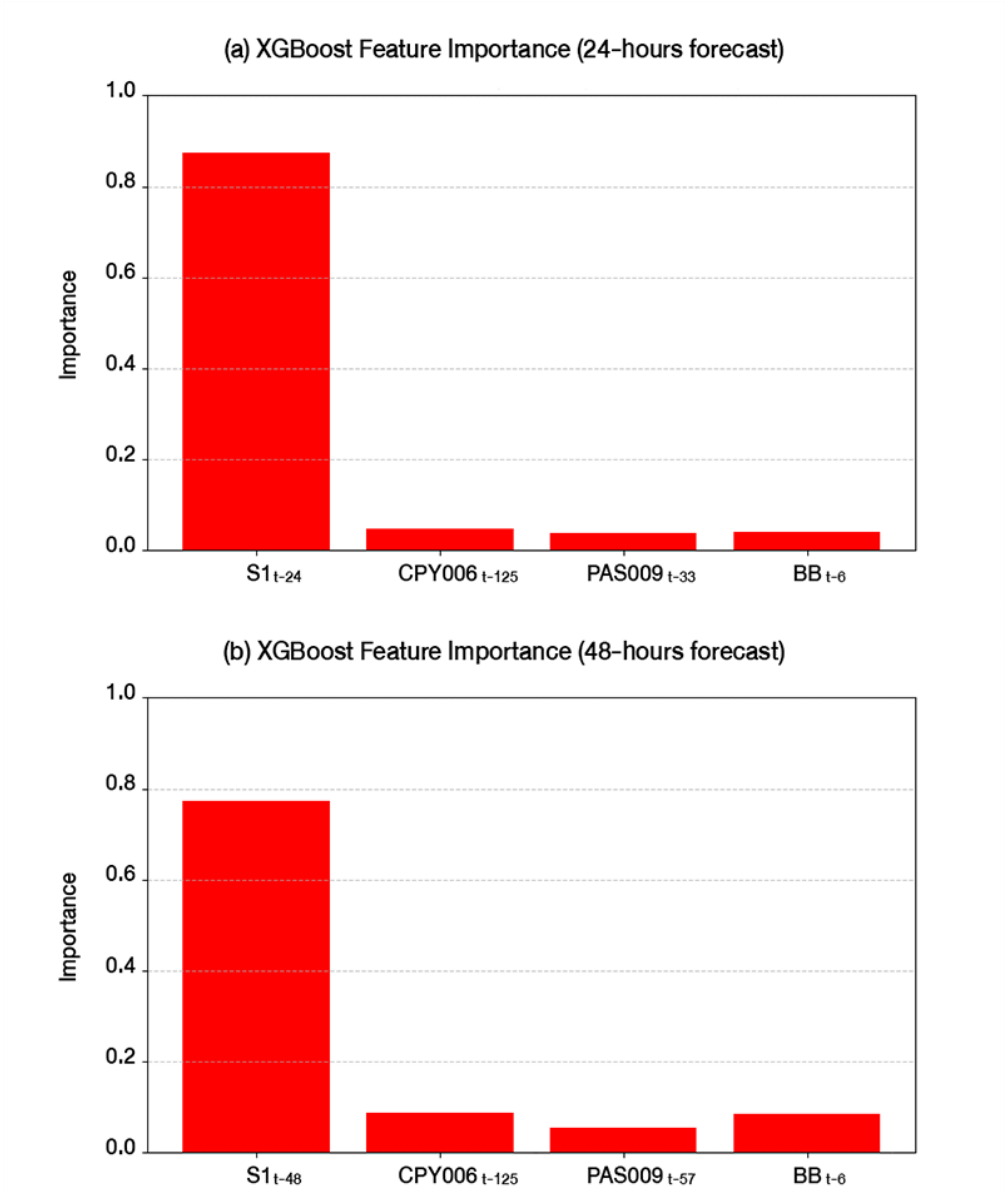

Figure S2. XGBoost feature importance for (a) 24-hour and (b) 48-hour forecast horizons.